Two faces are better than one: Face recognition in group photographs

Ohil K. Manyam [1] {omanyam}
Neeraj Kumar [2] {neeraj}
Peter Belhumeur [2] {belhumeur}
David Kriegman [1] {kriegman}

[1] University of California San Diego, La Jolla {@cs.ucsd.edu}
[2] Columbia University, New York {@cs.columbia.edu}

Abstract

Face recognition systems classically recognize people individually. When presented with a group photograph containing multiple people, such systems implicitly assume statistical independence between each detected face. We question this basic assumption and consider instead that there is a dependence between face regions from the same image; after all, the image was acquired with a single camera, under consistent lighting (distribution, direction, spectrum), camera motion, and scene/camera geometry. Such naturally occurring commonalities between face images can be exploited when recognition decisions are made jointly across the faces, rather than independently. Furthermore, when recognizing people in isolation, some features such as color are usually uninformative in unconstrained settings. But by considering pairs of people, the relative color difference provides valuable information. This paper reconsiders the independence assumption, introduces new features and methods for recognizing pairs of individuals in group photographs, and demonstrates a marked improvement when these features are used in joint decision making vs. independent decision making. While these features alone are only moderately discriminative, we combine these new features with state-of-art attribute features and demonstrate effective recognition performance. Initial experiments on two datasets show promising improvements in accuracy.

1. Introduction

With the advent of inexpensive digital cameras and social networking sites such as Facebook, millions of personal photographs are uploaded daily. Many photographs include multiple individuals. There is a strong desire to identify and tag faces in these photographs - automatically and accurately. And, unlike access control systems which typically include images with a single person, video surveillance images commonly include multiple people and even crowds. These scenarios require face recognition systems to identify multiple individuals in a single image and they have to be effective in unconstrained imaging conditions.

There has been a great deal of progress in recognizing people over pose and lighting variation [12, 15]. While these techniques seem reasonable for photographs containing single individuals, there is an opportunity to exploit the common imaging conditions across individuals in the same photograph. Figure 1 shows a few sample images from a television show. One can readily notice various correlations. Faces of taller people tend to appear higher than others in an image. Despite variations in lighting, fairer skinned individuals have brighter skin than darker skinned individuals. There is also similarity in the direction of shadows and gaze. Effects introduced by the camera are more subtle - color balance, spectral response, exposure setting, noise and even motion blur will all be similar for every face in a group photograph. Some of these variations such as height can potentially be used to aid in recognition. While others such as shadow or blur can be uniformly ignored for all people in the group shot.

Face recognition systems largely identify individuals independently. For an image with two detected faces, the recognition decision of one face does not influence the decision of the second face. For controlled imaging conditions, systems with remarkable performance have been developed [14]. In an unconstrained setting, the face recognition problem is harder. There is an emerging area that addresses this setting and evaluates performance on unconstrained datasets such as Labeled Faces in the Wild (LFW) [5] and PubFig [6]. While most methods operating in constrained settings have been based on aligning gallery and probe images in some way, including cross pose, some of

Figure 1. Collection of frames containing multiple individuals from the television show Buffy the Vampire Slayer

the most promising methods in unconstrained settings rely on extracting localized features (*e.g*. attributes, similes, etc.) that are trained with many examples images and are to a large extent insensitive if not invariant to imaging conditions. More recent techniques such as [2, 10] further push the accuracy envelope. Highly discriminative features are desired.

Some features, such as color, might be expected to be discriminative, but they are not in an unconstrained setting because image color is not only a function of face color, but also of illuminant color which varies. Yet when two faces are seen in the same image, the illumination color is very likely to be the same for both faces, and so the colors of the same pair of individuals as seen in two different images is expected to be highly correlated.

In this paper, we question the basic assumption of independence. Considering the many commonalities that people share in a group photograph, would it be beneficial to try and model these commonalities across groups of people? We try to answer this question by focusing on a few features that would usually be considered ineffective when considered individually (e.g. face color and the height of the face box in an image), and we show that there is significantly greater discriminative power of these features when used in joint decision making.

But do these features offer independent information when integrated in an existing system. To address this, we also extracted state-of-the-art attribute descriptors [6] that have been shown to perform well in an unconstrained setting, and we used both the attribute features and the new features for recognition in a generative (probabilistic) framework. We propose two models that attempt to recognize *pairs* of individuals in group shots. While one model computes the conditional probability of a person based on another person's feature vector (in addition to his own), the second model computes a joint probability for the pair of individuals. We evaluate this on two datasets containing groups shots, and show that joint decision making using the proposed methods exceeds independent decision making.

While recognizing multiple people simultaneously has been explored [13, 8], it was *meta-data* such as GPS locations and timestamps that was modeled. Our approach on the other hand models *visual* image features directly.

2. Relative Features

Group photographs contain multiple people in the same photograph – imaged in roughly the same illumination conditions with the same camera – and we would like to use features that can exploit this fact. In unconstrained settings, it is well known that a probe face can appear very different from its gallery examples, due to pose, expression, lighting, and imaging variations. For example, under bright illumination and heavy shadow, an intensity- or texture-based feature will produce a vastly different response than under more neutral conditions, leading to incorrect recognition.

However, if we look at pairs of faces in an image, we can use a *relative feature* such as the difference in brightness between the two faces. If one individual has a darker skin tone, then that person's brightness should generally be lower than the other's – regardless of the actual illumination conditions in the scene. Of course, the benefit of such relative features is not limited solely to lighting and color; a wide variety of attributes will show similar correspondences when multiple people are imaged together: relative heights, ages, degree of facial hair, nose size, eyebrow thickness, roundness of face, etc. While the variety of confounding effects in natural images will affect the appearance of (and therefore measurement of) these attributes, it will often do so to *all faces in the image* in a similar way, leaving their relative strengths roughly the same.

To explore the effectiveness of this approach, we use a number of different types of features, described in the following subsections. First, we use the describable visual attributes of Kumar *et al*. [6], which have been shown to be very effective on real-world face verification benchmarks such as Labeled Faces in the Wild (LFW) [5]. Second, we include color features that capture the median color and brightness of different parts of the face for each individual. These are expected to greatly benefit from using relative measurements. Third, we expand beyond the face to consider a feature that is heavily used by humans, but often ignored in computer vision – the height of a person. Of course, since we often do not see the full person in an image, we use the height of the face-detection box as a proxy for the person's height; however, we show that this is still effective.

2.1. Describable Visual Attributes

Kumar *et al*. [6] introduced a novel approach to face verification using classifiers trained to recognize the presence or absence of (and degree of) describable visual attributes such as age, gender, ethnicity, hair color, etc. These classifiers are trained using a supervised learning approach: hundreds of images are manually labeled for each attribute and used to train a binary classifier. The training process includes a greedy feature selection algorithm that automatically determines the most useful low-level features to use for a given attribute (via cross-validation), resulting in a system that can learn to classify new attributes efficiently and accurately given labeled training images.

We computed attribute values using the system described in [7], which consists of automatic face and fiducial point detection, affine alignment based on the fiducial points, and attribute classification using Support Vector Machines (SVMs). The complete list of 73 attributes used in this work can be found in [7].

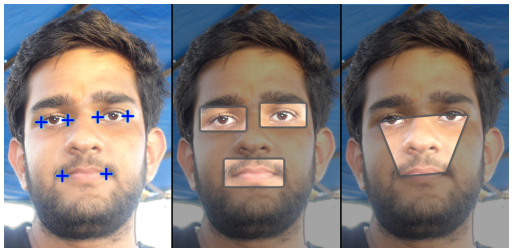

Figure 2. Detected fiducial points (left), eyes and mouth regions (center), and convex hull of fiducial points (right)

2.2. Color Descriptors

To capture characteristic color traits, we introduce four new descriptors. As described in the previous section, attribute computation includes fiducial point detection as a sub-step. This produces six keypoints corresponding to the corners of each eye and the corners of the mouth. With this, we define four regions from which we extract color information - two around the eyes, one around the mouth, and one on the convex hull of all 6 points. The first three are rectangular regions defined using two fiducial points each, with a width that is 115% of the distance between the two points and a height that is one half of this width. The latter encompasses primarily the nose and parts of the lower-eyes, upper lip, and cheeks.

Figure 2 shows the fiducial points automatically detected on the face, followed by the four regions. Each region is then converted to the HSV color space with a hue shift of $180°$. The median values for each color channel are then computed within the region, resulting in a three-component descriptor for each region. Finally, we note that if more feature points are available, additional regions can be defined in a similar way, *e.g.* around the forehead.

2.3. Height Descriptor

Similar to color, the height of a person can be a good relative descriptor. In addition, height is a form of non-facial information that is often ignored in recognition systems, despite its obvious usefulness, due to the difficulty in estimating it reliably from a single image. As an estimate of the height of an individual, we use the distance between the detected face box and the base of the image. Due to variations in camera position and ground level, a person's height so estimated may experience drastic fluctuations. Consequently, this height value may be a weak and even misleading descriptor for an individual; on the other hand, the face box of a taller person is more likely to be found higher in a photograph than that of a shorter person, and thus relative height values are expected to be discriminative.

To account for people closer to the camera appearing larger and taller, we normalize the distance of the face box from the base of the image by the size (height) of the face box itself. This ratio is treated as our height descriptor. Figure 3 illustrates this. Note that in the common case of individuals being photographed at the same distance from the camera (*i.e.*, many group photos), the size of all face boxes will be roughly identical and this measure reduces to the distance from the base of the image – which does indeed capture relative heights.

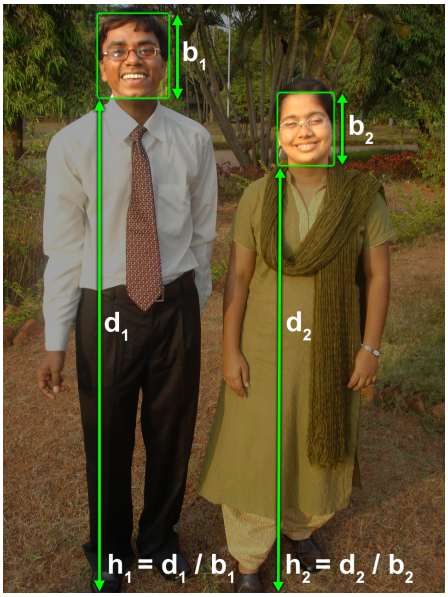

Figure 3. Computing height descriptors h_1 and h_2 for two people in a group photograph. Note that in the common case of people standing at the same distance from the camera, this measure provides an accurate measure of relative heights.

Thus, 73 attribute features combined with 12 color descriptors (3 for each of 4 regions) and 1 height descriptor together form an 86 dimensional feature vector for our experiments.

3. Recognition Models

We take generative Bayesian approach to recognition, starting with the likelihood of a feature vector $\vec{x}_1$ for person P_a as $P(\vec{x}_1 \mid P_a)$. For mathematical convenience, assuming equal prior probability on individuals, decisions can be said to have been made on $P(P_a \mid \vec{x}_1)$. Extending this, for a pair of people P_1 and P_2, we use feature vectors $\vec{x}_1$ and $\vec{x}_2$ as such:

$$\langle P_1, P_2 \rangle = \underset{\langle P_a, P_b \rangle,\ a \neq b}{\arg\max}\ P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) \qquad (1)$$

where P_a and P_b range over all K individuals in the dataset. We present three different approaches to modeling the joint $P(P_a, P_b \mid \vec{x}_1, \vec{x}_2)$ – the usual model assuming statistical independence, a conditional probability model and a joint probability model.

3.1. Baseline Model

A regular model that assumes statistical independence between detected faces (henceforth termed the baseline model) computes

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)\, P(P_b \mid \vec{x}_2) \quad (2)$$

i.e., the probability of feature vector $\vec{x}_1$ belonging to person P_a is treated independent of $\vec{x}_2$ belonging to P_b – even though $\vec{x}_1$ and $\vec{x}_2$ are derived from the same image.

Computing the class conditional density directly, we use a Gaussian as our baseline model. Mathematically, the model for person P_a is

$$P(P_a \mid \vec{x}_1) = \mathcal{N}(\mu_a, \Sigma_a) \quad (3)$$

where μ_a is the mean feature vector for P_a and Σ_a is the covariance of these vectors - both computed from training examples. Due to the relatively high dimensionality of our feature vectors, we employ Fisher's Linear Discriminant Analysis (FLDA) [1] to project them into a discriminative lower dimensional subspace. With K individuals in the dataset, this subspace has $K-1$ dimensions. The baseline Gaussian model is trained using vectors in this subspace.

3.2. Conditional probability model

Identifying two people P_a and P_b from their feature vectors $\vec{x}_1$ and $\vec{x}_2$ using a conditional probability model is formulated as

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)\, P(P_b \mid \vec{x}_1, \vec{x}_2, P_a) \quad (4)$$
$$= P(P_b \mid \vec{x}_2)\, P(P_a \mid \vec{x}_2, \vec{x}_1, P_b) \quad (5)$$

$P(P_a \mid \vec{x}_1, \vec{x}_2) = P(P_a \mid \vec{x}_1)$ is reasonably and implicitly assumed. Note that though equations 4 and 5 are mathematically equivalent, they may produce different results in practice due to differing numbers of training examples. Thus, it is advantageous to compute both and then combine them in some way.

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 the previous section. 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 very hard to estimate due to the scarcity of training data and because both $\vec{x}_1$ and $\vec{x}_2$ are real-valued vectors.

To bypass this problem, we define the approximate conditional in terms of a relative binary feature vector. For 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

$$C^i_{ab} = \begin{cases} 1 & \text{if } x^i_a \geq x^i_b \\ 0 & \text{if } x^i_a < x^i_b \end{cases} \quad (6)$$

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

The training phase involves learning the probability that feature component i is greater for person a than for person b. This involves simple counting over images where person a and person b occur in the *same* photograph. Mathematically, setting $z^i_1 = \text{count}(x^i_a \geq x^i_b)$ and $z^i_0 = \text{count}(x^i_a < x^i_b)$,

$$P(C^i_{ab} = 1) = \frac{z^i_1}{z^i_0 + z^i_1} \quad (7)$$

$$P(C^i_{ab} = 0) = \frac{z^i_0}{z^i_0 + z^i_1} \quad (8)$$

Intuitively, these equations express 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 that for person b, *i.e.*, one would expect $P(C^{male}_{ab} = 1) \approx 1$ and $P(C^{male}_{ab} = 0) \approx 0$. For attributes where the two individuals being compared have nearly similar values, then both these probabilities would be close to 0.5.

One drawback of such a counting estimate is that the probability $P(C^i_{ab})$ can vanish to 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 is usually a byproduct of having to work with limited data. As a simple fix to this problem, we perturb each probability estimate $P(C^i_{ab})$ towards 0.5 by a small amount:

$$P(C^i_{ab}) = P(C^i_{ab}) + \frac{\text{sign}(0.5 - P(C^i_{ab}))}{z^i_0 + z^i_1 + 1} \quad (9)$$

This changes the probability by a fraction $1/(z^i_0 + z^i_1 + 1)$ in such a way that values less than 0.5 are increased and values greater than 0.5 are decreased. Intuitively, for each probability estimate $P(C^i_{ab})$, we assume the existence of one additional data point belonging to the category with lower probability. This technique successfully eliminates perfect zero probabilities in a controlled manner.

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

$$P(P_b \mid \vec{x}_1, \vec{x}_2, P_a) \approx \prod_{i=1}^{D} P(C^i_{ab} = C^i_{12}) \quad (10)$$

This assumes individual components of the relative binary feature vector are independent of each other. In view of this, we apply FLDA to our feature vectors before computing the binary conditional probability. The FLDA projection is restricted to 20 dimensions. The value for $P(P_a \mid \vec{x}_2, \vec{x}_1, P_b)$ can be computed similarly.

This conditional approximation is used along with the baseline models, following equations 4 and 5. We use the geometric mean of the two estimates thus produced for making recognition decisions.

3.3. Joint probability model

The binary conditional model ignores the raw value of feature components – instead looking only at relative higher-lower relationships. A joint probability model, on the other hand, can learn raw feature values for two individuals as well as a correlation between them. Similar to our baseline model, we learn a single Gaussian for each pair of individuals. The input vector for such a model is simply $[\vec{x}_1, \vec{x}_2]$, a concatenated version of feature vectors from the two faces. We learn $K(K-1)$ pairwise models for K individuals in the dataset. The concatenated feature vector is twice the usual feature size and we apply FLDA to reduce its dimensionality. Instead of the default $K(K-1)-1$ dimensional subspace generated by FLDA (which can be quite huge), we restrict ourselves to a 20-dimensional subspace. Mathematically,

$$P(P_a, P_b \mid \vec{x}_1, \vec{x}_2) = \mathcal{N}(\vec{\mu}^{ab}, \Sigma^{ab}) \tag{11}$$

where $\vec{\mu}^{ab}$ is the 20-dimensional mean for the FLDA projection of the concatenated vectors $[\vec{x}_a \ \vec{x}_b]$, Σ^{ab} is the 20×20 diagonal covariance matrix for the same vectors, and $\mathcal{N}(\vec{\mu}^{ab}, \Sigma^{ab})$ is the single Gaussian learned for the pair of individuals P_a and P_b.

4. Datasets

Due to the popular implicit assumption of statistical independence between detected faces in a group shot, most face recognition datasets do not have images containing multiple individuals. To generate such a dataset, one option would be to use existing datasets captured under controlled conditions, such as CMU PIE [11] or Multi-PIE [4]. By using subsets of these datasets corresponding to constant lighting, background, or other imaging parameters, one could simply assume that photographs of two different individuals originated from the same group shot. However, we believe that such a synthetic dataset would be incompatible with our premise of trying to learn true correlations between face regions from a single photograph. Consequently, we use two real datasets: the "Buffy" dataset introduced by Everingham *et al*. [3] and a new dataset constructed from a personal photo album.

4.1. The Buffy Dataset

First used by Everingham *et al*. [3], the Buffy dataset consists of roughly $120,000$ total frames extracted from two episodes of the popular television series *Buffy the Vampire Slayer* (Season 5, Episodes 2 and 5). Manual annotations corresponding to the $50,000$ automatically-detected face boxes are provided for each image frame, covering 11 primary characters and a number of supporting cast and extras. Everingham *et al*. used this dataset to test their automatic character-naming system for TV shows. Their system used a variety of features, including intensity and SIFT [9]-like features computed around fiducial points, clothing-color descriptors, visual speaker identification, and speaker information from subtitles. They reported an accuracy of around 69% for recognizing all detected face images in both episodes, while accuracy was around 80% when labeling the 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 run through the attribute generation pipeline described in Section 2.1, followed by computation of the color and height descriptors from Sections 2.2 and 2.3. We use data from episode 2 for training and test on episode 5. We identify two subsets of feature data for each episode. The first, which we call *group-data*, consists of features computed for 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 from all occurrences of each individual in an episode (not just group shots). We call this subset *all-data* and it includes all of group-data as a subset. All-data contains 2 to 8 times more images than group-data for this dataset.

By design, our joint models will only be able to use group-data, and are hence trained on that subset. On the other hand, all-data is used to train the baseline models (which assume statistical independence) for a fair comparison. Both models are tested only on group-data from the test episode.

While the Buffy dataset is large and was already available for us to use, it is non-ideal in several respects for this work. Since the dataset is derived from a television show, heavy makeup and artistic camera effects are common. A large fraction of the show is shot at night with artificial directional lighting. Since frames are extracted from video, a small amount of motion blur is present. Also, frames capture a snapshot of character movements (talking, walking or even fighting), and in some sense, represent a wider variety of pose and expression than what one would encounter in many real-world situations, such as personal photo albums. Finally, as in any situation where actors are involved, most faces are non-frontal to avoid breaking the "fourth-wall."

4.2. A Personal Photo Album

To mitigate these issues, we also run experiments on a new dataset constructed from one of our own personal photo albums. This dataset consists of approximately $1,700$ 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. Unlike the Buffy dataset, most images have individuals posing for the

camera and hence contain frontal shots. Automatically detected face boxes were manually annotated to obtain 116 unique individuals.

We randomly select 70% of the images for training and use the rest for testing. (All faces in an image are considered part of the training or testing process irrespective of the number of individuals in the image.) In order to build good joint models, each with an appreciable amount of data, we restrict our experiments to two sets of people. The first set of individuals occur in at least 80 training images, while the second set occur in at least 65 training images. This constitutes 6 individuals (P_1 to P_6) and 12 individuals (P_1 to P_{12}), respectively. Unlike the Buffy dataset, most photographs here are group shots. Furthermore, every pair of individuals has at least 10 training images.

Although the 12-person set contains more individuals (and thus pairs) than the 6-person set, the number of training instances available is low for individuals P_7 to P_{12} and pairs involving them. This, coupled with the greater number of pairwise classes, means that 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.

4.3. Group Data Scarcity

Data scarcity is a major issue when building models to recognize pairs of individuals. Consider the case where the training dataset is a personal photo album comprising K individuals $P_1, P_2, \cdots, P_K$. Let n_p denote the number of photographs in which person p occurs. A conventional face recognition system, building independent models for each person would build K such models. For each individual p, it would be able to use all n_p images. On the other hand, a system recognizing pairs of people would have to build $K(K-1)$ models, corresponding to every *ordered pair* of individuals (p, q). Furthermore, the model for (p, q) can only use that fraction of n_p or n_q images where p and q occur together. On average, this would be $n_p/(K-1)$ or $n_q/(K-1)$ images. Thus, each pairwise model for person p will always have less training data than its corresponding independent model. However, with the ease of capturing and storing ever-increasing numbers of photos, this limitation may not have a practical impact in the near future.

5. Experiments

Given an image containing n individuals, $n(n-1)$ ordered pairs are possible. We treat each of these as separate pair-recognition problems. If a particular test pair had no corresponding training pairs, then this test pair is simply ignored. Accuracy is computed *per-person – i.e.*, if a model correctly recognizes one person, but makes a mistake with the other, this is counted as 1 correct and 1 incorrect recognition.

For each dataset, we experiment with three sets of feature vectors – attributes alone, our new color and height based descriptors alone, and both attributes and our new descriptors. We now present recognition results for each of our three models.

5.1. Baseline

While the conditional and joint probability models produce pair-recognition decisions, the baseline model by design identifies one person at a time. In order to enable easy comparison of accuracy values, our baseline model is presented with the same pairs of individuals as our other models. Due to the independence assumption, separate recognition decisions are produced for each individual in the pair. We apply FLDA as stated in Section 3.1 before learning Gaussian models. Results from this experiment can be found in Table 1.

	Buffy	**Photo Album**	
		6 people	12 people
Attrib. only	63.56	91.38	89.59
New descp. only	32.92	61.08	42.14
Attrib. + new descp.	64.60	92.86	89.82

Table 1. Gaussian baseline accuracy (in percentage)

As seen from the table, performance is poor using just our new descriptors. This is expected, as these descriptors are few in number and rather weak on their own. Using attributes along with the new descriptors is better than using attributes alone. We believe FLDA is largely responsible for this increase, as skipping it caused the new descriptors to have a detrimental effect when included with attributes.

5.2. Conditional probability

We implement the conditional model from Section 3.2 and use the best Gaussian model from the previous experiment as our baseline model (Gaussian trained on a combination of attributes and our new descriptors, with FLDA). Results of this experiment are presented in Table 2.

	Buffy	**Photo Album**	
		6 people	12 people
Attrib. only	72.00	91.87	88.92
New descp. only	66.66	93.10	89.74
Attrib. + new descp.	69.14	91.62	87.80

Table 2. Binary conditional model accuracy (in percentage)

Using our new descriptors alone for the binary conditional model, an improvement in accuracy over the best baseline can be seen for the Buffy dataset and the 6 person photo album. For the 12 person photo album, due to an increase in the number of recognition pairs and having less data, we notice a slight decrease in accuracy when compared to the best baseline. Also, the largest improvement in accuracy for the Buffy dataset is seen when the conditional model is trained on attributes alone. Attribute values are

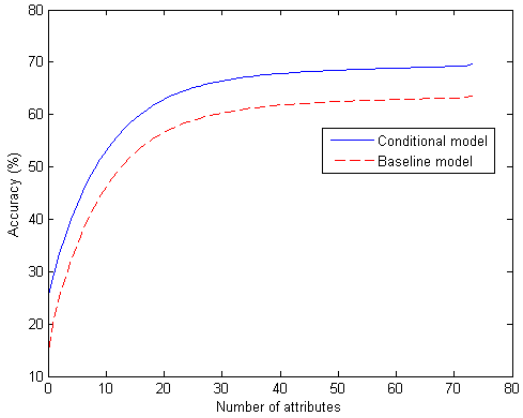

Figure 4. Variation in accuracy with number of attributes used for baseline (dashed red) and conditional (solid blue) models on the Buffy dataset. The latter consistently do better.

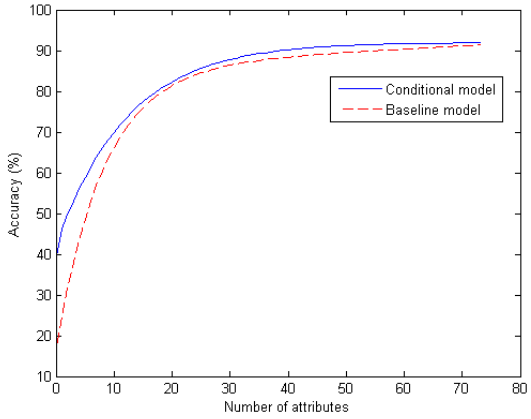

Figure 5. Variation in accuracy with number of attributes used for baseline (dashed red) and conditional (solid blue) models on the 6 people photo album dataset. The latter consistently do better.

intended as binary outputs from SVMs and we believe the binary conditional model is exploiting this.

To further illustrate the accuracy increases provided by the conditional model, we train a baseline model with a random subset of attribute features, for an increasing number of 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. The entire experiment is repeated and results from each run are averaged. Figure 4 shows results for this experiment on the Buffy dataset, and Figure 5 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 baseline models.

5.3. Joint probability experiment

Following the framework detailed in Section 3.3, we build a 20-dimensional Gaussian for pair of individuals. Due to group data scarcity and despite applying FLDA, the covariance matrix is singular for many models. So, we settle for a diagonal covariance matrix – effectively modeling each of the 20-dimensions separately. Results for this experiment can be found in Table 3.

	Buffy	**Photo Album**	
		6 people	12 people
Attrib. only	50.40	90.15	86.30
New descp. only	50.26	58.13	46.03
Attrib. + new descp.	53.52	91.87	88.55

Table 3. Gaussian joint model accuracy (in percentage)

Comparing this to Table 1, we notice that when using attributes alone or in combination with our new descriptors, net accuracy for all datasets drop when compared to corresponding baseline versions. This decrease is largely attributed to the reduced amount of data available for each pairwise model. For the Buffy dataset, reduction in training data is substantial and accounts for the large decrease in accuracy. Using our new descriptors alone, a 17% boost in accuracy is seen for the Buffy dataset and nearly 4% for the 12 person photo album when compared to similar baseline experiments. This shows that while the joint model is able to exploit natural correlations encoded in our new descriptors, providing an increase in accuracy, the greatly reduced amount of group data hurts performance more than the gains.

To further understand the performance boost due to our new descriptors, we compare the accuracy of a baseline model and a joint model both trained with just one of our new descriptors. Figure 6 shows results for this experiment on the Buffy dataset. Figure 7 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. While the overall accuracy of each descriptor is numerically low, every descriptor provides a boost in accuracy when used in a joint Gaussian model, often substantially so.

6. Conclusions and Future Work

Face recognition systems have traditionally built models for each individual in isolation. In group photographs, statistical independence is usually assumed between the detected faces during recognition. However, by virtue of the fact that all individuals in a group photo are in the same scene and captured by the same imaging system, there are a number of exploitable characteristics, such as common lighting, blur, *etc*. We have taken the first steps in using this information by building joint and conditional models for recognizing pairs of people in group photos more accurately. Our models use a variety of features, including describable visual attributes, median color and lighting in different regions of the face, and normalized height, which show the gains possible by using relative features. When

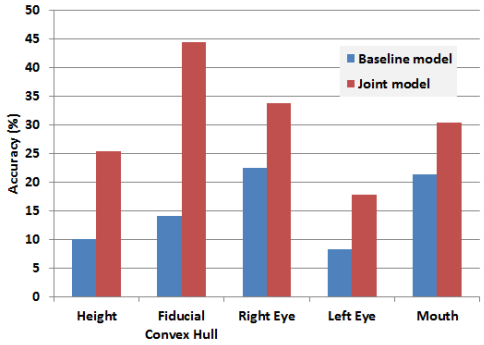

Figure 6. Comparison of accuracies provided by baseline (blue) and joint (red) models using our new descriptors on the Buffy dataset. The joint models consistently do significantly better.

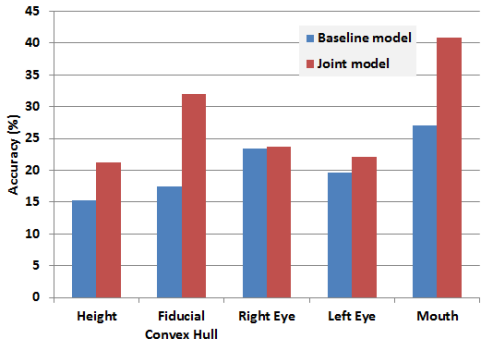

Figure 7. Comparison of accuracies provided by baseline (blue) and joint (red) models using our new descriptors on the 6 people photo album. The joint models consistently do better.

combined carefully (*e.g.*, using LDA), these models provide improvements over baseline techniques – the conditional model more so than the joint, which suffers due to data scarcity. Further, the conditional model can be easily added to existing face recognition systems, providing an accuracy boost when looking at group photographs.

We believe that a promising area for future work is in questioning many assumptions of traditional face recognition – independence of multiple faces, the use of only facial appearance, single-image processing – to exploit the numerous other sources of information present in many typical operating scenarios. First, it may be possible to formulate additional descriptors that capture commonalities in group photographs, including possibly those built on clothing or the body. Second, relative information can be inferred through transitivity to mitigate the data scarcity issue. For example, if two particular people are never seen together in group shots, but each is seen with a common third person, one could transitively infer relationships between the features of these two individuals. Indeed, one could extend this to chains of inference through multiple individuals. Relatedly, while we have shown how to exploit common information between pairs of people, more gains might be possible by using triplets or an even greater number of individuals simultaneously, *i.e.*, because there would be more constraints for each person. Finally, techniques that can squeeze more information from existing data would be very useful – for example, to learn from a single group photograph, or leverage estimates from all pairs (or triplets, *etc.*) in an image to form a single recognition decision for every person.

Acknowledgment: This work was supported in part by ONR MURI Grant N00014-08-1-0638.

References

[1] 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.

[2] D. Cox and N. Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *IEEE Intl. Conf. on Automatic Face Gesture Recognition and Workshops*, pages 8 –15, 2011.

[3] 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.

[4] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *IEEE Intl. Conf. on Automatic Face Gesture Recognition*, pages 1 –8, 2008.

[5] 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.

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

[7] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2011.

[8] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *Proc. European Conf. on Computer Vision*, pages 243–256, 2010.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 2003.

[10] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV (2)*, pages 709–720, 2010.

[11] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *IEEE Conf. on Automatic Face and Gesture Recognition*, pages 46–51, 2002.

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

[13] M. Zhao and S. Liu. Automatic person annotation of family photo album. In *Proc. Intl. Conf. on Image and Video Retrieval*, pages 163–172, 2006.

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

[15] 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.