THESIS

# SOFT BIOMETRICS ARE HARD


Submitted by

Hao Zhang

Department of Computer Science


In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2013

Master's Committee:

Advisor: Ross Beveridge

Committee Member: Bruce Draper
Committee Member: Geof Givens

# ABSTRACT

## SOFT BIOMETRICS ARE HARD

Soft biometrics typically refer to attributes of people such as their gender, the shape of their head, the color of their hair, etc. There is growing interest in soft biometrics as a means of improving automated face recognition since they hold the promise of significantly reducing recognition errors, in part by ruling out illogical choices. Here four experiments quantify performance gains on a difficult face recognition tasks when standard face recognition algorithms are augmented using information associated with soft biometrics. These experiments include a best-case analysis using perfect knowledge of gender and race, support vector machine-based soft biometric classifiers, face shape expressed through an active shape model, and finally appearance information from the image region directly surrounding the face. All four experiments indicate small improvements may be made when soft biometrics augment an existing algorithm. However, in all cases, the gains were modest. In the context of face recognition, empirical evidence suggests that significant gains using soft biometrics are hard to come by.

TABLE OF CONTENTS

# Chapter 1

# Introduction

This thesis explores the performance gain when using the information of soft biometrics to improve existing face recognition algorithms. A challenging data set for face recognition, the Good,the Bad and the Ugly data set is introduced, as well as two baseline algorithms, Local Region PCA and Cohort Linear Discriminant Analysis. The review of related work is also presented. Moreover, my own work of classifying soft biometrics and using them to improve the performance of the baseline algorithms is described. Two new soft biometrics, face geometry and face halo, are created and their performance on improving baseline algorithms is examined. In order to show the potential capability of perfect knowledge of soft biometrics, experiments using the ground truth of gender and race are conducted. A discussion of the experimental result, a conclusion of the whole thesis and future work conclude this thesis.

## 1.1 Soft Biometrics

In recent years, soft biometrics have generated considerable interest in the research community as a possible method for improving face recognition performance. It is stated in [12] that soft biometrics are defined as characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals. Wikipedia defines soft biometrics as physical, behavioural or adhered human characteristics, classifiable in predefined human compliant categories, established and timeproven by humans with the aim of differentiating individuals. Unfortunately, there is no universally accepted definition of the term *soft biometric*. In its strongest form, soft biometrics are discrete features that divide people into non-overlapping groups, such as gender, age, or eye color. Weaker definitions admit any non-facial feature of a person, for example weight or hair color. Still other researchers use the term to refer to any attribute of a face that is extracted and analyzed independently of the subject's identity. Examples of this range from small localized features such as moles or scars to slightly larger features such as periocular regions to wholistic features extracted from the face such as gender or race.

What these definitions have in common is the idea of a soft biometric as an attribute that is computed independently from the subject's identity. One difference between soft biometrics is that they can be presented in discrete or continue values. They can be a discrete category used to detect mismatches. Taking gender as an example, it can take only two discrete values - male and female. An image pair of different genders can be detected as a non-match pair without further computation. Soft biometrics can also be continuous numbers to be combined with other similarity measures. Still taking the example of gender, a gender predictor can give continuous positive numbers to indicate a male and negative numbers to indicate a female. It makes sense that some men look more manly than other men and some women look more feminine than other women. Those continuous numbers can be used to weight the similarity score of an image pair computed from certain face recognition algorithm. One might argue that soft biometrics are just a new name for facial similarity measures in this case. Another difference between soft biometrics is that some of them come from the face (e.g. eye color, gender or race), and some of them do not belong to the face (e.g. weight or hair color). Soft biometrics that belong to the face can be learned from face images while those do not come from face can not be learned given only face images.

## 1.2 Background

Researchers have used soft biometrics in various ways to tackle the problem of face recognition. Park and Jain [11], [12] use gender and local facial marks as soft biometrics and combine them with a traditional face recognition algorithm (FaceVACS). They were able to increase the recognition rate by about 1% on the FERET data set. Dantcheva et al. [6] looked at eye color in the visible spectrum as a soft biometric, but noted that 90% of the population have brown irises. Lyle et al. [17] analyzed images from the FRGC data set and classified periocular regions by gender and race. They were able to reduce the equal error rate by fusing the periocular soft biometric data from both eyes with LBP over the face region.

Kumar et al. [14] introduce the use of describable visual attributes, which are continuous labels that can be given to an image to describe its appearance, for face verification and image search. However, only a subset of what they are using are soft biometrics. Scheirer et al. [25] extend Kumar et al.'s work by introducing a Bayesian approach to combining descriptive attributes and producing accurate weighting factors to apply to match scores from face recognition algorithms based on incomplete observations made at match time. However, the descriptive attributes they use include a person's occupation, places they live and so forth. Table 1.1 shows some of the soft biometrics/attributes used frequently. The second column indicates their relationship to the face. Some of them are independent from the face, some are directly related to the face and some are indirectly related to the face but can be predicted from it. The third columns shows their category. They can be soft biometrics (S. B.), personal attribute (P. A.), or environmental attribute(E. A.).

| Attribute / Soft Biometrics | relationship | Category |
| --- | --- | --- |
| Gender | Indirect | S. B. |
| Race | Indirect | S. B. |
| Age | Indirect | S. B. |
| Hair | Direct | S. B. |
| Wearing Hat | Direct | S. B. |
| Eyebrow | Direct | S. B. |
| Eye Color | Direct | S. B. |
| Glasses | Direct | S. B. |
| Cheek Color | Direct | S. B. |
| Nose Shape | Direct | S. B. |
| Mustache | Direct | S. B. |
| Skin Color | Direct | S. B. |
| Face Shape | Direct | S. B. |
| Height | Independent | S. B. |
| Weight | Independent | S. B. |
| Occupation | Independent | P. A. |
| Lighting Condition | Independent | E. A. |
| Places | Independent | E. A. |

Table 1.1: **Attributes/soft biometrics used frequently**

## 1.3 Research Goals

Current face recognition algorithms are evaluated via face verification. Given a pair of two face images, the algorithms return a similarity score describing the similarity of two images. Ideally, a pair containing the same subject (a match pair) will be given a high similarity score and a pair containing different subjects (a non-match pair) will be given a low similarity score. Using soft biometrics can help compute the similarity score. As aforementioned, soft biometrics can be presented using discrete or continuous values. If they are in discrete values, they are simply used to prune the non-match pairs. An image pair with different values of the same soft biometric are treated as a non-match pair (e.g. an image of female and an image of male). If soft biometrics are in continuous values, they can be used to weight the similarity score from the face recognition algorithm. For a pair of two images, each of which yields a continuous value of certain soft biometric, these two values can be combined in certain way and the combined value can be treated as a weight of the similarity score computed from an existing face recognition algorithm.

Since a lot of the soft biometrics are directly or indirectly related to face images, the information they provide can not be seen as independent knowledge compared to what an existing face recognition algorithm already discovers. For instance, gender information is often implied in many face recognition methods. Current algorithms tend not to confuse a pair of female and male images.

This thesis tries to address 3 questions.

- Will using soft biometrics provide significant help in terms of improving the performance of existing face recognition algorithms?
- How much do soft biometrics help improve the performance?
- How much improvement can be achieved if the ground truth information of soft biometrics is available?

In order to answer these questions, I choose to work on a challenging data set, namely, the Good, the Bad and the Ugly (GBU) [22] data set. Two baseline algorithms have been developed to provide benchmark performance on this data set. The data set and these two baseline algorithms are introduced in the following sections.

## 1.4 The Good, the Bad and the Ugly Data Set

Kumar et al.'s work [14] on the Labeled Faces in the Wild (LFW) data set shows that using face attributes only can achieve comparable performance as the state of the art algorithms. Our first attempt of experiments also focuses on LFW data set. LFW provides two views. View 1 is for training and validation. People can do whatever they want on View 1. View 2 is for reporting performance. View 2 is randomly split into 10 sets. Each element in each set is a pair of images in View 2. However, the ground truth of the soft biometrics labels is not publicly available, which makes a supervised classification method difficult to implement.

Due to this reason, we switch to another data set created for face recognition by Phillips et al.. This data set is composed of frontal face images in various lighting and focus conditions, which makes it very challenging to recognize faces.

### 1.4.1 Description of the Data Set

This thesis uses a challenging data set, namely the Good, the Bad and the Ugly (GBU) [22] data set. It contains frontal face images in various lighting and focus conditions. GBU has three partitions, divided according to the difficulty. The difficulty of each partition is based on the performance of three top performers in the FRVT 2006 [23] Nominal performance evaluated by verification rate on the GBU is based on fusing the results from three top performers in the FRVT 2006. Verification is the process of a system declaring a person to be who they claim based upon the quality of match between a new face image of the person and a stored face image of the person. The Good partition contains pairs of images that are considered easy to recognize. On the Good partition, the base verification rate (VR) is 0.98 at a false accept rate (FAR) of 0.001. The Bad partition contains pairs of images of average difficulty to recognize. For the Bad partition, the VR is 0.80 at a FAR of 0.001. The Ugly partition contains pairs of images considered difficult to recognize, with a VR of 0.15 at a FAR of 0.001.

Each partition has a target set and a query set. Each of the target and query sets contains 1,085 images for 437 distinct people. The distribution of image counts per person in the target and query sets are 117 subjects with 1 image; 122 subjects with 2 images; 68 subjects with 3 images; and 130 subjects with 4 images. In each partition there is 3,297 match face pairs and 1,173,928 non-match face pairs.

In terms of soft biometrics, gender and race are two very important ones. They are relatively easy to classify and can provide useful information. In the GBU image set 58% of the subjects were male and 42% female; and 69% of the subjects were Caucasian, 22% east Asian, 4% Hispanic, and the remaining 5% other groups; and 94% of the subjects were between 18 and 30 years old with the remaining 6% over 30 years old. Figure 1.1 shows matching face pairs from each of the partitions.

### 1.4.2 Evaluation Method: Receiver Operating Characteristic

The performance of algorithms is evaluated via face verification result. Given a pair of two images, it is called a match pair if these two images come from the same person, and it is called a non-match pair if they come from different persons. The algorithm should predict whether these two images are a match pair or not. Then it is compared to the ground truth information. There are four possible outcomes from this evaluation method. If the outcome from a prediction is a match pair and the actual value is also a match pair, then it is called a true positive (TP); however if the actual value is a non-match pair then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are a non-match pair, and false negative (FN) is when the prediction outcome is a non-match pair while the actual value is a match pair.

A Receiver Operating Characteristic (ROC) curve [33] is defined by false positive rate (FPR) and true positive rate (TPR), where

$$FPR = \frac{\#FP}{\#FP + \#TN} \tag{1.1}$$

and

$$TPR = \frac{\#TP}{\#TP + \#FN} \tag{1.2}$$

# denotes the count. In the field of face recognition, FPR is also called false accept rate (FAR) and TPR is also called verification rate (VR).

Given a set of query images and a set of target images, a similarity matrix is computed using the face recognition algorithm. Each entry in this similarity matrix contains a similarity score of a specific pair of images, one coming from query and the other coming from target. If it is a distance matrix, we can simply reverse the sign of the score so that it is a similarity matrix. Higher scores indicate that the pair is more similar and more likely to be a match pair.

To determine whether or not a pair of images is a match pair, the whole similarity matrix is thresholded. Pairs of images whose similarity score exceeds the threshold is classified as a match pair and vice versa. The threshold slides from the minimum score to the maximum score, yielding all possible FAR and VR. Then an ROC curve can be plotted where the horizontal axis is FAR and the vertical axis is VR. A good face recognition algorithm should have high VR values with low FAR values. Beside the whole ROC curve, people also interested in some specific points, such as VRs at FAR $= 0.001, 0.01, 0.1$.

Figure 1.1: Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

## 1.5 GBU Baseline Algorithms

### 1.5.1 Local Region PCA

Local Region PCA (LRPCA) [22] is a refined implementation of the standard method based on principle component analysis (PCA) face recognition algorithm, also known as Eigenfaces [28]. It first extracts a cropped and geometrically normalized face region from an original face image. The original image is assumed to be a still image whose pose of the face is close to frontal. The face region in the original is scaled, rotated, and cropped to a specified size and the centers of the eyes are horizontally aligned and placed on standard pixel locations. In the baseline algorithm, the face chip is 128 by 128 pixels with the centers of the eyes spaced 64 pixels apart.

The PCA representation is conducted on thirteen local regions cropped out of a normalized face image and the complete face chip. The local regions are centered relative to the average location of the eyes, eyebrows, nose and mouth. Figure 1.2 shows a cropped face and the thirteen local regions. All the 14 face regions are normalized to attenuate variation in illumination. First, self quotient normalization is independently applied to each of the 14 regions [30]. The self quotient normalization procedure first smooths each region by convolving it with a two-dimensional Gaussian kernel and then divides the original region by the smoothed region. The influence of illumination can be reduced by self quotient normalization. A final normalization step adjusts the pixel values in each region to have a sample mean of zero and a sample standard deviation of one.

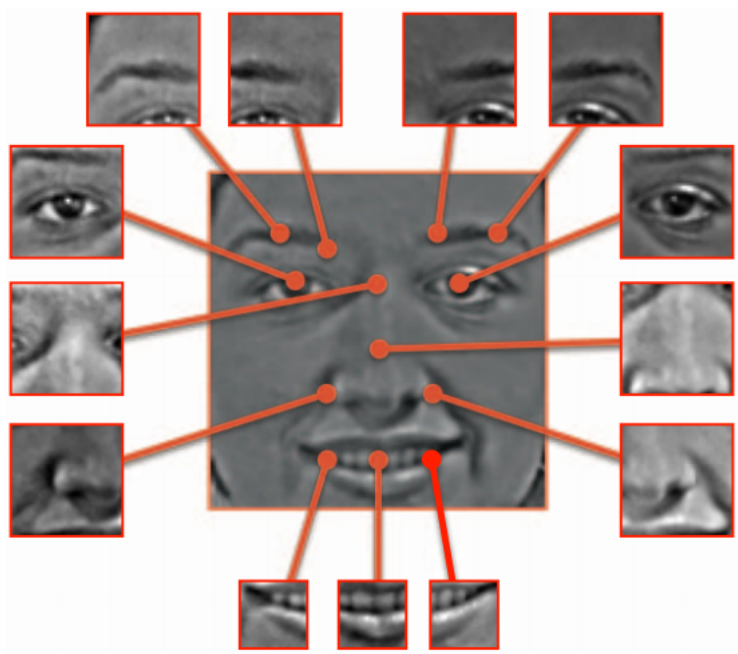

Figure 1.2: A cropped face and the thirteen local regions.

During training, PCA is computed for each of the 14 regions and the 3rd through 252th eigenvectors are retained to represent the face. As a result, a face is encoded by concatenating the 250 coefficients for each of the 14 regions into a new vector of length

3500. The representation is whitened by scaling each dimension to have a sample standard deviation of one on the training set. Then the weight on each dimension is further adjusted based on Fishers criterion. The Fishers criterion weight emphasizes the dimensions along which images of different people are spread apart according to the training set. During testing, coefficients computed from PCA projection of each of the 14 regions are concatenated as a vector. Each image corresponds to one vector. Similarity between pairs of faces is measured by computing the Pearsons correlation coefcient between pairs of these vectors.

### 1.5.2 Cohort Linear Discriminant Analysis

Cohort Linear Discriminant Analysis (CohortLDA) [16] is a Linear Discriminant Analysis (LDA) algorithm with color spaces and cohort normalization. The main differences between CohortLDA and standard LDA are two-fold. One is the preprocessing step and the other is that CohortLDA introduces a cohort set to normalize the score. Specifically, CohortLDA uses both the R channel from RGB color space and the I channel from YIQ color space to conserve the structure of the face and reduce the influence of strong illumination. Since the red channel is similar to the gray-scale image, it usually does not work well with large lighting variation. Therefore, logarithm transformation and z-normalization are applied after extracting the R channel. During training, it seeks a projection that maximizes the ratio of between-class scatter and within-class scatter in order to make the data belonging to the same cluster more similar and the data belonging to different clusters more different. Figure 1.3 presents the LDA faces computed from the red channel and I chrominance.

In face verification, a decision threshold is needed to determine whether a pair of faces is a match or not. Since some images are harder than others, a fixed threshold may not adapt well from image to image. As a result, a set of images called the cohort set is adopted to adjust the match distance. During face verification, a subset

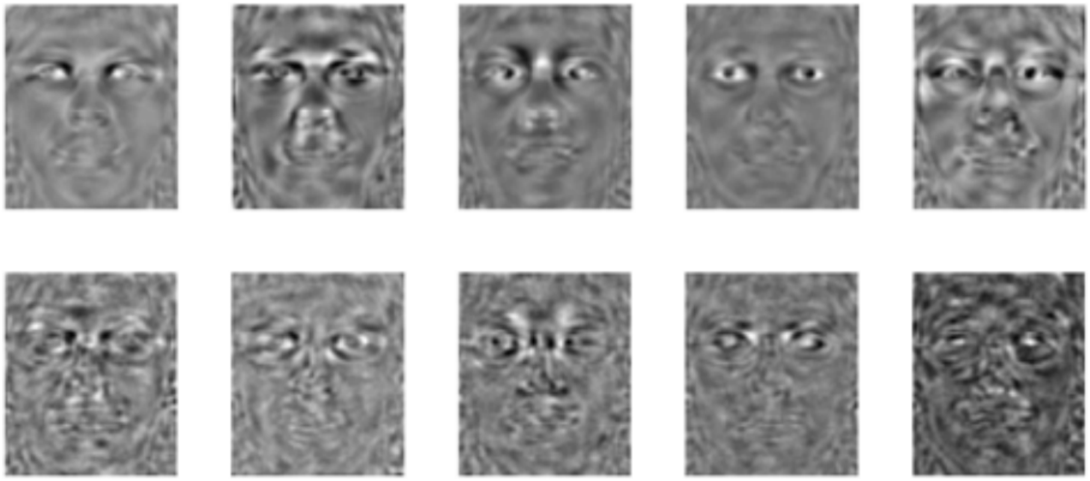

Figure 1.3: Top row: LDA faces acquired from the red channel after log and z-norm. Bottom row: LDA faces obtained from the I chrominance after znorm.

of cohort images is selected for each query image and target image using the k nearest neighbor rule. Then their average distance to the k neighbors is computed as an indication of the difficulty of the query/target image. The distance of the query and target image is calculated as their original distance subtracted by their difficulty. The final distance is obtained based on the red channel and I chrominance images using a simple sum rule.

This paper presents four studies meant to explore the potential of soft biometrics to improve face recognition performance. The first is a best-case analysis. If we have perfect information about a categorical soft biometric, in this case ground truth information about gender and race, how much does it improve face recognition performance on challenging data sets? The second looks at the same question in a more realistic scenario: how much do gender and race improve performance when they must be inferred from images via an imperfect classifier? The third study looks at the effectiveness of a non-categorical facial feature, facial geometry, as a soft biometric. Finally, the fourth study looks at adding non-facial information by analyzing the image region just outside the face – a region containing hair, ears and the neck – as a soft biometric.

Although by no means exhaustive, these studies (along with a couple of studies from the literature; see below) lead to a common conclusion. They suggest that soft biometrics can produce small increases in recognition performance on challenging data sets, particularly when used as soft weights rather than hard constraints. The small size of the performance gain, however, suggests that much of the information in soft biometrics is either already exploited by traditional face recognition algorithms or else redundant with other information in the face. Otherwise, we would expect the performance gains from integrating new and independent sources of information to be larger.

# Chapter 2

# Related Work

Google Scholar attests to the level of interest in soft biometrics. At the time of writing, a search for the terms *+"soft biometrics"* and *+face* produced 170 articles since the beginning of 2011. Although some of these papers only mention soft biometrics as a future research topic, most of the rest address how to automatically extract soft biometric features from face images, in particular gender, age and race, and focus on how to integrate the information of these soft biometrics into a face recognition algorithm. (See Lyle [17] and Shan [26] for tables comparing classification accuracies.) Researchers adopt different ways to automatically extract soft biometric features. Some use a universal framework and some use a specific algorithm for each specific soft biometric.

## 2.1 Automatic Estimation of Soft Biometrics

Gutta et al. [9] presented a mixture of experts for the classification of gender, ethnic origin, and pose of human faces. Ensembles of radial basis functions (RBFs) compose the mixture of experts. Besides the ensembles of RBFs, they also used a SVM classifier with RBF kernel for gating the inputs. Their gender classifier achieved an average accuracy rate of 96%, their ethnicity classifier yielded an average accuracy of 92% and their SVM classification rate on pose is 100%. These results on gender and ethnicity were reported on good quality face images from the FERET database where there are little the expression and pose variations.

Jain and Lu [15] proposed a Linear Discriminant Analysis (LDA) based scheme to tackle the two-class (Asian vs. non-Asian) ethnicity classification problem. Multiscale analysis and an ensemble framework based on the produce rule were adopted to enhance the classification performance. This scheme had an accuracy of 96.3% on a database of 263 subjects, each of which has 10 images (with equal balance between Asian and non-Asian classes).

Lin et. al. proposed a novel approach for recognizing the gender, ethnicity and age using face images. Their approach combined Gabor filter, Adabost learning and SVM classification. Facial features are extracted via banks of Gabor filters and Adaboost learning. Then SVM classifiers based on the features are trained to recognize soft biometrics. Their experimental results on FERET data set showed good performance. They also showed that a preprocessing step can further improve the performance.

Kumar et al.'s work [14], which is closely related to ours and will be elaborated later, used a large set of low level features to train an SVM classifier with RBF kernels to extract soft biometrics. A few low level features from this set which are most related to the specific soft biometric are picked as the input of that specific SVM classifier associated with that soft biometric. Reasonable classification rates are reported for the attributes they are classifying.

Lyle et al. [17] compute periocular texture from grayscale images using Local Binary Patterns. Then an SVM classifier is trained to classify the texture features. They conduct their experiments on the visible spectrum periocular images obtained from the FRGC face dataset. For 4232 periocular images of 404 subjects, they achieved a baseline gender and ethnicity classification accuracy of 93% and 91% in cross validation. They also showed that by fusing the periocular soft biometric data from both eyes with LBP over the face region, periocular recognition can be improved.

## 2.2 Integrating Soft Biometrics to Improve Recognition Performance

More relevant to this paper are efforts that use soft biometric features to improve recognition performance in challenging data sets. Besides different ways to extract soft biometrics, there are also different ways to fuse the information of soft biometrics into an existing face recognition algorithm. One is to use soft biometrics to prune the search space [7] [10]. Another is to use the information of soft biometrics as a weight to be added to or multiplied by the scores of an existing face recognition algorithm. [25] [11] [12] [13]. Moreover, a face recognition system can be built using only soft biometrics without any other face recognition alrogithm and still achieve good performance [14].

Jain et.al. [11] [12] explored the question on whether soft biometric traits can assist user recognition. They divided their recognition system into two subsystems. The first one is traditional biometric identiers called primary biometric system and the second is based on soft biometric traits called secondary biometric system. They formulate the recognition process from the perspective of probability conditioned on features corresponding to these two subsystems. Moreover, they also introduced a weighting of the two subsystems and a weighting of different soft biometric traits to pay more attention to the important features/soft biometrics. Preliminary experiments conducted on a ngerprint database of 160 users by synthetically generating soft biometric traits showed that the use of additional soft biometric user information signicantly improves ( 6%) the recognition performance. Their later experiments conducted on a database of 263 users show that the recognition performance of a ngerprint system can be improved significantly (about 5%) by using additional user information like gender, ethnicity, and height.

Jain and Park [13] use local facial marks such as freckles, moles and scars to improve the rank-1 face identification rate of a traditional face recognition algorithm

(FaceVACS). Active Appearance Model (AAM) was used to locate facial landmarks. Then they adopted Laplacian-of-Gaussian and morphological operators to detect facial marks. On the FERET data set, they were able to increase the recognition rate by about 1%. On the Mugshot data set, the improvement of recognition rate is 1.26%.

Dantcheva et al. [6] looked at eye color in the visible spectrum as a soft biometric. They examined the influence of illumination, presence of glasses and color perception of left and right eye. An automatic eye color detection system was built using iris localization and classification based on Gaussian Mixture Models (GMM) with Expectation Maximization (EM). It should be noted that 90% of the population have brown irises.

A successful example of applying soft biometrics is the work of Kumar et al. [14]. Their work gives me the most inspiration. I implemented a similar method to conduct general classification of soft biometrics. I also used the distance to the hyperplane as one of a few ways to estimate the soft biometrics.

In their paper, they introduce the use of describable visual attributes, which are labels that can be given to an image to describe its appearance, for face verification and image search. A part of these describable visual attributes is composed of soft biometrics such as gender, race and so forth. They used Amazon Turk, *http://mturk.com.*, to collect labels of thousands of images easily and with very low overhead.

Given an image, a pool of different low-level features are extracted according to local regions on a face, pixel representation methods, normalization methods and aggregation methods. Forward feature selection is applied to select a group of features tuned to the classification task of the specific attribute from the pool. SVM with an RBF kernel is adopted for the task. A grid search of the SVM parameters is performed, which is quite time-consuming. The classifiers learned from this training procedure is called attribute classifiers.

Besides attribute classifiers, they also construct another kind of classifiers named simile classifiers. They measure the similarity of a part of a person's face to the same part on a set of reference people. Support vector machines are trained for each reference person to distinguish a region (e.g., eyebrows, eyes, nose, and mouth) on their face from the same region on other faces. A set of possible features are manually selected and classifiers are trained for each reference person/region/feature type combination, without feature selection.

In order to decide whether two face images, $I_1$ and $I_2$ show the same person, one can train a verification classifier that compares attribute vectors and return the verification decision. Each image's distance to the SVM hyper-plane from attribute classifiers and simile classifiers is stored. Let $a_i = C_i(I_1)$ and $b_i = C_i(I_2)$ be the outputs of the $i$th trait classifier for each face ($1 \leq i \leq n$). The absolute difference of the distances and the product of the distances are considered important. Putting both terms together generates the tuple $p_i$:

$$p_i =< |a_i - b_i|, a_i b_i > \tag{2.1}$$

The concatenation of these tuples for all n attribute/simile classifier outputs forms the input to the verification classifier. Pairs of face images from the same person form positive examples and that from different persons form negative examples. The verification classifier is also SVM with an RBF kernel. The authors train this classifier with default parameters.

Although they avoid the term soft biometrics, Kumar et al. [14] use crowd sourcing to collect soft-biometric-like attribute labels for face images from the LFW data set, including gender, age, hair color, hair line, nose shape, face shape and attractiveness. They also collect labels for attributes that are clearly not soft biometrics because they are image specific, for example lighting conditions, quality of focus and whether the mouth is open or closed. They were able to generate impressive recognition results; unfortunately, it is difficult to know why, since they combined soft biometric and image-specific attributes and applied them to a data set where there are known correlations between imaging conditions and subject identity.

Scheirer et al.'s work [25] shares the same spirit of mine, i.e., improving the performance of baseline algorithms using soft biometrics. Two main differences exist between their work and my work. One is that they used contextual information such as occupation and places a person lives. The other is that they built a Bayesian network to combine attributes and produce a weight factor while I used normalized SVM scores as a weight factor.

They extend Kumar et al.'s work by introducing a Bayesian approach to combining descriptive attributes and producing accurate weighting factors to apply to match scores from face recognition algorithms based on incomplete observations made at match time. Specifically, given a gallery image with its corresponding attribute network and a probe image with observed attributes, the probability of them coming from the same person can be computed via a fast Noisy-OR formulation for streamlined truth value assignment and more accurate weighting.

They used the same classifiers constructed in the work of Kumar et al. [14], which extract low level features to build SVM classifiers for soft biometrics. Besides Kumar et. al.'s work, they also adopted a robust age estimation approach from Chen et al. [3] that has demonstrated the best performance.

They built attribute network enrollment records for each person in the gallery based on a list of visual attributes, and a list of contextual attributes. During matching, observations from the probe image are matched to each stored attribute network for each gallery entry. A weighting factor is generated after solving each network. This weighting factor is then applied to the original match score from baseline algorithms to improve the verification rate.

Their experiments on MBGC [20] data set yield an improvement over the baseline algorithm. It shows that incorporating descriptive attributes into the matching process significantly enhances face identification over the baseline by up to 32.8%.

However, the descriptive attributes they use include a person's occupation, places they live and so forth. These cannot be extracted from the image itself. It sparks a question on how well soft biometrics alone can do to improve the performance of face recognition algorithms.

Although Kumar et.al and Scheirer et al.'s work showed good improvement in performance when soft biometrics are integrated with additional appearance and contextual attributes, they did not present results of using only soft biometrics related to faces. My work will be focusing on using only face related soft biometrics to improve existing face recognition algorithms. Results on the Good, the Bad and the Ugly data set are presented in this paper indicating that using only these soft biometrics does not help much in terms of improving the face recognition performance.

# Chapter 3

# Using soft biometrics derived from images

## 3.1 Learning Soft Biometrics

In this section, I present a general method for estimating soft biometrics and then explore the effectiveness of incorporating soft biometrics into the recognize process.

Our approach here is inspired by the work of Kumar et al. [14], although for several reasons that will become apparent, ours is more limited in scope. Kumar et al. [14] defined a set of image features *a priori* along with 73 attributes. Some of these attributes describe properties of the face and are traditionally thought of as soft biometrics. Others pertain to imaging conditions, for example harsh versus soft lighting. In all cases, extensive hand labeled data is combined with a two stage SVM learning procedure to ultimately create a face recognition algorithm. A striking aspect of the work is that impressive face recognition results are demonstrated using the Labeled Faces in the Wild dataset and a classifier based upon the 73 attributes.

A precise adaptation of the 73 attribute algorithm to the GBU Challenge Problem is neither feasible nor even relevant to the topic of soft biometrics. It is not feasible due to a lack of detailed hand generated training data. It is not relevant in so much as attributes such as harsh versus soft lighting, whether a person's teeth are visible, or their mouth is open are not soft biometrics. I am interested in exploring the performance gain of using only face related soft biometrics, rather than the gains associated with lighting conditions. Therefore, the full set of attributes used in Kumar et. al.'s work are not adopted in my work. To be clear, Kumar et al. [14] never claimed that their attributes were all soft biometrics.

### 3.1.1 Low-Level Feature Extraction

We have implemented a reduced attribute system for learning soft biometrics. It begins with a four step procedure to extract low-level features. A complete feature type is constructed by first choosing a local face region. Then the pixel values in that region is converted to one of the pixel value types from RGB, HSV, image intensity, edge magnitude and edge orientation. Then mean normalization, or energy normalization, may be applied. Finally, there is an option to use histograms to aggregate the values from the previous steps. Local face regions are detected using Stasm [18]. Stasm is an extension to the Active Shape Model [4]. It is initialized using the eye coordinates provided with the GBU dataset. Figure 3.1 shows an example of fiducial points detection using Stasm. Figure 3.2 shows the local regions extracted from a face. A summary of options to extract low-level features is presented in Table 3.1.

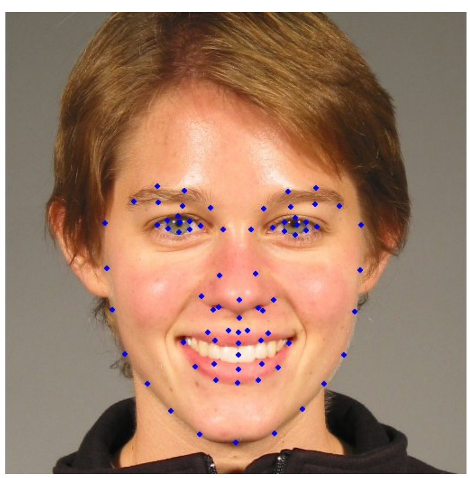

Figure 3.1: Stasm detection

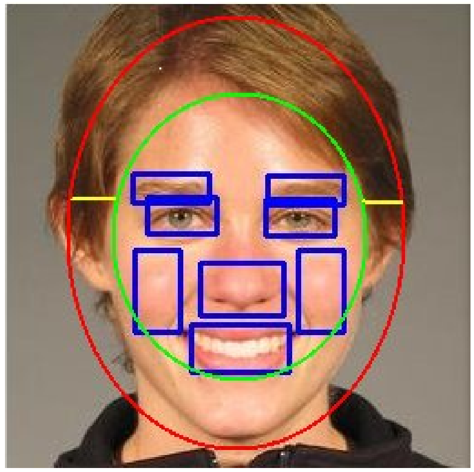

Figure 3.2: Extracted local regions

| Pixel Types | Normalization | Aggregation |
|---|---|---|
| RGB<br>HSV<br>Image Intensity<br>Edge Magnitude<br>Edge Orientation | None<br>Mean Normalization<br>Energy Normalization | None<br>Histogram |

Table 3.1: **Feature type options**

After low level feature extraction, I adopt Support Vector Machine (SVM) introduced by Vapnik [29] as my classifier to estimate soft biometrics. The description of SVM is presented below.

### 3.1.2 Classification Method for Soft Biometrics: Support Vector Machine

Given a set of $n$ input vectors $\mathbf{x}$ and outputs $y_i \in \{-1, +1\}$, one tries to find a weight vector $\mathbf{w}$ and offset $b$ defining a hyperplane that maximally separates the examples. This can be formalized as the minimization problem

$$\min_{\mathbf{w},b} \frac{||\mathbf{w}||^2}{2} \quad \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1, \ for\ \forall i \tag{3.1}$$

Using convex optimization theorem [2], it turns out that the solution for $\mathbf{w}$ can be written in the form

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

where the coefficients $\alpha_i$ are non-negative. The $\mathbf{x}_i$ with $\alpha_i > 0$ are called *support vectors*.

For more general SVMs one can consider kernels which implicitly map the $\mathbf{x}$ into a high-dimensional space called the feature space; $\mathbf{w}$ and $b$ then define a hyperplane in this space. There are a lot of kernels available, such as polynomial, Gaussian radial basis function, hyperbolic tangent and so forth. Kernels are usually chosen via experimental results or empirical experience. Moreover, if the data are not linearly separable – even in feature space – one needs to relax the constraints in eq. (3.1) then has to tune an extra parameter $C$ which controls how "soft" the constraint is made.

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \{ \frac{||\mathbf{w}||^2}{2} + C \sum_{i=1}^{n} \xi_i \} \\ &\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1 - \xi_i,\ \xi_i > 0\ for\ \forall i \end{aligned} \tag{3.2}$$

### 3.1.3 SVM Classification Results on Soft Biometrics

To augment the gender, age and race information already available for GBU, we hand labeled additional soft biometrics, specifically hair color (black/other), eyebrows (thin/other), head shape (chubby/thin) and eye color (dark/other). Race was reduced here to either white or other. Then for each of these soft biometrics and different combinations of low-level features as summarized in Table 3.1 an SVM is trained and evaluated using cross validation. Features that performed best were then manually chosen as the basis for or a small set of classifiers for each soft biometric. When multiple classifiers were selected, the final label is determined by a vote.

To avoid testing and training on the same people, the standard GBU training set, *GBU_Train_Uncontrolledx8*, was used throughout the soft biometric learning procedure. Not all of the soft biometrics initially considered proved useful. For example, many subjects in GBU are college students and of approximately equivalent age.

| | Train/Test | Good | Ugly |
|---|---|---|---|
| gender | 91.2% | 86.6% | 84.3% |
| race | 90.5% | 89.1% | 84.2% |
| eye color | 71.6% | 78.2% | 68.0% |
| hair color | 83.7% | 76.3% | 76.6% |
| eyebrow | 74.2% | 70.5% | 66.1% |

Table 3.2: SVM classification results for soft biometrics in cross validation training/test as well as the good and ugly GBU partitions.

Hence classification on age is trivial but not helpful. Table 3.2 provides classification results for the final SVMs selected. Columns indicate performance on the cross validation tests over the training data, followed by the good and ugly GBU partitions.

### 3.1.4 A comparison on Gender Classification: SVM vs. OpenCV Classification Using Fisherfaces

In order to show that our SVM classification method achieves reasonable results, I use the code available on OpenCV website[1] that is used to do gender classification and compare its result with my SVM result. The code used Fisherfaces [1] to tackle this two-class problem. It aimed to find a linear projection that minimizes the within-class variation in male and female classes and maximizes the inter-class variation between male and female in the meantime. The code was originally tested on a very small data set, which had only 8 male and 5 female subjects, each with 10 images. The cross validation performance on this small data set was very good, achieving a 98% recognition rate in a subject-independent cross-validation. A subject-independent cross-validation means images of the person under test are never used for learning the model.

[1]http://docs.opencv.org/trunk/modules/contrib/doc/facerec/tutorial/facerec_gender_classification.html#running-the-demo

| | Good | Ugly |
|---|---|---|
| gender | 79.7% | 74.0% |

Table 3.3: OpenCV gender classification on the good and ugly GBU partitions.

Given training and test data, the code can be separated into two parts. One is to learn Fisherfaces on the training data, and the other is to compute a score of test images after projection using Fisherfaces. I still use the standard GBU training set, *GBU_Train_Uncontrolledx8*, to learn Fisherfaces, and the test images are the Good and Ugly partitions. The input data are 200x200 face chips after careful alignment using ground truth eye coordinates. Table 3.3

It can be easily observed that SVM classification method is performing quite well compared to an open source gender classifier. Although there is no demonstration that the SVM classification result is the state of the art, its performance is reasonable and comparable. Moreover, this SVM classifier is a very general method, which can be also used to estimate race, eye color, hair color, etc.

### 3.1.5 Integrating Soft Biometrics in Existing Algorithms: A Soft Weighting Scheme

As discussed in Chapter , there are different ways to integrate soft biometrics given the result (a similarity/distance matrix) of an existing face recognition algorithm. I present results of using soft biometrics either as a weighting scheme to fuse their information with the similarity/distance matrix in this section and as a pruning method in the next section. Since a distance matrix can be easily converted to a similarity matrix, I will only mention similarity matrix from now on.

It is natural to think of soft biometrics as discrete values, such as male/female and Asian/Caucasian. However, assigning hard labels for these soft biometrics may not help identifying the person. The reasons are two-fold. One is that automatic classification methods make mistakes. The other is that even if a pair of face images share the same label of a soft biometric, the presence of that soft biometric may vary between them. For instance, one man can look more manly than another man. In cases like this, it would be more suitable to assign a score of that soft biometric indicating the degree of its presence, which could provide much richer information. Therefore, I adopt the distance to the SVM hyperplane as a score of the soft biometric and present the results. The results of using hard labels are described in the next section.

The procedure is described as follows. For each soft biometric, we store each image's distance to the hyperplane from each of the SVM classifiers corresponding to different features for a soft biometric. To combine these distances into a single one, we normalize each distance using the standard deviation of all the distances from the corresponding SVM and the distances after normalization are added together as a final distance. This is used to combine the features/distances from the same soft biometric.

Given a pair of images, we compute the absolute difference of their distance as a score representing how far these two images are in terms of the soft biometric. In order to combine this score, denoted as $s$, with similarity matrices, we turn this distance measure into a similarity measure using $s_{max} - s$, where $s_{max}$ is the maximum score of all image pairs. This similarity measure is then used to weight, through multiplication, similarity scores obtained by a face recognition algorithm.

Tables 3.4a and 3.4b show the verification rate at FAR=0.001 using these soft biometrics as weights combined with the three algorithms introduced earlier. The "no SB" column of each table shows the verification rate of the algorithm without any additional weighting by soft biometrics. The last column shows the verification rate when combining all these soft biometrics together as a weight matrix. Gender in most cases improves performance, although less then when ground truth gender information is used: Chapter 4. The other soft biometrics make little improvement

| | no SB | gender | race | eye color | hair color | eyebrow | all |
|---|---|---|---|---|---|---|---|
| LRPCA | 70.9% | 73.0% | 72.2% | 70.7% | 71.3% | 71.5% | 72.4% |
| CohortLDA | 84.1% | 84.2% | 84.2% | 82.6% | 82.2% | 82.3% | 83.8% |
| Fusion | 98.4% | 98.4% | 98.2% | 98.0% | 98.1% | 98.2% | 98.4% |

(a) Good partition

| | no SB | gender | race | eye color | hair color | eyebrow | all |
|---|---|---|---|---|---|---|---|
| LRPCA | 8.3% | 8.8% | 8.1% | 7.7% | 8.3% | 8.1% | 8.0% |
| CohortLDA | 11.4% | 12.0% | 10.7% | 10.9% | 11.5% | 12.2% | 11.7% |
| Fusion | 15.2% | 16.4% | 13.7% | 15.3% | 16.5% | 16.2% | 15.5% |

( b) Ugly partition

Table 3.4: Weighting using soft biometrics on the Ugly partition

or even degrade performance. Furthermore, even when all the soft biometrics are combined, the net change in verification is neither of practical significance nor even always positive.

### 3.1.6 Integrating Soft Biometrics in Existing Algorithms: A Hard Pruning Method

Besides using the distance to the hyperplane in the SVM as a weight, I also used a hard classification as a way to prune the similarity matrix. A pair of images with different discrete values for the same soft biometric can be marked as a non-match pair immediately without further computation. It not only reduces the computation burden, but also is the most suitable way to prune the similarity matrix when the ground truth labels are known. Of course, when it comes to an automatic classification method for the soft biometrics, the risks would be high discarding non-match pairs since it can mark a match pair as a non-match pair when it makes a classification mistake on one of the images. Since there are much more non-match pairs than match pairs, it would even hurt the performance if many match pairs are denoted as non-match pairs.

I pick the two most promising soft biometrics, gender and race, and conduct the following experiment. For a pair of images, SVM classification is applied based on the

| | No SB | G | R | G & R |
|---|---|---|---|---|
| LRPCA | 70.9% | 61.9% | 58.1% | 59.4% |
| CohortLDA | 84.1% | 71.4% | 65.0% | 66.2% |
| Fusion | 98.4% | 81.4% | 73.7% | 74.9% |

(a) Good partition

| | No SB | G | R | G & R |
|---|---|---|---|---|
| LRPCA | 8.3% | 7.6% | 6.4% | 6.6% |
| CohortLDA | 11.4% | 10.6% | 9.1% | 9.6% |
| Fusion | 15.2% | 15.7% | 12.8% | 12.6% |

(b) Ugly partition

Table 3.5: Verification rates with learned soft biometrics.

corresponding low-level features. Then I check the hard classification labels of them. If the labels are of different values, then this pair is regarded as a non-match pair and is pruned from the similarity matrix, meaning that its similarity score is set to -1000 (-1000 denotes a infinitely small score). The verification rate is then computed on the pruned similarity matrix.

Experimental results are shown in Table 3.5a and 3.5b. Except for the case that pruning using gender on the Ugly partition improves the performance of Fusion by 0.5%, performance in all other cases is hurt by this brute force pruning method. An important lesson is evident in Table 3.5. Making hard decisions, in other words pruning options, based on imperfect estimates of a soft biometric can easily do more harm than good.

## 3.2 Geometry

This experiment was designed to test another continuous (rather than categorical) soft biometric extracted from the face region. I was curious whether all the information in the face was already accounted for by the existing face recognition algorithms (particularly the fusion algorithm), or whether a new feature could provide non-redundant information. After piloting several possible features, I settled on the geometry of facial landmarks as the most promising. These landmarks do not vary much within the same face. Moreover, they tend to differ a lot between faces from different subjects. Therefore, I think they are a good feature to discriminate faces.

Face shape is not a new feature in the field of face recognition. [31] [5] [8] have already used this kind of features to identify faces. However, two issues arise that make me realize that face shape should be another good soft biometric to aid the baseline face recognition algorithms. One is that landmark localization was hard due to variations of poses and expressions. However, since GBU is only composed of frontal images, more accurate facial landmark localization should be expected. The other is that the two baseline algorithms, namely LRPCA and CohortLDA, do not explicitly explore by face shape. So it should introduce new and useful information when we add face shape to these algorithms. It can be expected that adding face shape information to face recognition algorithms which already take into consideration the locations of facial landmarks would not help much.

In particular, Stasm [18] is used to localize the landmarks on a face, ignoring landmarks around the mouth since they are too sensitive to changes in expression. To compare two face images, we use Stasm to find the landmark positions in each image, translate the landmarks so that they are centered around the origin, and then subtract the positions of the landmarks in the first image from the positions of the corresponding landmarks in the second image. The result is a vector of relative landmark displacements. An SVM is trained to distinguish displacement vectors generated from matching pairs from displacement vectors generated from mismatched pairs.

Some methodological details follow. Since there exist many more non-match pairs than match pairs, using them all would produce an unbalanced training set. To provide SVM training with a balanced input, bootstrapping is applied. We use all the match pairs and randomly sample the same number of non-match pairs. This is done

| | Good | Ugly |
|---|---|---|
| LRPCA | 70.9% | 8.3% |
| LRPCA + Geometry | 73.2% | 9.5% |
| CohortLDA | 84.1% | 11.4% |
| CohortLDA + Geometry | 83.0% | 13.4% |
| Fusion | 98.4% | 15.2% |
| Fusion + Geometry | 97.9% | 18.2% |

(a)

| | Good | Ugly |
|---|---|---|
| LRPCA w&w/o Geometry | 4.9e-8 | 2.4e-4 |
| CohortLDA w&w/o Geometry | 1.3e-2 | 1.5e-7 |
| Fusion w&w/o Geometry | 2.0e-2 | 2.9e-11 |

(b)

Table 3.6: Comparing performance with and without face geometry added to an existing algorithm. a) verification rates with and without additional constraint from face geometry, b) p-values for McNemar's testing statistical significance of improvement.

several times, each of which produces an SVM classifier with different parameters. Cross-validation is used to determine when to stop training the SVM, with the restriction that no subject can appear in both the training set and the validation set to avoid using person-affiliated information.

The kernel of our SVM is a radial basis function (RBF) and the parameters are found via grid search. Once training is complete, each SVM votes to decide whether a novel pair of test images is a match or non-match. Since some classifiers perform better in cross validation than others, there are some "wise" voters and some "not-so-wise" voters. Therefore, we weight votes by the classifier's cross validation score. The overall decision is the sum of the weighted votes.

Table 3.6a shows the verification rate of the 3 face recognition algorithms and their verification rates after weighting using the geometric soft biometric score information on GBU. As with the earlier experiments, adding the facial landmark soft biometric improved performance on the Ugly partition, but not dramatically. The performance of the fusion algorithm improved by 3%, while the performance of the other two algorithms improved slightly less. Performance on the Good data set actually dropped slightly for Fusion and CohortLDA, although in the case of the Fusion algorithm the drop was only 0.5%.

While not necessarily an improvement of great practical note, face shape as expressed through geometry yields the greatest gain of any measured soft biometric so far considered for the Ugly partition. In order to judge the statistical significance of adding the information provided by geometry, McNemar's test [32] is carried out and the resulting p-values are shown in Table 3.6b. This testing shows all differences are statistically significant, with the improvements on the Ugly partition unquestionably so.

## 3.3 Non-face Biometrics: Halo PCA

This study was inspired by an experimental result about human face recognition reported by O'Toole et. al. [24]. They showed pairs of face images from the ugly partition of GBU to human observers and asked them whether the faces matched or not, under two conditions. In the first condition, subjects were shown the oval of the face, but the image beyond the face oval was grayed out. This had the effect of blocking out not only the background but also the subject's hair, ears and neck. In the second condition, it was the oval of the face that was grayed out while everything else was preserved. Surprisingly, the human observers were about as accurate at matching faces when the face was obscured as when the face was present but everything else was obscured. This suggested to us that we should look for soft biometrics outside the oval of the face.

Traditional face recognition systems such as PCA [28] LDA [1] always pay much attention to the inner region of a face (face oval). It is true that humans can do a good job identifying persons well using only the face oval. There is no doubt that eyes, noses, mouths are important to the recognition of a face. However, Sinha et.

al. [27] pointed out that processing performed by the human vision system to judge identity is better characterized as head recognition rather than face recognition. It suggested that there is useful information in the region outside of the face oval and inside of the head region. It is common sense that different people have different hair styles and face contours and they do not change much within a period of time. Considering this cue may help with the recognition process. The background outside of the head region is often discarded since it varies a lot very quickly.

Another reason of looking at the region between head and face oval is that face recognition problem can become more challenging due to insufficient lighting, poor focus and so on, leading to the eyes, noses and mouths difficult to recognize. Therefore, using the inner region of the face to recognize subjects is not reliable. In cases like this, hair style and face contours come to be a more robust feature to rely on.

To construct a non-face soft biometric that was not dominated by useless background information, we extracted a "halo" which is a curved band just outside the face oval, as shown in Figure 3.3. Such halos show the hair and ears of the subject as well as part of the neck. The halo excludes the face and most of the background. This halo shaped region around the head does not vary a lot under these poor conditions. Moreover, it varies among different subjects, meaning that this can be a good feature to distinguish persons.

We then explored how discriminative such halos are. In particular, we adopted Principle Component Analysis (PCA) as our framework. Instead of using inner faces to generate eigenfaces, we used halo images to generate eigenhalos. Given a new face image, its halo was extracted and projected using eigenhalos. The distance between a pair of face images is defined as the angle between them after projection. Table 3.7 shows the verification rate of eigenhalos at a false accept rate (FAR) of 0.001 for the Good and Ugly partitions.

It is surprising that eigenhalos do as well as 4.4% on the Ugly partition, where

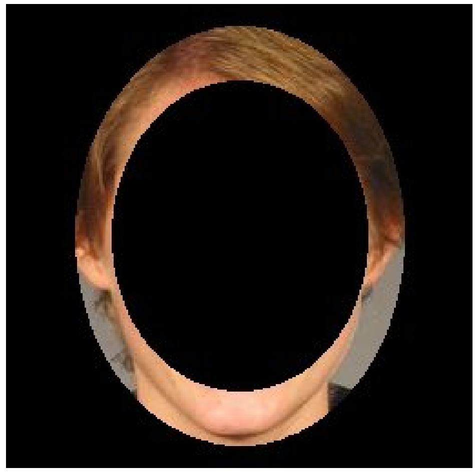

Figure 3.3: An example of the halo image

| | Good | Ugly |
|---|---|---|
| Halo PCA | 38.2% | 4.4% |

Table 3.7: Verification rate of Eigenhalos on the Good and Ugly partitions.

the best performance so far is 15.2%, given by the fusion of three commercial face recognition systems. After all, it is a simple feature *derived without reference to the face.* Nonetheless, the goal is not to build another face recognition algorithm, but to determine if soft biometric information from outside the face can be fused with traditional face recognition algorithms to improve performance. To determine the best weighting of eigenhalo similarity vs. face similarity, we plotted recognition as a function of relative weight, as shown in Figure 3.4 for the good partition and Figure 3.5 for the ugly partition.

We draw three conclusions from this experiment.

- There is a benefit to using the halo information as a soft biometric. Fusing HaloPCA with a face recognition algorithm almost always improves the verification rate of the original algorithm, as long as the weight on HaloPCA is small. (One exception is Fusion+HaloPCA on the good partitiion.)

- The benefits of the halo are small, however, particularly for the Fusion algo-

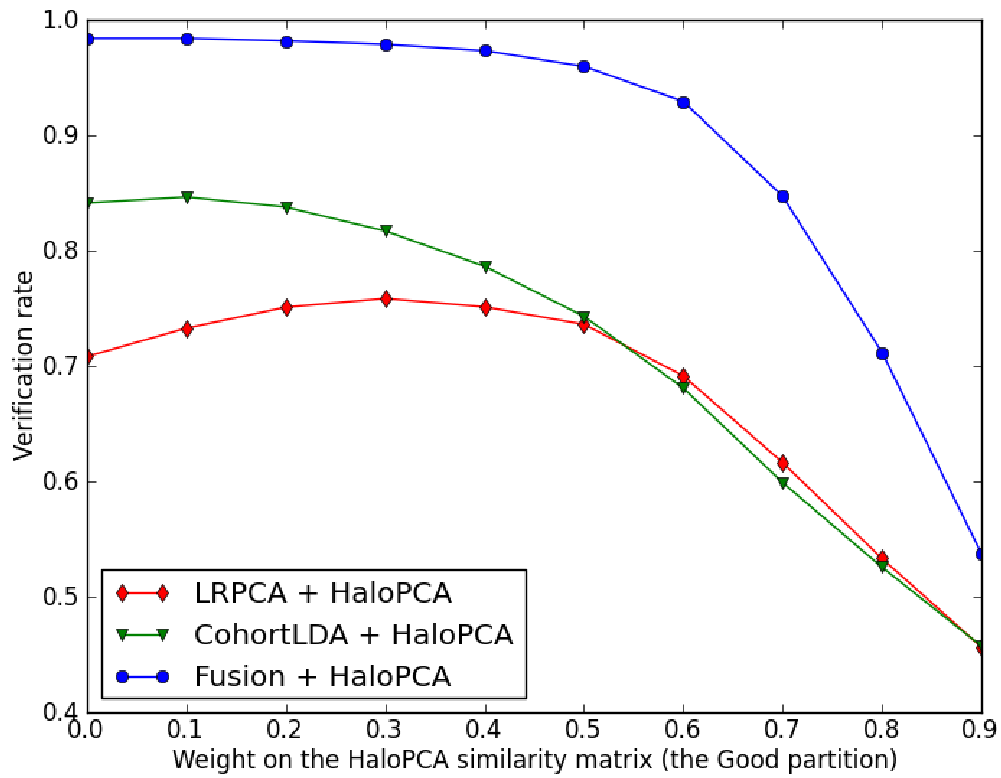


Figure 3.4: Fusing HaloPCA on Good

rithm, which is the best face recognition algorihtm tested.

- There is no "universal" weight for fusing HaloPCA with other algorithms. The weight depends on the algorithm and the data set.

## 3.4 Conclusion

In this chapter, I describe the SVM classification method to estimate face related soft biometrics based on low-level features. Two methods for integrating the information of soft biometrics in baseline algorithms are presented. One is to use the distance to SVM hyperplane as a soft weight and the other is to use the hard label as a pruning method. Comparing one with the other, it shows that soft weighting

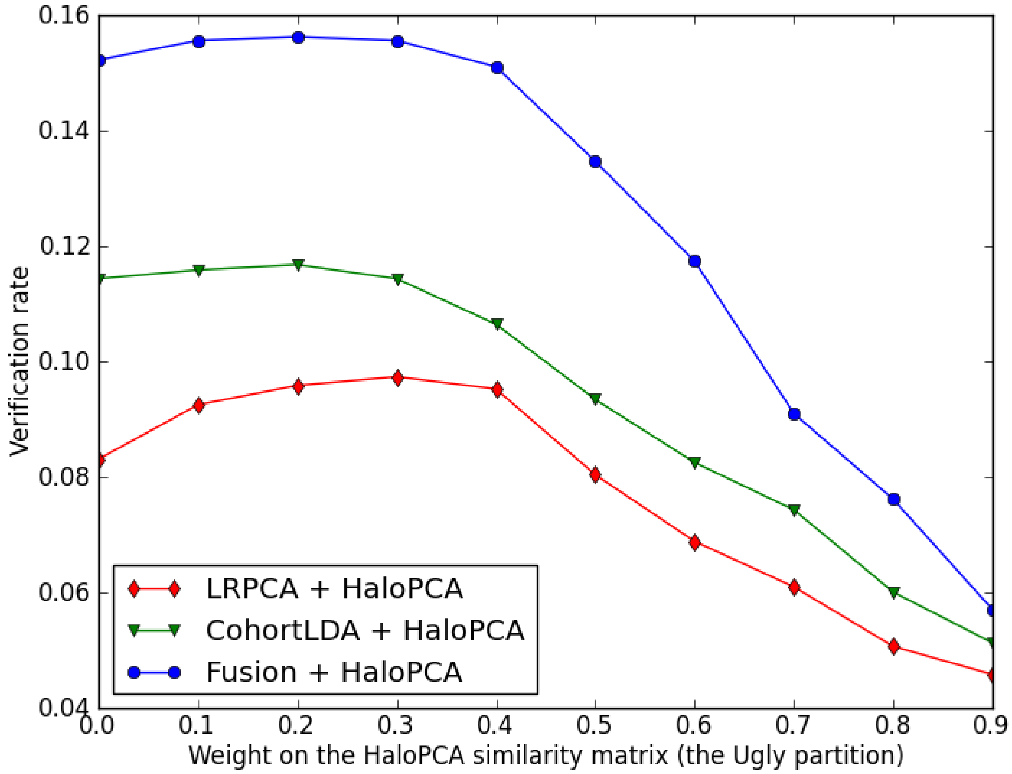


Figure 3.5: Fusing HaloPCA on Ugly

strategy works much better and using hard labels to prune the image pairs can easily do more harm than good given an imperfect classifier.

Furthermore, another continuous soft biometric, namely geometry is tested as a soft biometric. It improves the baseline algorithm performance better than other obvious soft biometrics since the two baseline algorithms do not consider the face shape much. It shows that whether a soft biometric is going to help improve the performance a lot depends on the algorithm itself. The soft biometric should introduce as much new information as possible in order for the improvement to be significant.

Finally, Halo PCA is introduced as a soft biometric, inspired by previous studies that claim there is useful information in the region between the face oval and the head. Just by looking at the halo itself, where the face oval and the background are masked out, we achieve a recognition rate of 4.4% on the Ugly partition, which is more than a half of what LRPCA achieves. Fusing Halo PCA with existing baseline algorithms also shows an improvement of the recognition rate.

# Chapter 4

# Best-Case Analysis: Gender and Race

## 4.1 Potential Improvement Analysis

In order for a soft biometric to help improve the performance of a face recognition algorithm, the algorithm itself is expected to have made mistakes on claiming a pair of images with different soft biometric labels to be a match pair. For example, a pair of female and male images are classified as a match pair. We analyze the mistakes made by LRPCA [22] and CohortLDA [16] and a fusion of three of the best performing commercial algorithms in the Face Recognition Vendor Test (FRVT) 2006 [21] regarding confusing gender and race on the good and ugly partitions. The procedure is described as follows. For each image in the query set, we find its nearest neighbor. If the nearest neighbor has the same gender/race as the query image, it is considered that the algorithm makes a correct decision with respect to gender/race. Otherwise, it is a wrong decision and can be fixed using the information of gender/race.

Table 4.1a and 4.1b show the percentage of query images whose nearest neighbor matches in terms of gender/race on the Good and Ugly partitions. It can be seen that all three algorithms do well internally on matching gender and race on the Good partition, while not so well on the Ugly partition. We will predict that using information of gender/race will not help much on the Good partition. However, it

Table 4.1: Potential improvement analysis on the Good and Ugly partition.

| | Gender | Race |
|---|---|---|
| LRPCA | 98.2% | 97.5% |
| CohortLDA | 99.4% | 99.4% |
| Fusion | 100% | 100% |

(a) The Good Partition

| | Gender | Race |
|---|---|---|
| LRPCA | 71.9% | 69.8% |
| CohortLDA | 87.6% | 84.8% |
| Fusion | 89.9% | 86.5% |

(b) The Ugly Partition

will improve the performance of the algorithms on the Ugly partition by a notable amount.

## 4.2 Best-Case Analysis: Gender and Race

The goal of this study was to determine how much recognition performance might be improved in practice using common soft biometrics. We chose to analyze performance with regard to the Good, Bad and Ugly (GBU) Challenge Problem [22], since it allowed us to analyze the effects of soft biometrics on data for which existing face recognition algorithms perform well (the so-called good partition) as well as data on which existing algorithms perform poorly (the ugly partition). The data also contains ground truth information for a couple of soft biometrics, namely gender and race[1].

We analyzed the performance of three algorithms on the good and ugly partitions: two open-source algorithms (local region PCA (LRPCA) [22] and CohortLDA [16]), and a fusion of three of the best performing commercial algorithms in the Face Recog-

[1]The GBU data also includes ground truth data on age, but there is so little age variance among subjects in the data that it is not a factor worth analyzing.

Table 4.2: Verification rates without and with pruning by gender and/or race.

| | No SB | G | R | G & R |
|---|---|---|---|---|
| LRPCA | 70.9% | 72.9% | 73.0% | 74.7% |
| CohortLDA | 84.1% | 85.0% | 85.1% | 86.4% |
| Fusion | 98.4% | 98.5% | 98.5% | 98.6% |

(a) The Good Partition

| | No SB | G | R | G & R |
|---|---|---|---|---|
| LRPCA | 8.3% | 9.7% | 9.6% | 11.3% |
| CohortLDA | 11.4% | 13.1% | 13.0% | 15.3% |
| Fusion | 15.2% | 17.9% | 17.7% | 20.1% |

(b) The Ugly Partition

nition Vendor Test (FRVT) 2006 [21]. For every algorithm, we looked at its performance without using gender or race as a soft biometric, and compared it to the algorithm's performance when similarity scores between images with incompatible soft biometrics were disallowed. In other words, matches between men and women or between people of different races were pruned from the set of possible matches.

Tables 4.2a and 4.2b shows the results for the good partition and ugly partitions respectively. The first column indicates the algorithm and the second the verification rate at FAR=0.001 without the aid of either gender or race information. The third column shows the results when matches across gender are disallowed, and the fourth column shows the results when matches across race are disallowed. The last column shows the results when both gender and race are used to disallow a match.

In all cases, performance gets slightly better when the soft biometrics are included. Since we are using error-free ground truth values for gender and race, performance could not logically have gotten any worse. But the gains in performance are very small. The Fusion algorithm, in particular, does very well on the good partition without using soft biometrics; it is right 98.4% of the time. Therefore, it doesn't make many mistakes for the soft biometrics to correct. Moreover, when it does make a mistake it apparently does not confuse men and women or people of different races very often, because eliminating these mistakes only raises performance 0.2%. The LRPCA and CohortLDA algorithms are not as good and therefore get a bigger increase in performance, but the differences are still small: using both gender and race improves the performance of CohortLDA by 2.3% and LRPCA by 3.8%.

One might expect, therefore, that information about gender and race would be more useful on harder data sets, where the algorithms make more mistakes. Surprisingly, Table 4.2(b) shows the results of the same type of analysis over the ugly partition. In this case, the baseline recognition rates of the three algorithms are 8.3%, 11.4% and 15.2%, respectively. Nonetheless, adding ground truth information about gender and/or race leads to only small improvements. Adding gender information improves performance by only 1.4%, 1.7% and 2.7%, respectively. Similarly, adding race information only improves performance by 1.3%, 1.6% and 2.5%. Even combining both biometrics only creates improvements of 3%, 3.9% and 4.9%. Apparently, although these algorithms make many mistakes on the ugly partition, they are not in general confusing men and women or people of different races.

I hypothesize that gender and race are not terribly useful as soft biometrics because they are global properties that determine many aspects of a person's facial appearance. Therefore, no matter what face recognition algorithm is used, it is fairly rare for a woman to be confused with a man or someone who is asian to be confused with someone who is white. Therefore eliminating these mismatches has only a small overall effect. Some identification documents include height information. If height Ire available as a soft biometric is might improve performance more, since height is presumably relatively uncorrelated to facial appearance.

# Chapter 5

# Discussions, Conclusions and Future Work

## 5.1 Discussions

As can be seen from those experiments, the improvement of existing face recognition algorithms using soft biometrics is not significant. The reasons are three-fold. Firstly, soft biometrics cannot be treated as additional information independent from the algorithm. They can be redundant. The algorithm itself may implicitly or explicitly encode the information of soft biometrics such as gender and race. In these cases, the recognition performance may even degrade if the soft biometrics are not used properly. Secondly, perfect classification of the soft biometrics can not be achieved currently. Considering the first reason, the improvement can still be algorithm-dependent even if the classification is perfect. Even though the information of soft biometrics is independent of the recognition algorithm, a poor classification can also degrade its performance. Thirdly, there are only a few soft biometrics that can be extracted on a single face image. The lack of the number of available soft biometrics limits the improvement gain of face recognition algorithms.

## 5.2 Conclusions

This paper presents experiments that try to improve face recognition performance using soft biometrics. At one level, all of them succeed: face recognition performance increases many cases. But the performance gains are never large; in fact, they could be described as disappointingly small. Nor is there any reason to expect a large performance gain: the first experiment shows that even perfect information about gender and race yields only a small gain in performance. The gains seen in the other three experiments are smaller because the soft biometrics themselves are noisy. This is consistent with the results reported by Park and Jain [19].

This paper is not meant to quash research in soft biometrics. Soft biometrics remains an interesting and important research topic. After all, we tested only a handful of soft biometrics and a handful of methods for fusing soft biometric information. Moreover, most of the soft biometrics tested helped performance a little bit; enough soft biometrics might lead to a qualitative improvement in performance. Nonetheless, this paper should dampen the expectation that a few obvious soft biometrics will lead to large performance gains on challenging data sets. To significantly improve performance, soft biometrics must be carefully designed or trained, and breakthroughs in soft biometric feature extraction may still be needed. In short, as a means of improving face recognition, soft biometrics are hard.

## 5.3 Future Work

There are some existing challenges in using soft biometrics to improve the performance the face recognition algorithms. One is that classifying soft biometrics is a supervised task. Since a number of training images are required to train robust classifiers for soft biometrics, labeling them can be a difficult task. Though Amazon MTurk can be used sometimes, it is possible that researchers do not have access to it in some cases due to security or financial reasons. A better method for classifying soft biometrics that requires less human interaction is needed.

The second challenge is that most soft biometrics are manually defined. Some of them are distinguishable but some are not good enough. There are cases some soft biometrics are ambiguous even for a human. We would like to focus on soft biometrics that help identify a person greatly and ignore ones that are not helpful. However, no effective rules are available to measure how good a soft biometrics is to help with identification. A good soft biometric should be data driven rather than manually defined.

The third challenge is that many existing face recognition algorithms have already encoded the information of certain soft biometrics. Soft biometrics work well for improving one face recognition algorithm can be redundant for another one. It is hard to determine which soft biometrics would help improve certain face recognition method especially when the recognition method is a black box.

Given the above three challenges, we will be working on extracting soft biometrics on a data driven manner. Both the distribution of the training images and the performance of the face recognition algorithm would be evaluated. The question of which Soft biometrics will be used can be answered by investigating the entropy of each potential soft biometric in the training data and the original verification result of all the match and non-match pairs. A soft biometric which can separate training data into clusters with close sizes and correct many errors made by the face recognition algorithm is considered useful.

# References


[1] Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Cuixian Chen, Yaw Chang, Karl Ricanek, and Yishi Wang. Face age estimation using model selection. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 93–99. IEEE, 2010.

[4] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[5] Timothy F Cootes, Christopher J Taylor, David H Cooper, Jim Graham, et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[6] Antitza Dantcheva, Nesli Erdogmus, and Jean-Luc Dugelay. On the reliability of eye color as a soft biometric trait. In *IEEE Workshop on Applications of Computer Vision*, pages 227–231, 2011.

[7] Antitza Dantcheva, Carmelo Velardo, Angela DAngelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

[8] Gareth Edwards, Timothy Cootes, and Christopher Taylor. Face recognition using active appearance models. *Computer VisionECCV98*, pages 581–595, 1998.

[9] Srinivas Gutta, Jeffrey RJ Huang, P Jonathon, and Harry Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *Neural Networks, IEEE Transactions on*, 11(4):948–960, 2000.

[10] L Hong, A Jain, and S Pankanti. Can multibiometrics improve performance?[online], available: citeseerx. ist. psu. edu/viewdoc/download? doi= 10.1. 1.3. 7621&rep= rep1&type= pdf, 2011.

[11] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. Can soft biometric traits assist user recognition? In Anil K. Jain and Nalini K. Ratha, editors, *Biometric Technology for Human Identification*, volume SPIE 5404, 2004.

[12] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication*, volume LNCS 3072, pages 731–738, 2004.

[13] Anil K Jain and Unsang Park. Facial marks: Soft biometric for face recognition. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 37–40. IEEE, 2009.

[14] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[15] Xiaoguang Lu, Anil K Jain, et al. Ethnicity identification from face images. In *Proceedings of SPIE*, volume 5404, pages 114–123, 2004.

[16] Yui Man Lui, D Bolme, PJ Phillips, JR Beveridge, and BA Draper. Preliminary studies on the good, the bad, and the ugly face recognition challenge problem. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 9–16. IEEE, 2012.

[17] Jamie R. Lyle, Philip E. Miller, Shrinivas J. Pundlik, and Damon L. Woodard. Soft biometric classification using periocular region features. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7, 2010.

[18] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *Computer Vision–ECCV 2008*, pages 504–513, 2008.

[19] Unsang Park and Anil K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, 2010.

[20] P Phillips, Patrick Flynn, J Beveridge, W Scruggs, Alice OToole, David Bolme, Kevin Bowyer, Bruce Draper, Geof Givens, Yui Lui, et al. Overview of the multiple biometrics grand challenge. *Advances in Biometrics*, pages 705–714, 2009.

[21] P. J. Phillips, K. W. Bowyer, and P. J. Flynn. Comment on the casia version 1.0 iris dataset. *IEEE trans. Pattern Analysis and Machine Intelligence*, 29(1869-1970), 2007.

[22] P. Jonathon Phillips and et al. An Introduction to the Good, the Bad, and the Ugly Face Recognition Challenge Problem. In *IEEE Conference on Automatic Face and Gesture Recognition*, pages 346 – 535, March 2011.

[23] P Jonathon Phillips, W Todd Scruggs, Alice J O'Toole, Patrick J Flynn, Kevin W Bowyer, Cathy L Schott, and Matthew Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831–846, 2010.

[24] Allyson Rice, Jonathon Phillips, Vaidehi Natu, Xiaobo An, and Alice O'Toole. Unconscious use of the body in identifying the face. *Journal of Vision*, 12(9):979–979, 2012.

[25] Walter J Scheirer, Neeraj Kumar, Karl Ricanek, Peter N Belhumeur, and Terrance E Boult. Fusing with context: A bayesian approach to combining descriptive attributes. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011.

[26] Caifeng Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33:431–437, 2012.

[27] Pawan Sinha and Tomaso Poggio. I think i know that face... 1996.

[28] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[29] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 1999.

[30] Haitao Wang, Stan Z Li, Yangsheng Wang, and Jianjun Zhang. Self quotient image for face recognition. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 2, pages 1397–1400. IEEE, 2004.

[31] Laurenz Wiskott, J-M Fellous, N Kuiger, and Christopher von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):775–779, 1997.

[32] Wenyi Zhao and Rama Chellappa. *Face Processing: Advanced Modeling and Methods: Advanced Modeling and Methods*. Academic Press, 2005.

[33] Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical methods in diagnostic medicine*, volume 569. Wiley-Interscience, 2009.