12-2014

Gender and Ethnicity Classification Using Partial Face in Biometric Applications

Jamie Lyle
Clemson University, jlyle@g.clemson.edu

Gender and Ethnicity Classification Using Partial Face in Biometric Applications

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Computer Science

---

by
Jamie Lyle
December 2014

Accepted by:
Dr. Damon L. Woodard, Committee Chair
Dr. Shaundra Daily
Dr. Juan Gilbert
Dr. Jason Hallstrom

Abstract

As the number of biometric applications increases, the use of non-ideal information such as images which are not strictly controlled, images taken covertly, or images where the main interest is partially occluded, also increases. Face images are a specific example of this. In these non-ideal instances, other information, such as gender and ethnicity, can be determined to narrow the search space and/or improve the recognition results. Some research exists for gender classification using partial-face images, but there is little research involving ethnic classifications on such images. Few datasets have had the ethnic diversity needed and sufficient subjects for each ethnicity to perform this evaluation. Research is also lacking on how gender and ethnicity classifications on partial face are impacted by age. If the extracted gender and ethnicity information is to be integrated into a larger system, some measure of the reliability of the extracted information is needed. This study will provide an analysis of gender and ethnicity classification on large datasets captured by non-researchers under day-to-day operations using texture, color, and shape features extracted from partial-face regions. This analysis will allow for a greater understanding of the limitations of various facial regions for gender and ethnicity classifications. These limitations will guide the integration of automatically extracted partial-face gender and ethnicity information with a biometric face application in order to improve recognition under non-ideal circumstances.

Overall, the results from this work showed that reliable gender and ethnic classification can be achieved from partial face images. Different regions of the face hold varying amount of gender and ethnicity information. For machine classification, the upper face regions hold more ethnicity information while the lower face regions hold more gender information. All regions were impacted by age, but the eyes were impacted the most in texture and color. The shape of the nose changed more with respect to age than any of the other regions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The world today is becoming more and more identity-driven. As the number of applications which require identification increases, the chance of identity-theft also rises. Modern biometric applications were developed as a solution to this problem. Identification can now take place by who you are, physically and behaviorally, as opposed to what you have or what you know. Biometric applications, which uniquely identify the individual, are useful, but not perfect under all conditions. Sometimes the data available to the application is not ideal and identification fails.

An example of this non-ideal information is grainy surveillance footage of a gas station robbery. The police probably will not be able to identity the suspect on the video, but they can determine certain information. This information could be the approximate height, gender, or ethnicity of the suspect. These descriptions will not identify the individual to the police, but they will narrow the possible suspects to the people who match whatever information is extracted. This type of information, which describes the individual but does not uniquely identity them, is known as soft biometric information.

Soft biometric information covers a wide range of details about a person. It can include height, weight, hair color, eye color, gender, age, and even clothing color. Some soft biometric traits are more permanent than others, gender versus clothing color for example. Some traits change naturally over time, such as height, weight, and age. Even though some traits are not entirely stable, certain applications might only need short-term information where the soft biometrics would be unlikely to change, or they could use the more permanent information such as gender and ethnicity. Which traits are the most useful depend on the application and the type of data it collects.

Soft biometric information can be used in various aspects of culture today including advertising, computer interaction customization, surveillance and tracking, as well as biometric applications. Advertisers today target ads to specific populations. Web advertisements can appear based on previous browsing behaviors. Targeted advertising even shows up in movies; for example, shoppers are identified based on iris scans and their purchase history made available in order to personalize the store's interaction with them [50]. Most likely, soft biometric information would allow targeted ads based on demographic groups as opposed to a specific identity, but the idea remains the same. Human Computer Interaction (HCI) is a growing field of study which can benefit from the use of soft biometrics. Suppose the computer could determine an age and take appropriate steps based on that age, such as enlarge the font for an older individual or enable parental controls for a younger individual. Law enforcement uses soft biometric information to describe subjects as mentioned previously and algorithms exist that can track an individual across camera views using soft biometric information [19].The main focus of this paper will be an analysis of gender and ethnicity classifications on various regions of the face, culminating in using the results of the analysis in an effort to improve the performance of a biometric application using face information.

Park and Jain [56] group the uses of soft biometric information in biometric and computer applications into four main categories. The first category is using soft biometric information to supplement existing biometric systems to improve identification accuracy. The second category is enabling faster image retrieval which could be used in a biometric application or an image database. The next category mentioned is enabling matching or retrieval of facial images that are occluded or captured at an angle. The last category is using soft biometric information to provide more descriptive evidence about the similarity or dissimilarity between facial images, which could be useful in judicial settings. These categories were created with face images in mind, but can generalize well to other biometric modalities. This work seeks to further the understanding of the use of partial face in the first two categories to enable future use in all categories.

The face modality is a well-researched modality in the biometrics community. Most human recognition takes place based on the face and it is a relatively easy modality to acquire. The face is a good source of gender and ethnicity information, which is the soft biometric information of interest in this work. Partial face has not been researched as in-depth as full face for both gender and ethnicity classification, but deserves more attention. As surveillance increases and awareness of surveillance increases, the difficulty of acquiring a good quality full face image increases as well.

The probability that some part of the face will be occluded or blocked from the camera is high. Therefore, an analysis of what information, namely gender and ethnicity, can reliably be found in different regions of the face is needed. Since surveillance is not limited to any certain age group, an investigation of how age impacts the partial-face gender and ethnicity classification is needed as well. Before the details of the analysis are discussed, a short overview of biometric systems is included for background information.

1.1 Biometrics Overview

Biometrics, within the security and computing fields, is defined as the science of identifying an individual based on physiological or behavioral characteristics [6]. Physiological characteristics belonging to an individual that have been used for biometric purposes include fingerprints, face, iris, and hands. Behavioral characteristics include voice, signature, and the way a person walks (gait). The specific characteristic used is known in the literature as a biometric modality or simply a modality. Biometric applications can use a single modality, such as in face recognition systems, or can use multiple modalities, such as a system that combines face and gait information to identify an individual. The type of modality used will depend on the problem the application is trying to solve.

The two main types of problems for biometric applications are verification and identification. Applications that deal with both problem types have a similar structure, which can be seen in Figure 1.1. First, the system needs data, which is captured by a sensor. The sensor could be a camera or an audio recorder. The data is preprocessed to get it in the right form for feature extraction. In facial recognition this step includes face detection, alignment, and contrast enhancement. Once the data is in the right format, the system extracts features from the data. These could be texture, shape, or color features from an image. The features are then used to generate a template. If the system is in the learning or enrollment phase, the templates are stored in a database along with an identity. If the system is in testing phase, the template, also known as the probe, is compared against other stored templates, known collectively as the gallery. A match score is given for each comparison. The system uses these scores to make a decision. The decision is different for verification and identification systems.

Access control is an example of a verification problem. Only certain individuals are permitted access and their identities must be verified before they can gain access. The person who wishes

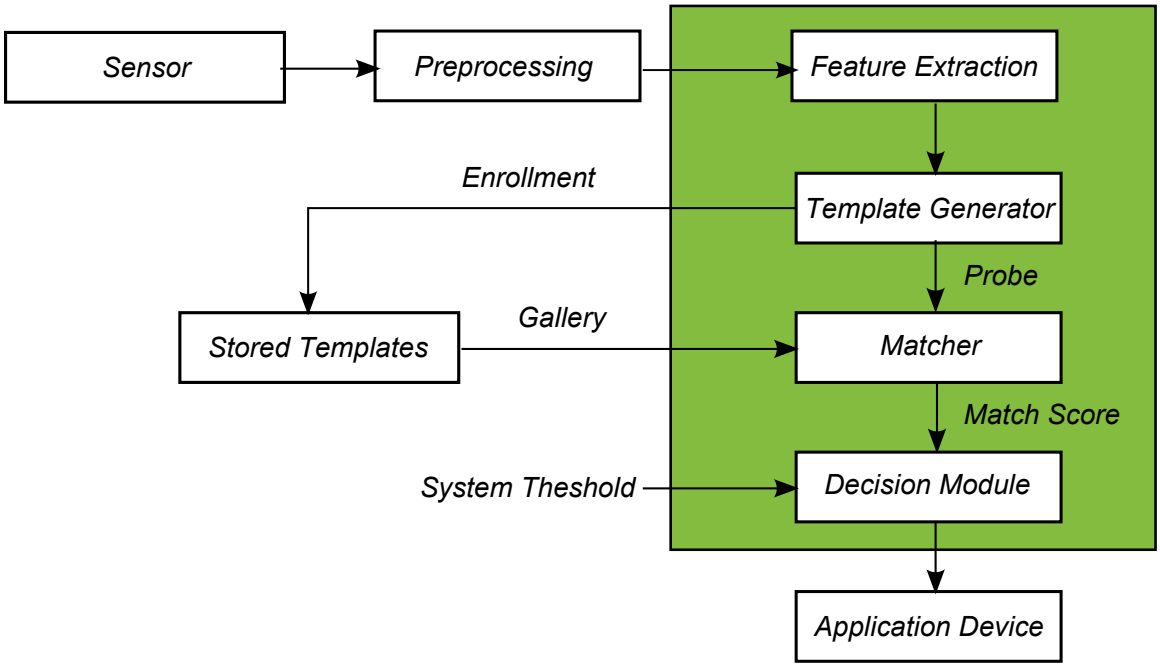


Figure 1.1: Typical biometric system

access claims an identity. The biometric system processes the data and compares the probe to the stored template that corresponds to the claimed identity. The system has a threshold for match scores in its decision module. If the score is above the threshold, the system decides that the person and the identity are the same and grants access. If the score is below the threshold, the system decides that the person and the claimed identity are not a match and denies access.

In an identification problem, the unknown person makes no identity claim. Since no claim is made, the probe is compared against multiple, possibly all, gallery templates stored during the enrollment phase. The match scores are sorted. The corresponding identities are ranked from the most likely match to the least likely. The most likely identity can be returned or a list of most likely identities can be returned for a human expert to make the final decision.

Some cases arise where identity is not the main goal or is not plausible with the given data. In those cases classification by some other criteria, such as age or gender, is desired. This soft biometric classification may be the end result of the system or it could be incorporated into a biometric application in some way. Classifications on the probe could be used as another feature to increase confidence in a possible identity match. They could also be used as a filter to only compare the probe with identities in the gallery that have matching soft biometric information. A

soft biometric classification system starts the same as either a verification or identification system. A sensor collects the data which is then preprocessed. Features are extracted from the preprocessed data. In place of enrollment, a classification system undergoes a training phase. The training can either be supervised, where class labels are provided for the training data, or unsupervised, where labels are not provided. The trained system will take the features extracted from the data and output a class label, such as male or female.

With a general understanding of biometrics and classification systems, the proposed analysis will be placed in context with previous research involving partial-face classification of gender and ethnicity.

1.2 Previous Work

Much of biometric research focuses on using the face for recognition purposes. The face is very visible in most cultures and is a large part of how humans recognize one another. In many instances, not even the entire face is needed for a human to recognize an individual. When looking at face recognition by machines, the quality of the face image is an important factor. Blurred images [39], occluded faces, and low or uncontrolled lighting [7] can all negatively impact the results of facial recognition algorithms.

Identity is not the only information to be found in the human face. Previous studies have shown that the face holds both gender [24, 29, 72] and ethnicity [28, 29, 30, 45] information in addition to identity. In the event that a face recognition system is not confident of its result, gender and ethnicity information can be used to increase the confidence of the system. In an image of a face though, there can be noise which would inhibit an accurate determination of gender or ethnicity. Studies have been performed to show the usefulness of various subregions of the face for gender [5, 9, 35, 47] and ethnicity [40, 47] determination; however, most of the studies are on a small scale and do not carry out experiments on all regions of the face. The goal of the research detailed herein is to gain an understanding of the limitations of partial-face regions for gender and ethnicity classifications. This understanding will including learning how reliable each part of the face is for gender and ethnicity classification and how each region is impacted by age. Knowing these limitations, a method using machine-based gender and ethnicity classifications made on partial face is proposed to improve performance in a facial biometric application.

1.2.1 Reliability

What parts of the face hold reliable gender and ethnicity information for machine applications?

The first question to be considered is which parts of the face hold reliable gender and ethnicity information for machine applications. It is possible that a subset of the face holds more gender or ethnicity information than another subset. Some regions of the face may be gender or ethnic neutral. Another consideration is what type of features encode the gender or ethnicity information. Texture, color, and shape features can be found in facial images. Knowing which regions of the face hold reliable gender and ethnicity information and what features encode the information best will provide a foundation for improving gender and ethnicity classifications based on partial face.

1.2.1.1 Face Regions

Previous studies have focused mainly on gender when working with partial face [5, 8, 35, 36, 44, 55, 69]. Details for partial face studies mentioned can be found in Table 1.1. Some details of the databases used can be found in Table 1.2. Most of these studies use human-defined regions to subdivide the face, such as the eyes, nose, mouth, and chin. Regions containing the eyes ranked among the top two regions for several studies [5, 8, 44]. These results were derived from three different datasets, so the good performance of the eye region is not a peculiarity of a specific dataset. It is more likely that the eye region holds more gender information than other regions tested. Other best performing regions were the nose in FERET experiments and the mouth in XM2VTS experiments [5]. The nose performed the worst in the XM2VTS experiments. This could be due to the different ethnic compositions of the datasets, or the authors suggested that the nose was more susceptible to illumination than the mouth. The differences between the best performing regions in the various datasets suggest that more research is needed in this area. It could be that the sample/experiment size was not large enough. Most of the studies perform experiments with less than 1,000 subjects, as seen in Table 1.3. Lapedriza *et al.* [37] have one of the larger studies based on image number, but the number of subjects is unknown and their research focuses more on features outside the face such as the ears, forehead, and hair, instead of the eyes, nose, mouth, and chin. Aside from the nose [5], mouth [5, 36] and eyes [5, 8, 44, 47] mentioned above, another best performing region of the face is the jaw or chin [35, 36].

Study	Feature	Classifier	Rate	Modality	Trait	Dataset
Ozbudak [55]	DWT+PCA	FLD	93% (52%)	Face (w.o. nose)	G	FERET, SUMS
Kawano [36]	FDF	LDA	93.7% (89.9%)	Face (Jaw)	G	Softopia Japan
Andreu [5]	Pixels+PCA	SVM	90.6% (84.0%)	Face (Mouth)	G	XM2VTS
Andreu [5]	Pixels+PCA	SVM	95.2% (86.4%)	Face (Nose)	G	FERET
Buchala [8]	Pixels+PCA	SVM(RBF)	86.5% (92.3%)	Face (Composite)	G	FERET, AR, BioID
Lu [44]	Pixels	SVM(RBF)	92.5% (92.9%)	Face (Upper)	G	CAS-PEAL
Lu [43]	2D-PCA	SVM(RBF)	90.4% (95.3%)	Face (Fusion)	G	CAS-PEAL
Hu [31]	Curvature	SVM	93.5% (94.3%)	Face (Fusion)	G	own, UND (3D)
Lapedriza [38]	Face Fragments	SVM	94.2%	External Face	G	FRGC
Manesh [47]	Gabor	SVM(RBF)	94.0%	Face Fusion	G	FRGC, CAS-PEAL
Lapedriza [37]	Face Fragments	JointBoost	96.8% (96.7%)	Face (External)	G	Controlled FRGC
Lapedriza [37]	Face Fragments	JointBoost	91.7% (90.6%)	Face (External)	G	Uncontrolled FRGC
Thomas [68]	Texture	DT	75%	Iris	G	Custom
Lyle [46]	Texture , Color	ANN/SVM	97.3%	Periocular	G	FRGC
Lyle [46]	Texture	ANN/SVM	90%	Periocular	G	MBGC
Merkow [48]	Texture, pixels	LDA+SVM	85%	Periocular	G	web
Manesh [47]	Gabor	SVM(RBF)	97.4%	Face Fusion	E(2)	FRGC, CAS-PEAL
Qiu [59]	Gabor features	AdaBoost	86%	Iris	E(2)	CASIA, UPOL, UBIRIS
Qiu [60]	Iris-Textons	SVM	91%	Iris	E(2)	CASIA-BioSecure
Li [40]	Eyelash direction	1-NN	93.2%	Periocular	E(2)	CMU-PIER, UBIRISv1
Lyle [46]	Texture , Color	ANN/SVM	94%	Periocular	E(2)	FRGC
Lyle [46]	Texture	ANN/SVM	89 %	Periocular	E(2)	MBGC

Table 1.1: Details and results for partial face experiments. Traits are gender (G) and ethnicity (E) with (x) classes. Results from best performing subregion were included if applicable.

Database	Images	Subjects	Short Description
BioID	1,521	23	Grayscale face images captured under real-world conditions
UPOL	384	64	Color iris images
CMU-PIER	321	107	Iris images collected for ethnicity classification based on eyelash direction
MBGC	149 (videos)	114	Multi-modal face and iris database captured under near infrared illumination, video and stills
AR	4,000	126	Color face image captured under varying illumination and facial expressions. Occlusions were created with sunglasses and scarves.
UBIRIS (v1)	1,877	241	Noisy iris images captured under visible light conditions (color)
XM2VTS	1,180	295	Multi-modal face database including high-quality color face images
Softopia Japan	1,200	300	Color face images, equal male and female subjects
SUMS	400	400	Grayscale face images from equal male and female subjects from Stanford University Medical students
FRGC	39,329	568	High-quality color face images from students and faculty at University of Notre Dame
CASIA (v3)	22,051	700	Grayscale iris images captured under near infrared illumination
CAS-PEAL	30,871	1,040	Grayscale Chinese face database captured under varying pose, expression, accessory, and lighting conditions
FERET	14,126	1,199	Color face images captured under semi-controlled conditions.
MORPH	55,134	13,314	Color face images collected from public records taken under real-world operating conditions (semi-controlled)
Pinellas	1,447,607	403,619	Color face images collected from Pinellas County Sheriff's Office

Table 1.2: Short description and details from databases found in previous work in partial-face gender and ethnicity classification and those to be used.

Other studies broke the face into regions differently, both larger and smaller than the previously mentioned human-defined regions. The area around the eye still performs well in these studies [44, 47]. In a study by Manesh *et al.* [47], the eyebrows and the area between the eyes performed best for gender classification. The upper region of the face, which includes the eyes, also performed well in a study by Lu *et al.* [44]. The drawback of using the larger regions is that it is unclear which smaller portion of the face actually encodes the gender information.

Ethnicity research on partial face are not as widespread as gender research, but some research has been done in the past. The eyes are still a good choice for ethnicity classification according to

Study	Trait	Type	Subjects	Images	Datasets
Ozbudak [55]	Gender	Texture	–	480	2
Kawano [36]	Gender	Shape	300	1,200	1
Andreu [5]	Gender	Pixels	203	1,378	1(XM2VTS)
Andreu [5]	Gender	Pixels	834	2,147	1(FERET)
Buchala [8]	Gender	Pixels	–	400	3
Lu [43, 44]	Gender	Pixels	–	800	1
Hu [31]	Gender	Shape	321	945	2
Lapedriza [38]	Gender	Texture	–	2,640	1
Manesh [47]	Gender, Ethnicity	Texture	1,691	1,691	2
Lapedriza [37]	Gender	Texture	–	3,440/1,886	1
Thomas [68]	Gender	Texture	–	57,137	1
Qiu [59]	Ethnic	Texture	–	39,982	3
Qiu [60]	Ethnic	Texture	60	2,400	1
Li [40]	Ethnic	Texture	214	642	2
Lyle [46]	Gender, Ethnicity	Texture, Color	404	4,232	1
Lyle [46]	Gender, Ethnicity	Texture	60	350	1
Merkow [48]	Gender	Pixels,Texture	–	936	1

Table 1.3: Details of partial face experiments found in literature.

Manesh *et al.* [47], but their results suggest that the cheeks are also useful for determining ethnicity. Manesh *et al.* [47] also have the largest number of subjects in their gender and ethnicity experiments with 1,691 subjects from the FERET and CAS-PEAL datasets. However, both datasets were needed to gain the desired ethnic diversity. The majority of each ethnic class was found in only one of the databases. The authors took steps to try to minimize the influence of using two different datasets, but their system could have learned the image properties of the database instead of the ethnic class.

Studies on just the iris and periocular regions have been performed in both gender [46, 48, 68] and ethnicity [40, 46, 59, 60]. This is the region of the face where most ethnic research has been performed. The eye region is able to obtain at least 85% on most binary gender and ethnicity classifications, indicating that it holds a large amount of gender and ethnicity.

Taking all this previous work under consideration, experiments will be performed on the eyes, nose, mouth, and chin regions of the face in reliability experiments. Most of the facial regions, at one point or another have all been the best performing region on some dataset. It is hoped that experiments within this work will have some measure of agreement and be able to relate back to previous works. Regions including the eyes are expected to perform well for both gender and ethnicity classification across all datasets as they are among the best regions for the majority of previous studies.

1.2.1.2 Features

Various types of features have been used in the past for gender and ethnicity classification. The main types of features found in images can be categorized into pixels, texture, shape, color, and key-point. Pixel intensity values are one of the most basic features used for these purposes [44, 48]. This type of feature is commonly used in combination with a projection into a different space, such as Principal Component Analysis (PCA) [5] or Independent Component Analysis (ICA) [32], for feature reduction.

Local texture features are another type of feature that can be extracted from facial images. Image texture can be characterized as a set of repeating patterns. Within the face, texture can range from very fine skin texture to larger texture caused by hair and wrinkles. Texture feature extraction techniques used in partial-face research include Gabor features [47, 59], iris texture [60, 68], Histograms of Oriented Gradient (HOG) [46], Discrete Cosine Transform (DCT) [46], and the texture of the face fragments used by Lapedriza *et al.* [37] is derived using Gaussian derivative filters. The Local Binary Patterns (LBP) feature extraction method is another widely used texture representation that has been used for gender and ethnicity classification in both full face [41, 66, 75] and partial [46, 48].

Aside from local texture information, shape information can also be found in the face. 3D shape information has been utilized for gender classification [31]. Active Appearance Models (AAMs), which combine shape and texture information, can be used to represent a face. AAMs model the statistical shape of an object, as well as texture, and can be used to find a shape, such as the face, in an image as well as for matching or classification [15]. AAMs have been used successfully for gender classification [64]. Another approach which uses the distance and angles between facial landmarks is called face metrology. This method has been used to successfully predict gender in facial images [10].

Another type of feature representation is Scale Invariant Feature Transform (SIFT). This technique is different from the others in that it only looks at specific key-points on the image instead of calculating statistics over the entire image or finding a sequence of points (shape). SIFT has been used to successfully perform gender classification using full or occluded facial images [69].

The last category is color. Local Color Histograms (LCH) are a simple color representation that have been used for both gender and ethnic classification [46]. Color is often an important cue

for humans in determining ethnicity. One drawback to using color is that it can be easily changed by various factors such as temperature, cosmetics, or illumination. Because of this, it might be useful to combine color features with features from another category, such as texture or shape.

Since texture is a large part of the information to be found in partial face regions, several different texture representations will be investigated to see if any one is better suited to a particular region than another. The LBP feature extraction technique was chosen due mainly to its success in gender classification using full face. HOG is another interesting texture representation which was first developed for pedestrian detection [18]. Its discriminating capability has been shown to be successful in facial recognition [20, 65], as well as periocular gender and ethnicity classification [46]. Local Phase Quantization (LPQ) is the last local texture representation chosen for investigation. LPQ is a recently proposed texture descriptor [54] which is said to be robust to image blurring. It has been used successfully for facial recognition [2, 11, 12]. It is expected that its descriptive capabilities will also work well for gender and ethnicity classification on partial-face images. LPQ was also chosen because the robustness to blur would mean that image quality might not be as important in order to achieve good performance. Local Color Histograms (LCH) will be used to represent color information based on previous performance [46]. Shape features will be investigated using a version of face metrology [10] adapted to partial face. A more detailed description of these methods can be found in the following chapters.

1.2.1.3 Classification

Early work in gender classification of faces relied on neural networks [1, 16, 22, 24, 67] to classify the data. The classifiers most widely used for gender and ethnicity determination in more recent works are still various forms of of Artificial Neural Networks (ANN) as well as Support Vector Machines (SVM) which can be seen in Table 1.1. In order to give a comparison, both SVM and ANN classifiers will be investigated as classifiers for this work. A simpler classifier, nearest neighbor, will also be used as a baseline with which to compare ANN and SVM classification. A more detailed description of these classification methods can be found in Section 2.5.

1.2.2 Impact of Age

How is machine classification of gender and ethnicity in partial face impacted by age?

The second question to be considered is how age impacts gender and ethnicity classification

for each part of the face. The human face changes as an individual ages. Younger faces are fuller, lacking the fine lines and wrinkles the face acquires as it ages [14]. The skin starts to sag as the face ages, losing it elasticity, and fine lines become more apparent. The eyebrows can even fall to the level of the brow. The eyes seem to sink into the face and "crow's feet" appear at the eye corners. The cheeks hollow as the fat deposits sink or are redistributed in the face. These changes are natural in the aging process and can occur at different times for different individuals. How do these changes affect gender and ethnicity classification using partial face? Is there a specific region of the face that works best for a certain age range or one that works well across various ages? Being able to characterize how performance is impacted by age will allow for better design choices in a biometric application when the demographics of the population are known in advance.

1.2.2.1 Facial Regions

Research in gender classification on full face shows that gender classification is impacted by age. The results of multiple studies [26, 62] indicate that gender classification is easier on adult faces than younger or more senior faces. The results in [26] were 10% higher on adult faces than seniors or children, with subjects ranging from 0 to 93 years of age. Successful recognition is possible on younger faces though. One machine classification method [71] achieved better gender recognition in toddler faces than humans performing the same task. Not very many researchers have looked at how age affects gender classification using partial face images. Kawano *et al.* [36] only looked at how age affected their best performing partial-face region, which was the jaw. They found that the best gender performance on the jaw was on subjects in their 30's. The error rate for subjects that were either older or younger increased. This is very similar to full face results; however, it would be interesting to see if the trend holds for other regions of the face, not just the jaw.

Not as much research has been done on ethnicity classification and the impact that age has on performance. Ethnicity results in [27] on full face were not as affected by age as the gender results mentioned above, but the age range of the subjects is different from the gender experiments. The subjects in the dataset used for the ethnicity experiments were limited mostly to adult faces with ages ranging from 16 to 67 years of age, while the gender study included children as well as adults up to 93 years old. No partial face research was found which looked at ethnicity classification in the presence of age.

A specific study of how gender and ethnicity classifications on the various facial regions are

affected by age would be beneficial. Knowing how each individual part is affected by age would contribute to the confidence level for gender and ethnicity classifications. If there is a specific part of the face in which gender and ethnicity information is least impacted by age, that would be the best region when age is a factor. Ethnicity classification on partial face is expected to be less impacted by age than gender classifications based on the literature for classification of full face.

1.2.2.2 Features

The feature extraction methods used when looking at gender and ethnicity over an age range include many of the same ones mentioned in the previous section. Pixel intensity values, HOG features, and LBP features are among the features compared by Guo *et al.* [26]. A different feature, named Biologically Inspired Features (BIF), is used by Guo *et al.* [26, 27] for both gender and ethnicity classification in the presence of aging. BIF are based, at the lowest level, on the results of Gabor filtering on an image. The other feature extraction most used in the presence of aging is Active Appearance Models (AAM) used by Wang *et al.* [71] to classify the gender of toddlers. Shape features are important because the change in features from infants to adults are mainly shape-based. Texture is the next big change as the adult face ages. This is when the skin loses its elasticity and fat layer underneath, allowing the wrinkles to appear. The main features used for gender and ethnicity classification over age fall into the categories of pixels, shape, and texture.

Experiments on how age impacts performance will use the same features mentioned in Section 1.2.1.2. These features belong to the color, texture, and shape categories. Since facial texture changes as an adult ages, it is expected that results for texture features will be impacted by age. Color also changes some with age, but will probably not be as impacted by the age range present in the datasets used for experiments.

1.2.3 Application

To what extent can automatically determined gender and ethnicity information improve performance in a biometric experiment using the face or partial face modality?

The final question to consider is how to use gender and ethnicity classifications on partial faces to improve the performance of a biometric application. Various studies have investigated fusing soft biometric information with a biometric experiment in an effort to improve performance [3, 33, 34, 45, 76]. Jain *et al.* [34] were the first to introduce the terminology "soft biometrics"

and used gender, ethnicity, and height information to improve the performance of a fingerprint experiment. They achieved an improvement of approximately 5% when including the soft biometric information. In a later experiment involving facial features [33], the extracted gender and ethnicity information failed to improve facial performance, while the simulated height information improved performance by approximately 5% again. Jain *et al.* [33] concluded that soft biometric information improves performance only if the information is complementary, or independent, of the primary modality used in the experiment.

Another way to utilize the soft biometric information is to use it to retrieve a candidate list from the stored templates, similar to indexing a database. Park *et al.* [56] use facial marks, such as freckles, moles, scars, and birthmarks, along with gender and ethnicity information to filter the stored templates in their experiments. Using these soft biometric traits, they were able to increase performance by 1.5%. An improvement is seen, even though the soft biometric information was not independent of the facial features used in the biometric recognition experiment. However, the gender and ethnicity information used here was not automatically extracted in the course of the experiment.

If the soft biometric information does not need to be independent of the primary modality, it is possible that gender and ethnicity information can improve the performance of a face experiment. The best methods from the previous sections will be combined with face features in an effort to improve performance. The machine-based soft biometric classifications will be used as a filter for the stored templates to see what improvement can be achieved. These combinations can be useful for both recognition and verification applications.

Each of these research questions will be addressed in the subsequent chapters, along with the design of the experiments for each question (Chapter 2). The analysis on reliability and impact of age will be performed for the three main types of features found in face images: color (Chapter 3), shape (Chapter 4), and texture (Chapter 5). The application of the research to use the acquired gender and ethnicity information to improve an everyday face biometric system (Chapter 6) will utilize conclusions from the analysis. Overall conclusions from each chapter will provide the basis for future work (Chapter 7.)

Chapter 2

Research Design

2.1 Data

Three databases will be used for experiments in this work: the Facial Recognition Grand Challenge database, the Craniofacial Longitudinal Morphological Face database, and a face database collected by the Pinellas County Sheriff's office. The collection and composition of each database are discussed in the following sections as well as the motivation behind including these databases.

2.1.1 Facial Recognition Grand Challenge

The Facial Recognition Grand Challenge database (FRGC) [57] was collected at the University of Notre Dame. The database is composed of full frontal still images. Images were collected from 568 subjects under varying lighting conditions and with different facial expressions, see example images in Figure 2.1. The images were collected on campus over the 2002-2003 and 2003-2004 academic school years. As a result, the majority of subjects in the database are between the ages of 18 and 27 years old. The approximate ages of the subjects at the start of collection in 2002 can be seen in Figure 2.2a. With a median age of 19, this database will not be used in age experiments; however, the distribution of the subjects by gender, Figure 2.2c, and ethnicity, Figure 2.2b, show enough diversity for FRGC to be used in reliability experiments for gender and ethnicity classification.

The FRGC database was chosen for reliability experiments for several reasons. First, the FRGC database is widely used in facial recognition research [49, 58] and gender classification [37,

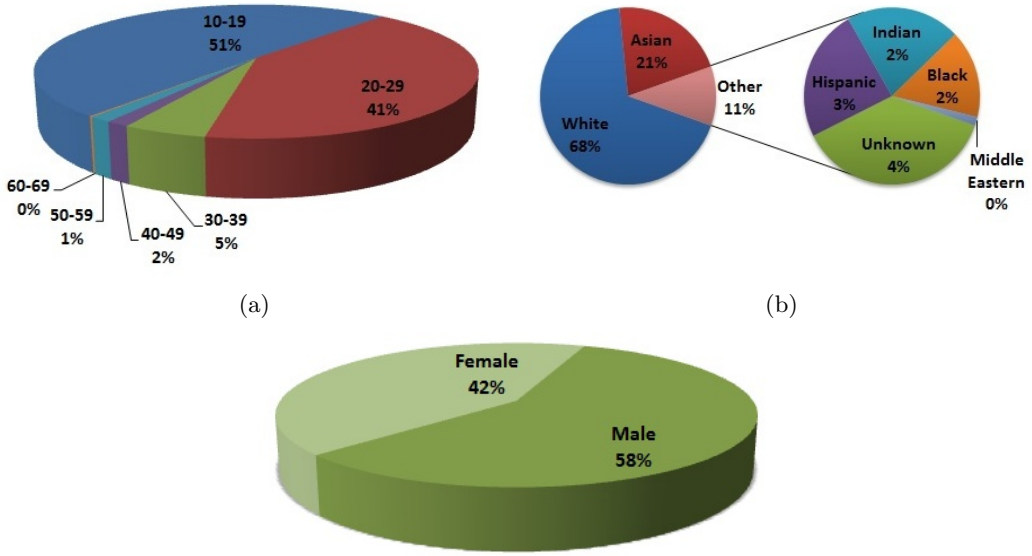
Figure 2.1: Example images from the FRGC database.

38, 47]. This database was included to provide a basis of comparison with some of the previous full-face and partial-face studies performed in gender and ethnicity classification. Second, the images are high-resolution. The high-resolution allows for skin texture to be captured in detail. Experiments on the high resolution images will provide a baseline for experiments performed on the other databases which are captured at a lower resolution.

Not all the data in FRGC is usable, and the data that is usable must first be processed. The subjects corresponding to the *Unknown* ethnicity label cannot be used in ethnicity experiments, so those images are discarded for both gender and ethnicity experiments. Only one subject exists in the database in the *Middle Eastern* ethnicity label. Since distinct subjects cannot be used for training and testing with only one subject, this class will be discarded. Images taken under uncontrolled lighting conditions were discarded, as well as images where the subject was wearing glasses. Four images were used for each subject. Some subjects did not have four which met these requirements and were excluded. The total number of subjects for FRGC experiments is 535. With four images per subject, the number of face images in FRGC experiments is 2,140. The average interocular distance for these images is 270.32 pixels. Images were not sorted on expression, so a subject can have images with either a neutral or smiling expression. A breakdown of the subjects in the experiment set can be seen in Table 2.1. Other considerations and preprocessing steps will be discussed later.

2.1.2 Craniofacial Longitudinal MORPHological Face

The Craniofacial Longitudinal Morphological Face database (MORPH) [63] is a collection of facial images taken from public records. The images were taken under real-world conditions and



(a) (b)

(c)

Figure 2.2: Distribution of the FRGC dataset according to a) age, b) ethnicity, and c) gender.

FRGC				**MORPH**			
Ethnicity		*Gender*		*Ethnicity*		*Gender*	
White	379 (4)	Male	309 (4)	White	2,666 (2.7)	Male	11,182 (2.7)
Asian	119 (4)	Female	226 (4)	Asian	50 (2.6)	Female	2,132 (2.7)
Hispanic	15 (4)			Hispanic	529 (2.8)		
Indian	12 (4)			Black	10,056 (2.7)		
Black	10 (4)			Native American	13 (3.1)		
Total	535 (4)		535 (4)		13,314 (2.7)		13,314 (2.7)

Table 2.1: Breakdown of subjects in the FRGC and MORPH experiment sets by gender and ethnicity. The average number of images per subject in each class is given in parentheses.

not by researchers under controlled conditions. Example images from this database can be seen in Figure 2.3. The MORPH database was formed for the purpose of studying age progression. It contains multiple images of an individual spanning both large and small age gaps. The MORPH database has two subsets, Album 1 and Album 2. The MORPH public release [61], used for this work, contains all of Album 1 and a subset of Album 2. This version contains over 55,000 images of more than 13,000 subjects. Subjects in the dataset range from 16 to 77 years old with a median age of 33. Gender and ethnicity information are included along with the age information for this dataset, as seen in Figure 2.4, providing ground truth for later experiments. The resolution of images in MORPH is either 200×240 or 400×480.

The MORPH database, chosen for reliability and age experiments, was included for several

Figure 2.3: Example images from the MORPH database.

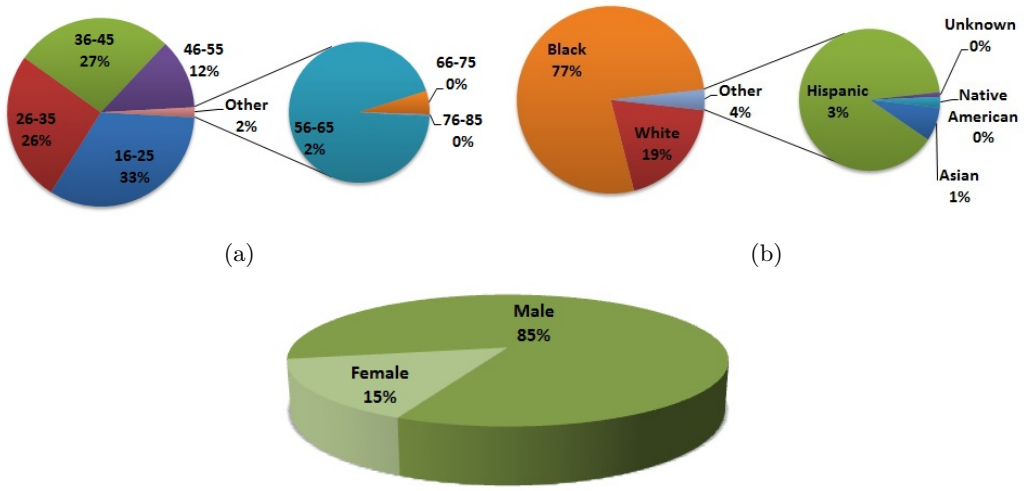
reasons. MORPH Album 2 has been used in previous work [27] for ethnicity estimation on faces. MORPH includes five ethnicity labels that provide the diversity needed for ethnicity classification. Also, since the data was captured under real-world conditions and not in a research environment [51], results from MORPH should provide a more accurate measure of performance of the proposed methods in real-world applications. The demographic distribution provides the opportunity to see if performance is impacted by any specific demographic, which is the final reason MORPH was chosen for this work.

As with the FRGC database, some of the data in MORPH cannot be used. Images of subjects belonging to the *Unknown* class will be discarded. The *Unknown* class is not useful for supervised learning techniques. The image list for MORPH was created by selecting a maximum of four images per subject. Some subjects do not have four images in the database. These images were examined to determine if the point fit was good and the subject was not wearing glasses. A subject was excluded from the experiment only if *all* of his or her candidate images did not meet the criteria mentioned above. The experiments on the MORPH data set included 35,601 face images of 13,314 distinct subjects. The breakdown of subjects in the chosen experiment set can be seen in Table 2.1 along with the average number of images per subject. The average interocular distance for images used in MORPH experiments was 93.51 pixels.

2.1.3 Pinellas

The Pinellas database is a collection of booking (mug shot) images from the Pinellas County Sheriff's Office[1] in Florida. Example images can be seen in Figure 2.5. With approximately 1.4 million facial images from over 400,000 subjects, this database is one of the largest facial databases

[1]The mug shot data used in these experiments were acquired in the public domain through Florida's "Sunshine" laws. Subjects shown in this manuscript may or may not have been convicted of a criminal charge, and thus should be presumed innocent of any wrongdoing.



(a) (b)

(c)

Figure 2.4: Distribution of the MORPH dataset according to a) age, b) ethnicity, and c) gender.

Figure 2.5: Example images from the Pinellas database.

to date. The distribution of images according to age, gender, and ethnicity can be seen in Figure 2.6. Images in the database average 480 × 600 pixels.

This dataset was chosen for its size as well as the ethnic and age diversity present in the images. Pinellas is much larger than MORPH and provides plenty of images for the application experiments. It was also included for reasons similar to MORPH. The data was captured under real-world conditions, at the sheriff's office, and so can provide a more accurate measure of performance of the proposed methods in the real world.

As with the MORPH and FRGC databases, images of subjects belonging to the *Other* class will be excluded from experiments. Further details on the selection of images from this dataset can be found in Chapter 6.

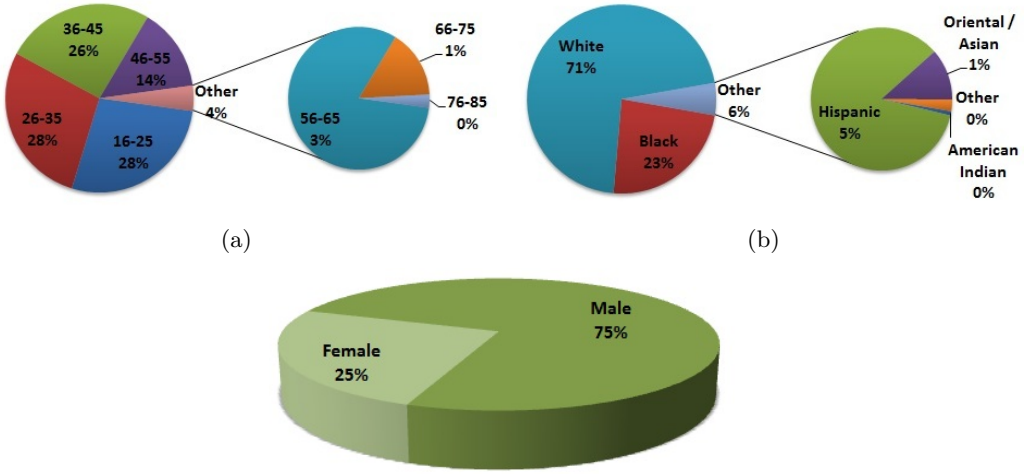


Figure 2.6: Distribution of the Pinellas dataset according to a) age, b) ethnicity, and c) gender.

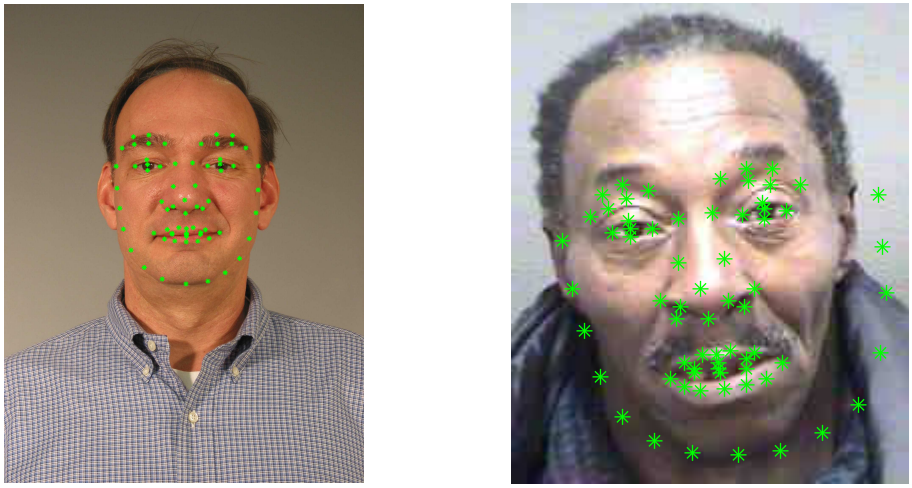
2.2 Preprocessing

The first steps of preprocessing are to annotate the facial images and correct for in-plane rotation. All images were processed with a script using the VeriLook 5.4 Standard SDK. Included in the SDK is the FaceExtractor component which extracts 68 facial feature points. These points include locations for the eyes, nose, mouth, and chin, and can be seen in Figure 2.7. The fit of the points was another consideration for including images in the experiment sets. If all 68 points were not detected in an image, it was automatically discarded. Using the eye centers from the detected feature points, the faces are rotated so the eyes are level in the image plane. The rest of the points are updated as well to be used later for separating the face into its subregions and extracting the face. More information on VeriLook can be found at the end of this chapter.

Figure 2.8 shows example facial regions of the chin, mouth, nose, nose tip, and eyes. A subset of the 68 points detected by VeriLook, shown in Figure 2.9, were used to extract facial regions. The eye regions are centered on the eyes (points *4* and *5*) with a width equal to the distance between the centers and a height equal to the vertical distance from the eye centers to the center of the nose (*12*). The nose tip region starts midway between the eye centers (*4*, *5*) and the nose center (*12*) and extends to midway between the nose center and the mouth center (*11*). The left and right boundaries correspond to the x-coordinates of the eye centers. The nose region starts at the y-coordinate of point *6* and extends to midway between the nose center (*12*) and the mouth center (*11*). The

	FRGC		MORPH		Pinellas	
	Avg size	*Exp size*	*Avg size*	*Exp size*	*Avg size*	*Exp size*
Face	482×486	480×480	195×196	200×200	246×248	250×250
Eyes	220×120	220×120	91×41	90×40	114×68	120×70
Nose	146×172	150×170	69×68	70×70	80×94	80×100
N Tip	220×119	220×120	91×47	90×50	114×64	120×70
Mouth	261×144	260×150	102×59	100×60	124×70	130×70
Chin	261×90	260×90	102×37	100×40	124×45	130×50

Table 2.2: The average size of the regions from raw images and the resolution the regions were resized to for experiments. The images used in calculations were those that had a face and all 68 points detected by VeriLook script: 39,250 of 39,328 for FRGC, 53,964 of 55,608 for MORPH, and 1,436,799 of 1,447,607 for Pinellas.



(a) (b)

Figure 2.7: Example feature points detected by VeriLook SDK a) good fit on FRGC image, b) bad fit on MORPH image

Figure 2.8: Example facial regions used for experiments, extracted from an FRGC image.

left and right boundaries are 5 pixels outside the x-coordinates of points *7* and *8* on the outline of the nose. The mouth region starts halfway between the mouth (*11*) and nose (*12*) centers and extends to halfway between the mouth center (*11*) and the chin (*2*). The left and right boundaries are midway between the x-coordinates of the mouth corners (*9*, *10*) and the closest points on the face contour (*1*, *3*). The chin region keeps the same left and right boundaries of the mouth region. The chin region starts midway between the mouth center (*11*) and chin (*2*) and extends 5 pixels below the chin tip (*2*). Table 2.2 show the average resolution of each of the facial regions for each dataset. Based on these calculations, an experiment size was chosen for each region and dataset. All face regions were resized to the experiment resolution before preprocessing continued and feature extraction began. The extracted face region for baseline comparison experiments extends over the entire are were points were detected, from the eyebrows to the chin. Prior to further processing an elliptical mask of neutral color is placed around the face to minimize the effect of the background on performance.

The next preprocessing step is to enhance the contrast of the image. Two methods are used. For texture-based features that will be extracted from a grayscale image, a simple histogram equalization is performed. The contrast enhancement procedure is slightly more complex for color-based features. In order to preserve the relative color information between color channels, the image is first converted to the L*ab color space. Illumination is stored in the L channel, while the other channels hold the color information. Histogram equalization is performed on the L channel and the image is converted back into RGB space. An example of the preprocessing steps for the face and one of the eye regions can be seen in Figure 2.10.

A local, patch-based approach is used for feature extraction with both the texture- and color-

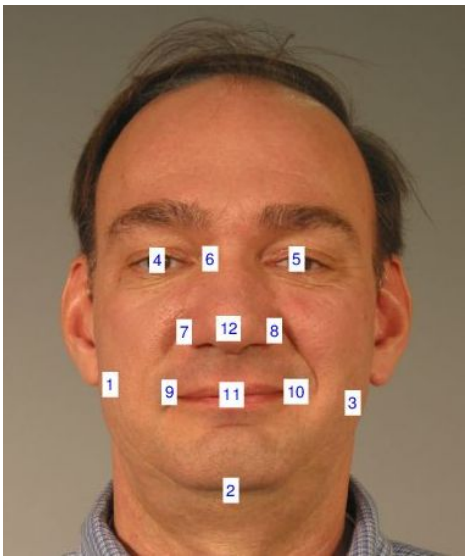


Figure 2.9: Point subset from VeriLook annotation used to extract facial regions.

based features. Each facial region image is divided into smaller images, called patches. Features are extracted from each patch and the feature vectors are concatenated to form the feature vector for the entire image. The MORPH and Pinellas experiments use a patch size of 10×10 pixels and the FRGC experiments use a patch size of 20×20 pixels. Since the FRGC images are of greater resolution, the patch sizes are larger so that the patch size relative to the image size is similar to MORPH and Pinellas. This also allows for a comparable number of features to be extracted from each of the datasets.

2.3 Feature Extraction Methods

Three main types of information can be found in 2D facial images: texture, color, and shape. Face images provide texture in both the skin and hair portions of the image. Color information can be gathered to provide hair, eye, and skin color. Moles and birthmarks can also be indicated by color. The outside contour of the face and position of the eyes, nose, mouth, and chin within the face give shape information that can be used for classification. The following chapters detail feature extraction methods and experiments performed in each of the three main categories of features.

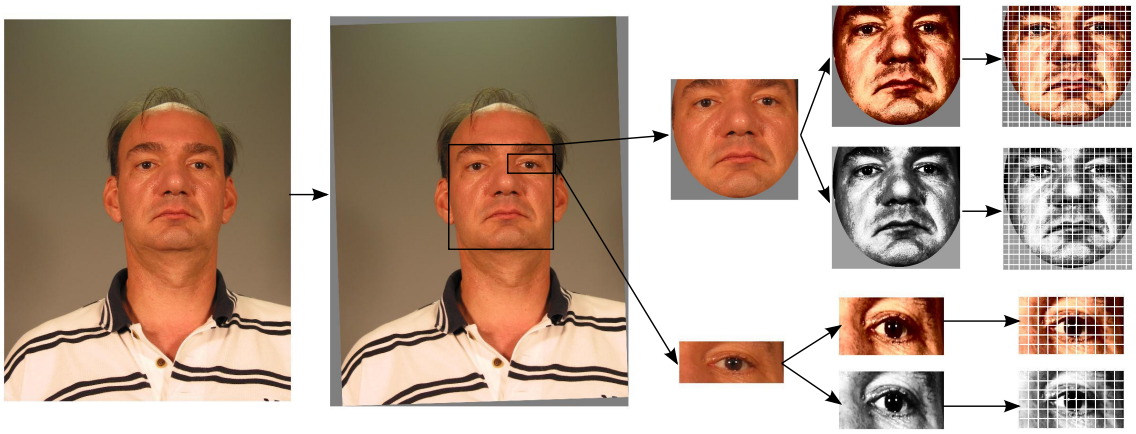


Figure 2.10: Preprocessing of a FRGC image following facial annotation. The original image is to the far left. The next image is corrected for in-plane rotation. The LEYE and FACE regions are extracted. The smaller images to the right represent the preprocessing steps for color (top) and texture (bottom) features for each region. The first images are contrast enhancement and resizing. The final images represent the patches used for feature extraction.

2.4 Feature Reduction Methods

Feature reduction techniques seek to eliminate noise and non-essential information from the feature vectors being used for classification. In this sense, feature extraction methods can be considered feature reduction techniques when the output of the extraction method is smaller than the original number of pixels in the image. Non-essential information could still be useful information, just not for the current problem. For example, information about a mole might be useful to help identify an individual, but not for gender classification. For this work, there is more interest in retaining features that encode gender or ethnicity information or even a combination of both, rather than identity.

Principal Component Analysis (PCA) is widely used for data compression and feature reduction. It is very useful for reconstructing data and is the basis for the well-known facial recognition algorithm "EigenFaces" [70]. PCA, known originally as the Karhunen-Loéve transform, finds the directions of maximum variance for a training set, regardless of class, using the eigenvectors of the covariance matrix.

For feature reduction using PCA, some of these eigenvectors are discarded; those corresponding to the smallest eigenvalues. The larger the eigenvalue, the more the variance represented by that eigenvector is present in the training set. In practice, researchers normally keep the eigen-

vectors that account for 95% of the variance. This can be determined by how many eigenvalues, from largest to smallest, are needed to have 95% of the sum of all the eigenvalues. The retained eigenvectors are used to build a projection matrix. Features are projected into the PCA space by multiplying by the projection matrix. The Eigen library [25] was used to implement PCA for this work.

Another type of feature reduction is Linear Discriminant Analysis (LDA). This method is designed for classification problems with multiple examples of each class. This is another transform, similar to PCA, which projects the features into a lower-dimensional space; however, LDA takes into account the different classes that are present in the data. In this transform, the variance between classes is maximized, while the scatter within each class is minimized. By incorporating class information into the training, LDA does better at discriminating between classes than PCA in most classification cases. The number of features retained is equal to the number of classes used in the analysis. In preliminary experiments on FRGC, using LDA for feature reduction resulted in a large increase in training time, a large decrease in features used, but only small increase in performance for some features. Performance on some actually decreased. For this reason, feature reduction was limited to PCA.

2.5 Classification Methods

2.5.1 k-Nearest Neighbor

The k-nearest neighbor (k-NN) is one of the simplest classification algorithms available to researchers. k-NN is a simple classifier which relies on the training data. No training is necessary before classification starts. A test sample is classified according to the class of its k closest neighbors. Each neighbor votes and whichever class has the most votes is the class assigned to the sample. In practice, k should be odd to avoid ties, although this is not enforced within the algorithm. The FLANN library within OpenCV was used for experiments with k-NN classification.

The closeness of the neighbors can be determined by any distance metric. The choice of distance metric often depends on the type of data and the area of the problem. Common distance metrics used for any real-numbered data are the Manhattan, Euclidean, and Maximum distance metrics. These are also known as the L_1, L_2, and L_∞ norms respectively. These three distance

metrics are derived from the Minkowski or p-norm measure, which can be defined as

$$L_p(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}},$$

where x and y are vectors with length n. Euclidean distance is the case where $p = 2$ and corresponds to the intuitive idea that the distance between two points (in two dimensions) is a straight line. Manhattan distance, also known as city-block or taxicab, is the case where $p = 1$. The more colorful names stem from visualizing the distance measure on city streets. Manhattan distance is the distance a cab would have to travel on the street around square city blocks to get between two points, assuming all streets are two-way. Maximum distance is the case as p approaches infinity. This simplifies to simply the maximum of the distances between each dimension. This is also known as Chebyshev distance.

Many of the features presented in this work will be in the form of histograms. Widely used distance measures for histogram data are the Chi-Square, Histogram Intersection, and Hellinger distances. The Chi-Square (χ^2) distance is a weighted form of Euclidean distance with differences between larger elements being less important than differences between smaller elements. This distance can be defined as

$$\chi^2(x, y) = \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}.$$

The Histogram Intersection distance measure starts out as a similarity metric, measuring the overlap between the two histograms. It is transformed into a distance metric by normalizing and subtracting from 1, as given by the follow equation:

$$HI(x, y) = 1 - \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\min(|x|, |y|)}.$$

The Hellinger distance, very similar to the Bhattacharyya distance is used in probability and statistics to provide a measure of similarity between two probability distributions. Hellinger distance between two discrete distributions is calculated as follows:

$$H(x, y) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{x_i} - \sqrt{y_i})^2}.$$

2.5.2 Support Vector Machine

Support Vector Machines (SVM) are a popular classification method in gender and ethnicity classification [5, 9, 31, 38, 44, 47, 60]. As a supervised machine-learning algorithm, the SVM requires labeled training samples. The underlying idea of SVM training is to find the hyperplane that best separates the training samples. In kernel-based SVMs, the samples are transformed by the kernel into a higher dimensional feature space. This allows more room for the best separation to be found. The algorithm searches for the best hyperplane that separates the data with the largest margins on either side of the hyperplane to the training data. The training samples used to define the best hyperplane are known as support vectors. These vectors are stored with the equation of the hyperplane for testing.

SVMs are inherently binary classifiers, but can be modified to work with multiple classes. The LIBSVM library [13], which was used for SVM classification, implements the one-against-one approach. If there are k classes, the multi-class SVM consists of $\binom{k}{2}$, or $\frac{k(k-1)}{2}$, binary classifiers. The final class-decision of the multi-class SVM is the class that 'won' the most binary decisions.

Experiments performed with SVM classification used linear kernels. Radial Basis Function (RBF) kernels were investigated since literature shows that RBF SVM have equal or greater performance than linear SVMs. Unfortunately, these SVMs are very sensitive to chosen parameters. Using a tool provided in LIBSVM and features from one image per subject, for a given subset of subjects, a grid search was performed for cost (C) and gamma (γ) parameters. The cost parameter is the penalty given for misclassifying a sample and γ is a kernel specific parameter. Features from all images were not used to increase the speed of the parameter search and to avoid over-fitting. After choosing parameters with this technique, test results were mostly of only one class, which indicates the chosen parameters were not in the correct parameter space to achieve optimal performance. For this reason, RBF results are omitted.

2.5.3 Artificial Neural Network

Artificial Neural Networks (ANN) are a supervised machine-learning algorithm developed to imitate the way scientists believe the human brain works. Individual neurons are highly interconnected and are the basic blocks which compose the brain. In an ANN, a neuron is modeled by a weight vector of its incoming edges and an activation function. Networks can have any number of

layers of neurons, but many have just three, an input layer, a hidden layer, and the output layer. ANNs use labeled training data. Training is an iterative process with labels being computed, an error function evaluated, and the weights for each neuron being updated. The final ANN models the relationship between inputs and outputs and can capture patterns present in the data.

Classification for multiple classes can be achieved by using an output layer of c neurons where c is the number of classes. Each class corresponds to a position in the output vector. The position that has a positive value indicates membership in the corresponding class, while all other positions are zero or negative. This is the method used in the classification experiments within this work. ANN classification was implemented using the machine-learning section of OpenCV 2.3.1 as well as the Fast Artificial Neural Network Library (FANN) [23, 52]. All ANNs used in this work are 3-layer with weights from training calculated by the resilient backpropagation (RPROP) algorithm. The RPROP algorithm is the default training method for OpenCV and is usually faster than classic backpropagation. Hidden layer sizes of 50, 100, and 200 were chosen for experiments to show performance with small and larger hidden layers.

2.6 General Experiment Setup

2.6.1 Performance Measures

In many classification experiments, the overall accuracy is used to report the performance. Overall accuracy is defined as the percentage of instances classified correctly regardless of the class or label for each instance. Another performance measure used in classification and machine learning provides a finer level of detail. A confusion matrix, or matching matrix, looks at both the predicted class and given class of an instance and places it into the correct position of the table. The rows of the table represent the given class while the columns represent the predicted class. This allows an observer to see the similarity between classes by looking at the misclassification rates of a given class to each of the other classes. Within the confusion matrix, the correct class-wise classification can be found along the diagonal. This is useful in instances when the test data is not balanced by class. An average class-wise accuracy would give a better idea of the performance of the classifier across the classes instead of the overall average which can be biased by an unbalanced set.

In many cases, a box plot of the class-wise accuracies will be shown. A box plot is a way to graphically represent groups of numbers. The box itself represents the first and third quartiles of

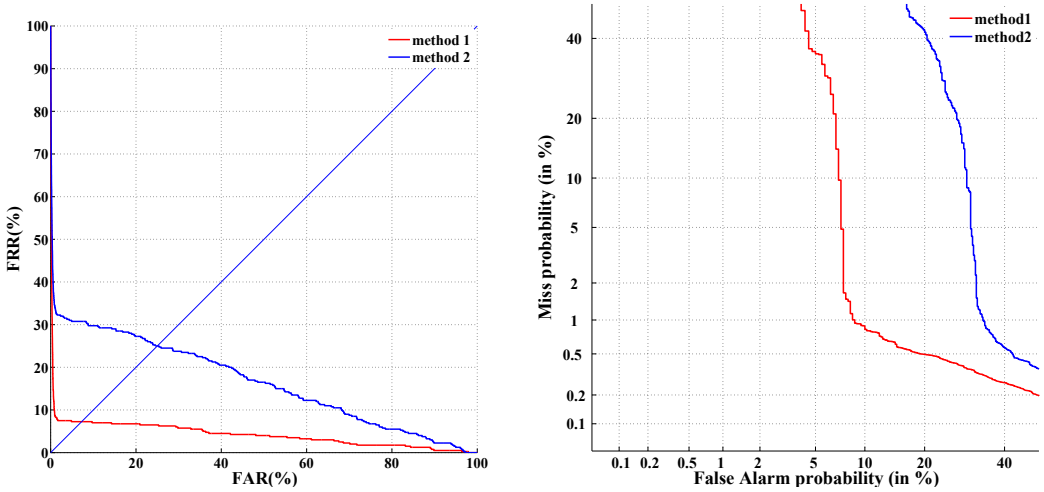


Figure 2.11: Example ROC and DET curves. (L to R): ROC, DET.

the data while the line inside represents the median of the data. The whiskers of the box are the largest and smallest values within 1.5 × IQR (interquartile range) which is the difference between the first and third quartiles. Points that do not fall within ±IQR of the box are marked as outliers and plotted as points on the graph.

In biometric verification/authentication problems, a curve known as the Receiver Operating Characteristic (ROC) is used as a measure of performance. In this problem, a threshold is used on the match score by the system to determine whether two samples are from the same individual or different individuals. The ROC curve is created by varying the threshold from its lowest to highest possible values. At each threshold, a certain number of individuals are accepted or ruled as the same person, when they should not be. These are known as False Matches or False Accepts. At the same time, genuine matches are ruled to be different people and rejected. These are known as False Rejects or False Non-matches. The axes on the ROC curve are the False Accept Rate (FAR) and the False Reject Rate (FRR). An example two ROC curves can be seen in Figure 2.11. The curve which is closest to the origin is considered the best. In this instance 'method 1' outperforms 'method 2'. Another measure to evaluate the performance is the Equal Error Rate (EER). This is the point on the graph where the FAR is equal to the FRR, in other words, the point where the curve intersects the diagonal in the graph. It is generally accepted that better systems will have a lower EER.

Another curve, the Detection Error Trade-off curve (DET), can be used to display the FRR

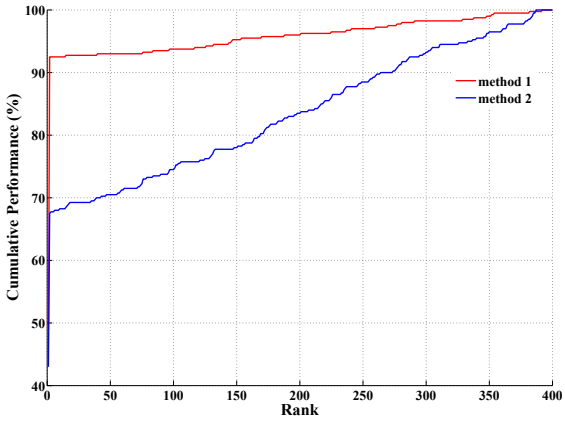


Figure 2.12: Example CMC curve

versus the FAR in a different format from the ROC. Instead of scaling the axes linearly, a logarithmic transformation is used on both the x- and y-axes. This results in a more linear trade-off curve than those seen in the ROC curves. An example can be seen in Figure 2.11. The EER can still be found by where the curve intersects a line going diagonally from the origin to the upper right corner. The curve which is closest to the origin is generally considered the best in these graphs.

In biometric recognition/identification problems, a curve known as the Cumulative Match Characteristic (CMC) is used as a measure of performance. The CMC curve is created by sorting the match scores for each given probe. The axes on the curve are the rank in the sorted list ($1 \ldots N$, where N is typically the number of entries in the gallery) and the performance of the system at the given rank. In other words, the performance at Rank-K is the percentage of probes where the true match is found in the first K entries of the sorted match scores. The higher the Rank-1 performance and the faster the curve approaches 1 or 100%, the better the system when comparing CMC curves. An example of two CMC curves can been seen Figure 2.12. In this figure, 'method 1' has a higher Rank-1 performance than 'method 2'. Both methods reach 100% very late, but 'method 1' would still be considered the better of the two, since its curve is higher than the curve for 'method 2' throughout the graph.

2.6.2 Cross-Validation Evaluation

Reliability and age experiments for color, shape, and texture are evaluated using a stratified five-fold cross-validation (CV) approach. In a five-fold CV experiment, the subjects are divided into

five parts, keeping the proportions of classes in each part approximately equivalent to the proportions of the whole set. In each fold, a smaller experiment takes place. Images, or features from images, corresponding to the subjects from four parts are used for training while the other set is used for testing. For each fold, the part used for testing changes. In this way, each subject appears in the test set only once in the whole CV experiment. Cross-validation is done based on subjects instead of images so that images of the same subject will not appear in both the training set and testing set for any particular fold when multiple images are used per subject. The performance of a CV experiment is normally the average overall classification performance of the folds. Confusion matrices will also be shown to view the distribution of classifications for each class.

2.6.3 Biometric Application Experiments

Application experiments will have a different set-up than cross-validation. Specific images are partitioned to be used solely for training the classifiers. A list of gallery images was created with two images per subject. A list of probe images was created with one image per subject using the same subjects as the gallery list. Subjects in the probe and gallery sets are distinct from subjects in the training sets. Classification results on the probe will be used to filter which gallery entries will be used in recognition and verification experiments. ROC and CMC graphs will be used to measure the performance of these experiments as well as the EER and Rank-1 performance. Classification results will be reported with confusion matrices and average class-wise accuracies.

The match scores for the base face experiment will be calculated using the VeriLook 5.4 Standard SDK previously mentioned during preprocessing. This is a commercial software development kit which supplies biometric functionality to its users. Functionality includes face detection and annotation, template generation, gender prediction, and template matching. All images that have templates generated for matching must pass a quality check, including lighting conditions and a minimum size requirement. More information can be found at the Neurotechnology website, `http://www.neurotechnology.com/vl_sdk.html`.

This concludes the description of data and methods to be used. The next three chapters will include results and discussion from reliability and age experiments for both gender and ethnicity classification using color, shape, and texture information. Color experiments will be discussed first.

Chapter 3

Color

Images in digital color formats are normally saved in three channels. Therefore, each location in the image, known as a pixel, has three values associated with it. The values for the pixel depend on the color space the image is stored in. The most prevalent color space, or more accurately color model, is the RGB space. In this color space the values for each pixel represent the amount of red, green, and blue present in that particular location in the image. Some color spaces were developed to be independent of the device the image was captured with while others were developed as more device specific. With the exception of the RGB color space, the color spaces mentioned here divide luminance- or lighting-type from the chroma or color information.

3.1 Feature Extraction Methods

Color has been a useful feature in the area of content-based image retrieval. One simple color representation is Local Color Histograms (LCH), used in previous work for gender and ethnicity classifications on pericular images [46]. As mentioned before, each color pixel in an image has an intensity value for each channel, red, green, and blue for example. For this work, in the RGB color space, the red and green intensity values are quantized into four levels. Using both values to count occurrences in a two-dimensional histogram, a feature vector is produced of length 4×4 or 16 elements per patch. These parameters were chosen based on findings in preliminary work [74]. Features from other color spaces will use the two channels that contain color information.

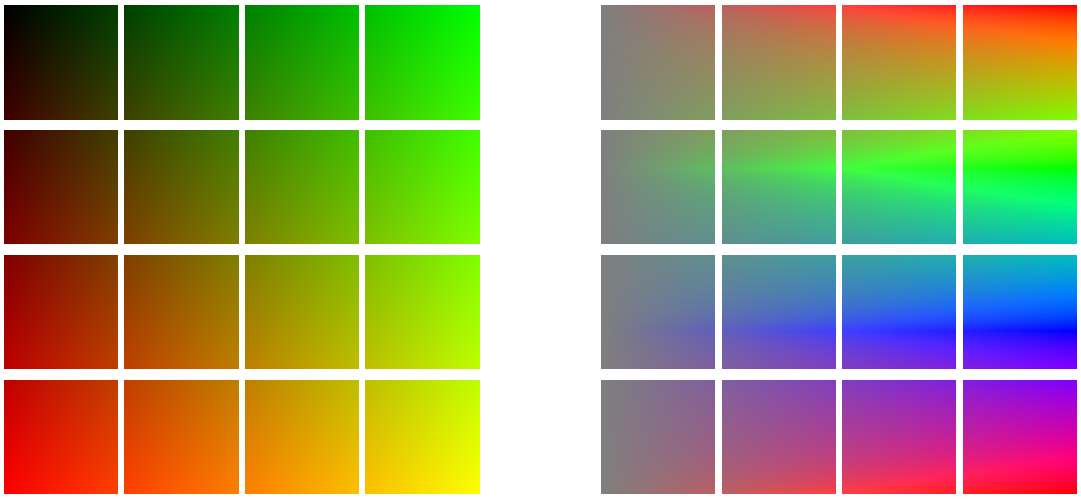


Figure 3.1: RGB and HSI histogram visualization. Left: RGB, RG Histogram. B at 0%. Right: HSI, HS Histogram. I at 50%.

3.2 Color Spaces

A brief description for each of the color spaces that are used in this work follows, as well as a visualization of the colors found in each histogram bin. Since the histograms involve two channels, while images are composed of three, the value in the third channel is held constant.

3.2.1 RGB

The RGB color model is an additive color model. The vales for red, green, and blue are added together to form all the other colors available in this space. This model is based on the light spectrum where light of different wavelengths combine to produce color. As mentioned before, RGB is not a specific color space, but rather a color model. In a typical RGB image, values are quantized to 256 different levels. Figure 3.1 shows the various colors that can be found in each bin of a 4×4 red-green histogram with no blue added. Other colors are possible with different levels of blue.

3.2.2 Device Dependent

3.2.2.1 HSI

The HSI color space is a cylindrical representation of points found in the RGB color model. The purpose of this arrangement is to more closely imitate how artists choose colors with a color wheel or a palette. The information for each pixel is divided into hue (H), saturation (S), and

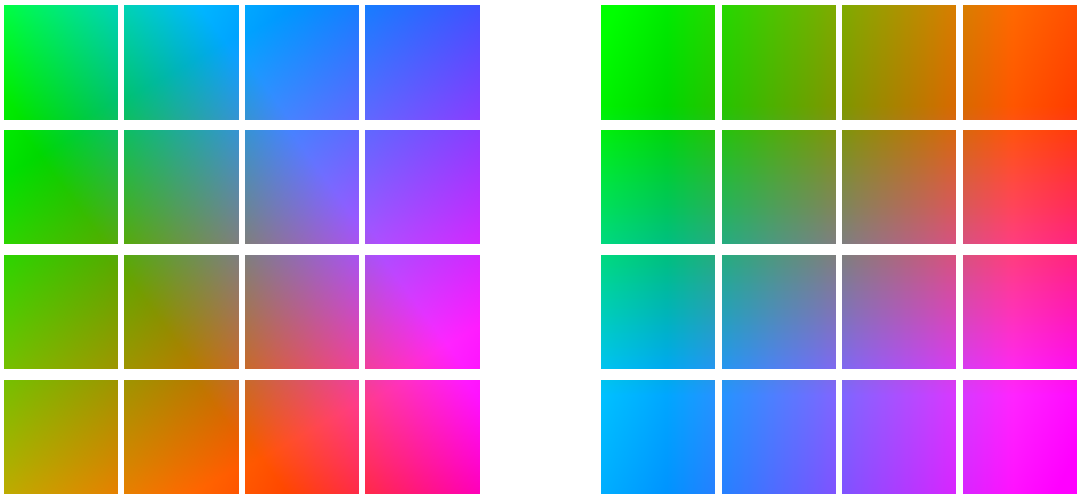


Figure 3.2: YIQ and YCbCr histogram visualization. Left: YIQ, IQ Histogram. Y at 25%. Right: YCbCr, CbCr Histogram. Y at 25%.

intensity (I) values. The intensity value gives the height on the cylinder, saturation the distance from the center, and hue gives the angle from the center of the cylinder. Other variations of this color space are the HSV and HSL spaces. The value (V) and lightness/brightness (L) channels are slightly different than the I channel of HSI, but mostly all represent the grayscale image with no color information. Hue information is roughly the same for all of the spaces, but the definition of saturation between the three color spaces is very different. Figure 3.1 shows the various colors that can be found in each bin of a 4×4 histogram with 50% illumination.

3.2.2.2 YIQ

The YIQ color space was originally developed for use with televisions. The main usefulness was that images could be sent to both color and non-color television sets because the luminance information was roughly captured in the Y channel. The I stands for in-phase and Q stands for quadrature, referring to components used in quadrature amplitude modulation. The I channel encodes color in the orange-blue range while the Q channel encodes color in the purple-green range. Figure 3.2 shows the various colors that can be found in each bin of a 4×4 histogram with 25% illumination.

3.2.2.3 YCbCr

The YCbCr color space is widely used for digital video storage and is related to the YIQ color space. The luminance information is found in the Y channel, just as it is in the YIQ space. The Cb and Cr channels encode the color information. The color information is stored as difference components. The Cb channel is the difference between the blue component of a pixel and a blue reference value, while the Cr channel is the same with the red component of a pixel. Figure 3.2 shows the various colors that can be found in each bin of a 4×4 histogram with 25% illumination.

3.2.3 Device Independent

The color spaces in this section were developed by the International Commission on Illumination (CIE). For this reason, these color spaces are often prefixed by CIE. To transform an image from an RGB space, the image must first be converted to the CIE XYZ color space. The XYZ space was one of the first mathematically defined color spaces, specified in 1931 by the CIE, and is the basis for the following spaces. Once there, conversions may be made to all the other CIE color spaces.

3.2.3.1 LAB

The intention of the CIELAB, or L*a*b*, color space was to produce a color space where a change in the color value should produce the same amount of change in the perceived color. It was designed to approximate human vision and was adopted by the CIE in 1976. The CIELAB space was used during the contrast enhancement of the color images. The luminance information is stored in the L* channel while the color information was stored in the other two channels. The a* channel holds information in the red to green range, while the b* channel holds information in the blue to yellow range. The reason this color space was not used for color histograms was that no hard boundaries could be found for the color channels a* and b*.

3.2.3.2 LUV

The CIELUV color space also attempts to gain perceptual uniformity as mentioned for the CIELAB color space. It was adopted by the CIE in 1976. This space is widely used in computer graphics and other applications which use colored lights, as it is an additive space. It is related to

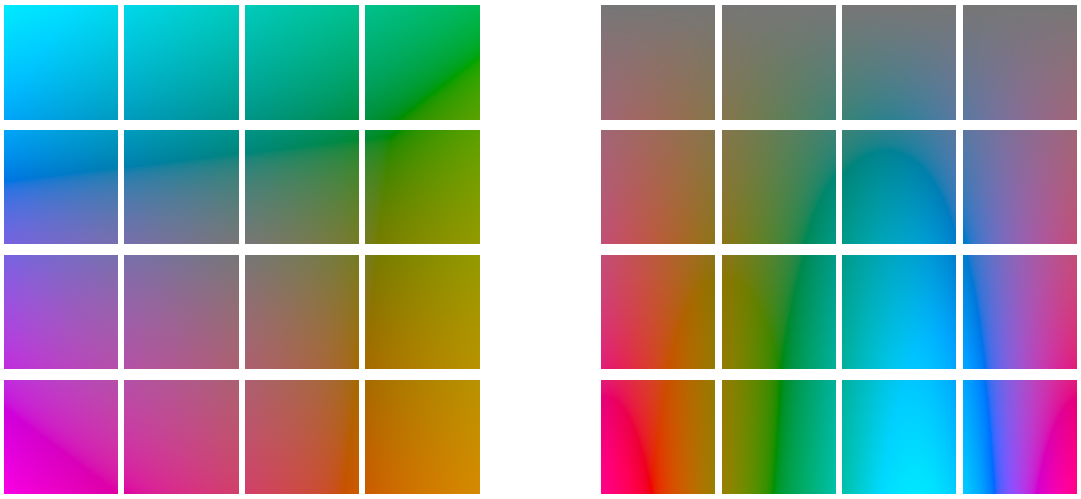


Figure 3.3: LUV and LCH histogram visualization. Left: LUV, UV Histogram. L at 25%. Right: LCH, CH Histogram. L at 50%.

CIELAB, in that each of the color spaces preserve the same L channel, but the chroma information is represented differently. Figure 3.3 shows the various colors that can be found in each bin of a 4×4 histogram with 25% illumination.

3.2.3.3 LCH

The CIELCH color space is related to the CIELAB color space. It is a cylindrical version of CIELAB where C is the chroma channel and H is the hue channel. As with the CIELAB and CIELUV color spaces, the LCH color space attempts to gain perceptual uniformity for changes in color values. Figure 3.3 shows the various colors that can be found in each bin of a 4×4 histogram with 50% illumination.

3.3 Experiment Setup

The experiments for this section will be performed on subsets of the FRGC and MORPH databases, using the color images. Table 2.1 shows the breakdown of the subjects according to gender and ethnicity for the subsets used. Features are extracted using the RGB, HSI, LCH, LUV, YCbCr, and YIQ color spaces. Classification of gender and ethnicity for all six color spaces will be performed over all regions using the k-NN classifiers (L_1, L_2, and L_∞). SVM and ANN classification will be performed on three of the color spaces, RGB, HSI, and LCH. The total number of color experiments

is 180 for each dataset and demographic. Five runs of a stratified cross-validation experiment are performed for each feature, region, and classifier combination.

3.4 Analysis

Global feature vectors were created to look at the color histograms within each region and color space. The two-dimensional histograms were flattened by taking rows from top to bottom. Figure 3.4 shows the average global features from the MORPH dataset. Many of the global feature vectors look similar over multiple regions. For all but the RGB features, the value in the majority of the bins is close to zero, indicating that colors within the region were limited. The range of skin tones is so small compared to the entire range of the color space, that a 4×4 histogram over the entire space may not be sufficient to capture differences due to ethnic group or gender. Since the global vectors for each region are similar, the colors found in each region are similar to each other as well. This indicates that the different regions of an individual's face are of a similar color. This is logical provided make-up, tattoos, and birthmarks do not dominate any specific region of the face. Differences between all combinations of regions will be calculated to further investigate the similarity of color over the entire face.

The features from the YCbCr color space are mostly concentrated in a single bin for the FRGC database, so this color space may not be very useful for classification purposes with the granularity used. It is possible that a 2D histogram with a finer granularity would capture more differences in the classes, but for now, it is not likely that the YCbCr features will work well for gender and ethnicity classification. In the MORPH dataset another very small peak is apparent in these features, which may increase the usefulness of the YCbCr space in differentiating between classes.

Differences between the global feature vector for each region were computed for each facial image to gauge the similarity of color over the face. The distance measures used were L_1, L_2, L_∞, Histogram Intersection, Hellinger, and χ^2. Means were calculated according to each gender and ethnicity class. Similar trends were seen with each of the distance measures. Figure 3.5 shows the differences between regions in the RGB and HSI spaces using the Histogram Intersection distance measure for the FRGC dataset. The LCH, LUV, and YIQ color spaces had results similar to the RGB graphs shown.

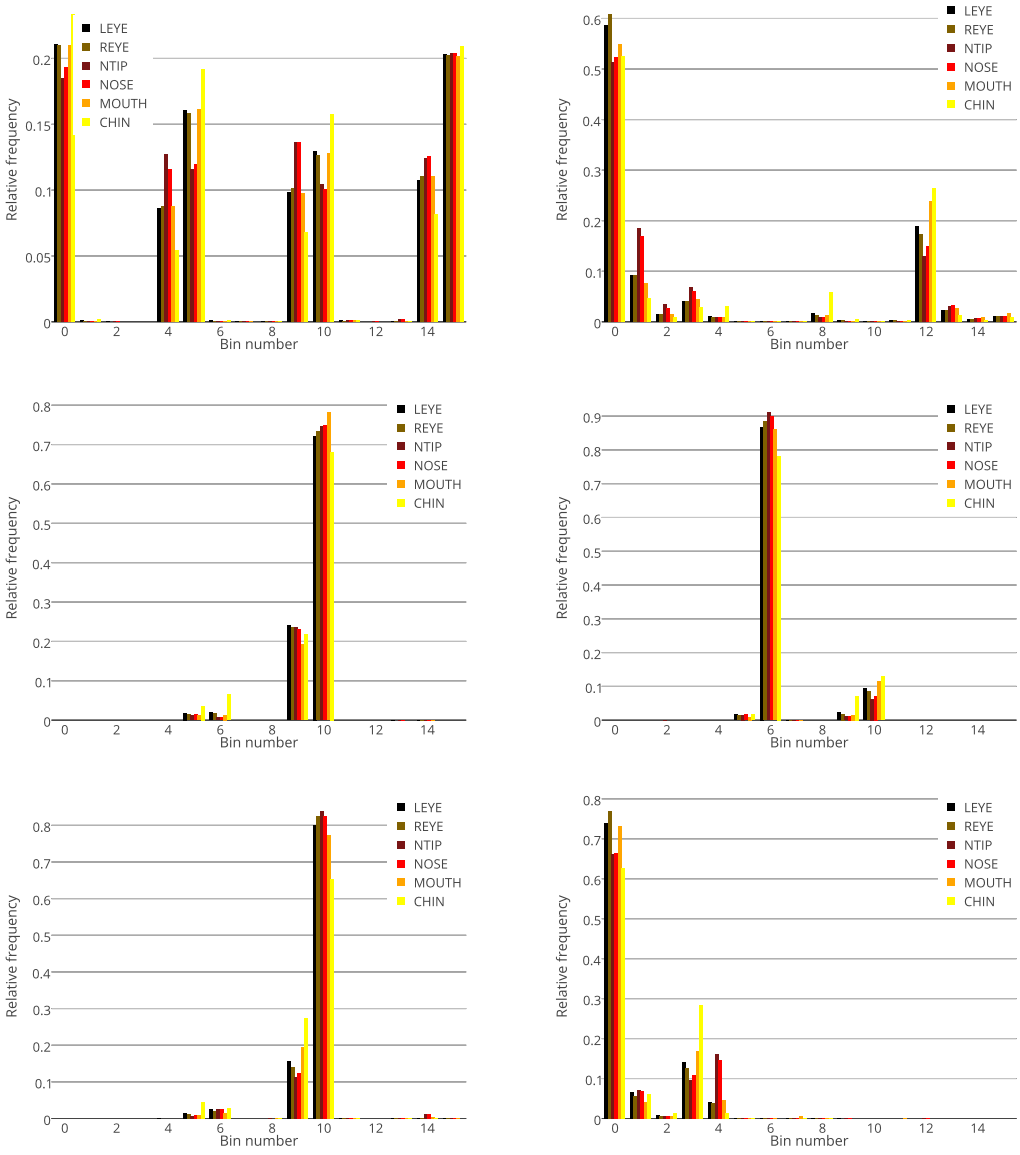


Figure 3.4: Average global feature vector for MORPH data. These are the normalized color histograms for each region. The x-axis corresponds to the bin in the flattened histogram. The y-axis corresponds to the relative frequency of values found in each bin. Top row (L to R): RGB, HSI. Middle row (L to R): YIQ, YCbCr. Bottom row (L to R): LUV, LCH.

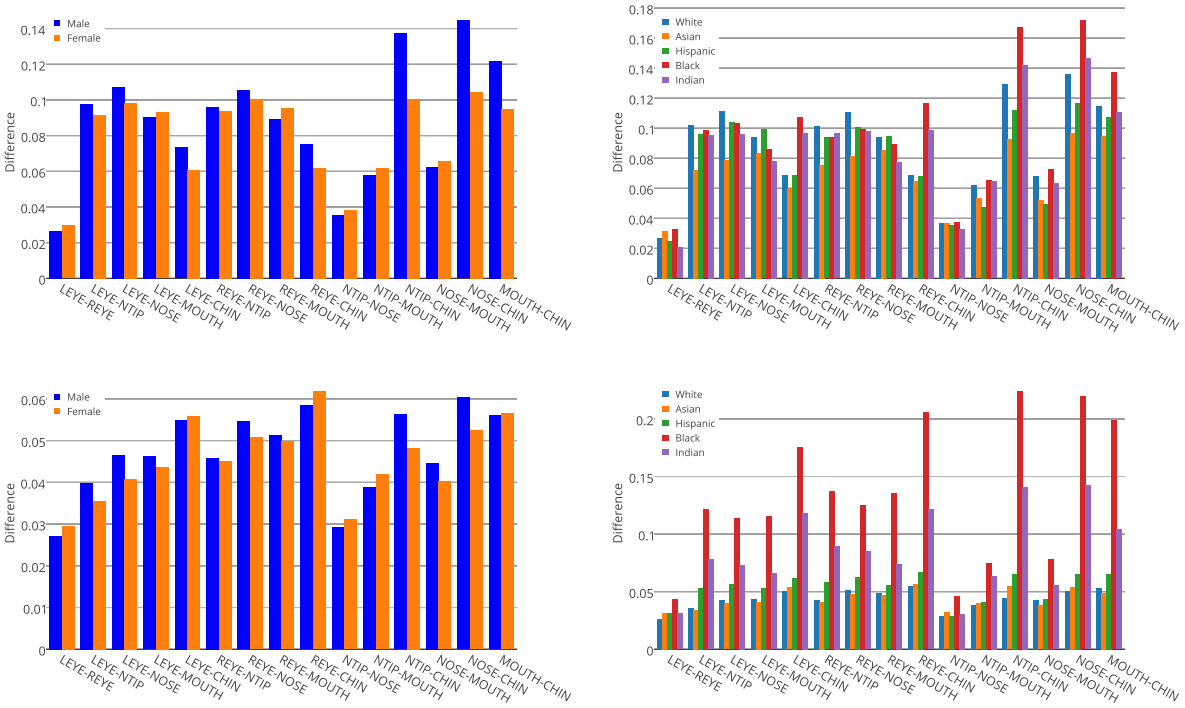


Figure 3.5: Mean difference between FRGC global color region vectors by demographic. Differences shown are for the RGB (top) and HSI (bottom) color spaces using the Histogram Intersection distance.

Overall, the differences between the regions were small, further indicating that color is fairly stable over the human face. The two pairs of regions most similar to each other in all the color features were LEYE/REYE, and NOSE/NTIP. The eye regions both hold an eye and for the majority of the population, both eyes are the same color. Congenital heterochromia iridis, two different colors in the eyes from birth, only occurs in approximately 6 out of 1,000 births [17], and in some cases is hardly noticeable. In this work, the nose and nose tip regions overlap. For these reasons, it makes sense that these two pairs of regions are the most similar out of all combinations.

The two pairs of regions least similar to each other in the RGB, LCH, LUV, and YIQ color spaces were the NOSE/CHIN and NTIP/CHIN pairs. In the HSI color space, the chin region was furthest away from all the other regions. The chin region is likely to have different colors for several reasons. In males, this region likely has facial hair which provides a larger difference than when comparing regions in females. For all classes, the chin is the only facial region that is on the border of the face, so some neck or clothing color could also be included, depending on landmark detection accuracy. The nose regions are not likely to have facial hair, unless the subject has a mustache, which would partially be included in the nose regions. The nose is also prone to illumination which may wash out the color information in the original RGB image.

In looking at the means of each gender class, the difference between the chin and any other region is noticeably larger for males than females. None of the other regions have a noticeable trend. Besides the chin region, color varies similarly between both males and females. This indicates that color is slightly more stable around the face for females than males.

In many of the HSI region comparisons and several in each of the other color spaces, the FRGC *Black* and *Indian* classes have much higher region differences than the other classes. This could indicate that color is not as consistent across the face in these demographics, but it could also indicate that the colors found in these regions lie close to the histogram bin borders. Small changes in color could produce very different histograms in that case. Another reason for the large differences could be that with just 10-12 subjects in the FRGC set, the mean is representing more of an individual case than the class as a whole. Color around the face is much more stable in HSI for *White*, *Asian*, and *Hispanic* classes, whereas in the *Black* and *Indian* classes it resembles the variance found in the other color spaces. In the MORPH dataset, the *Black* class spikes in the CHIN region comparisons which is likely the influence of the large majority of males in the class.

In the FRGC YCbCr space, the mean differences were much, much smaller than the other

Age	**Male**	**Female**	**Black**	**White**	**Hispanic**	**Asian**	**Native American**	***Total***
16-20	5,882	817	5,199	1,028	428	36	8	6,699
21-25	5,408	879	4,681	1,161	384	59	2	6,287
26-30	3,865	752	3,466	859	276	13	3	4,617
31-35	3,894	878	3,484	1,091	184	9	4	4,772
36-40	3,919	936	3,564	1,169	112	6	4	4,855
41-45	3,439	843	3,267	957	49	0	9	4,282
46-50	2,034	382	1,816	563	26	5	6	2,416
51-55	1,019	180	907	277	11	2	2	1,199
56-60	292	49	241	98	0	0	2	341
61-77	119	14	89	44	0	0	0	133
Total	29,871	5,730	26,714	7,247	1,470	130	40	35,601

Table 3.1: The number of images present per demographic label and age range for the MORPH dataset.

color spaces. The minimum distance between regions recorded was zero. This is the effect of what was seen in the global feature vectors, all colors concentrated in just one bin for a large portion of the subjects. In this color space, the mouth region was the most separated, likely due to the lips providing a color not found in the same histogram bin as the rest of the face. For MORPH, this color space was more similar to the other color spaces.

MORPH features will also be analyzed by age. The subjects' ages at the time of the image capture were recorded and provided with this dataset. The ages of the subjects ranged from 16 to 77 years old. The ages were grouped in age ranges of 5 years with the last group, 61-77, covering 17 years. This choice was made based on the low number of images present for that particular age group. Table 3.1 gives the number of images per class over the ages present in the experiment set. As the age of the subject goes above 60 years of age, the number of images present in the age range drops dramatically. The over 60 subjects are also not represented in three out of five ethnic classes. Since experiment sets were chosen without regard to age, it is most likely that performance on images of older subjects will be lower than that of younger subjects. A large majority of the training images will be on younger subjects. This may distort the training space if the features themselves are not age-invariant.

Discounting the 61+ group, the RGB features vary less than 5% across the age groups. The rest of the color spaces vary slightly more than 5% in pairs that include the chin or the mouth. This indicates that there is a difference of color in the chin and mouth regions with respect to age. As MORPH is a predominately male dataset, this could be facial hair. The eye regions do not

necessarily include the eyebrow, which would possibly increase the similarity. The RGB color spaces features seem to be slightly more stable with respect to age than the other color spaces.

The region differences in the younger age groups tend to be higher in the region comparisons with the most variance. The exceptions to this trend are the region comparisons involving the mouth. In looking at the mean differences per class, older males tend to have more variation in the comparisons with the mouth than younger males, while younger females have more variation in these comparisons than the older ones. The male variation is likely due to facial hair, while the female variation could be a result of lipstick or make-up. More work would be required to verify a relationship between age and facial hair or age and lipstick, so this remains a theory. One result of aging that could account for this is that lips thin over time [4], providing more skin tone and less lip color to help stabilize color in older females. As previously mentioned, most of the images are of younger subjects and a larger sample space allows for more variance within the classes.

3.5 Gender

3.5.1 Reliability

Nearest neighbor classification was performed using the L_1, L_2, and L_∞ distance measures to explore which features had classes that grouped together. PCA was performed on the features for feature reduction, retaining 95% of the variance. Figure 3.6 shows the 1-NN results using gender on the FRGC dataset. The values plotted in the box plot are the class-specific accuracies from five runs of each cross-validation experiment.

For most of the features in the different color spaces, the plots are fairly high and small, indicating that the performance for both *Male* and *Female* was good. The whiskers are all very short as well. The class-specific accuracies do not tend to vary much between cross-validation experiments. This results in two sets of bunched numbers. With just the two classes, these two bunches determine the height of the box plot, and the numbers that fall in the first and fourth quartiles will not fall far outside the box, resulting in short whiskers. This indicates that, most of the time, the features for the *Male* and *Female* classes have a close neighbor of the same class, and the results are not highly influenced by the random choice of training samples. The MOUTH and EYE regions perform the best out of the regions for several color spaces.

The class accuracies for the CHIN region and the YCbCr color space were more spread out

than the other experiments. For the YCbCr features, this is the result of all the features looking very similar and being concentrated mainly in one bin of the histogram. The nearest neighbor algorithm does not deal with ties. Depending on the implementation, the "closest" neighbor could be either the first neighbor the algorithm saw at that distance or the last. In this instance, samples in the *Male* class outnumber samples in the *Female* class, so the chances of being the first or last neighbor are better for the *Male* class. This results in most samples being classified as *Male*, giving the *Male* class a very high performance rate and the *Female* class a very low performance rate, which results in a large box plot. The CHIN region has a higher classification rate for *Males* than the other regions, but a lower rate for *Females*.

In a large majority of the experiments, the *Male* class had a higher classification rate than the *Female* class. With more samples in the training space, the likelihood of the closest neighbor in an overlapping area is greater for the *Male* class than the *Female* class. The exceptions to this trend are found in the YCbCr experiments. Using the L_2 distance measure, the *Female* class has a higher accuracy in every region. The L_1 and L_∞ distance measures in this color space both follow the larger trend of higher *Male* classification rates.

Figure 3.7 shows the results of gender nearest neighbor classification on the MORPH dataset using the L_1, L_2, and L_∞ distance measures. Performance on the MORPH dataset declines from the high-quality dataset results. The bias towards *Male* classification is more evident with larger discrepancies between the class-wise accuracies. In every instance, the class-wise accuracy is higher for the *Male* class than the *Female* class. This may not be due to the drop in image quality though. The ratio of female to male samples went from approximately 3:4 in FRGC to around 1:5 in the MORPH dataset, which could account for part of the difference, if not all. The RGB features have the smallest discrepancy between the class-wise accuracies, but it is still a large discrepancy.

Figure 3.8 shows the results of gender classification using the more sophisticated machine-learning classifiers ANN and SVM for both datasets. The RGB, LCH, and HSI color spaces were chosen for these experiments based on performance in 1-NN experiments and analysis. The performance gap between classes increased a little in the FRGC LCH and HSI results, but the performance increased overall. The RGB features increased performance and decreased the performance gap between genders. These results indicate that a more sophisticated classifier would be useful in further gender classification. For MORPH, performance in the RGB space improved for the most part by using a more sophisticated classifier. The most improvement was seen in the MOUTH region. The

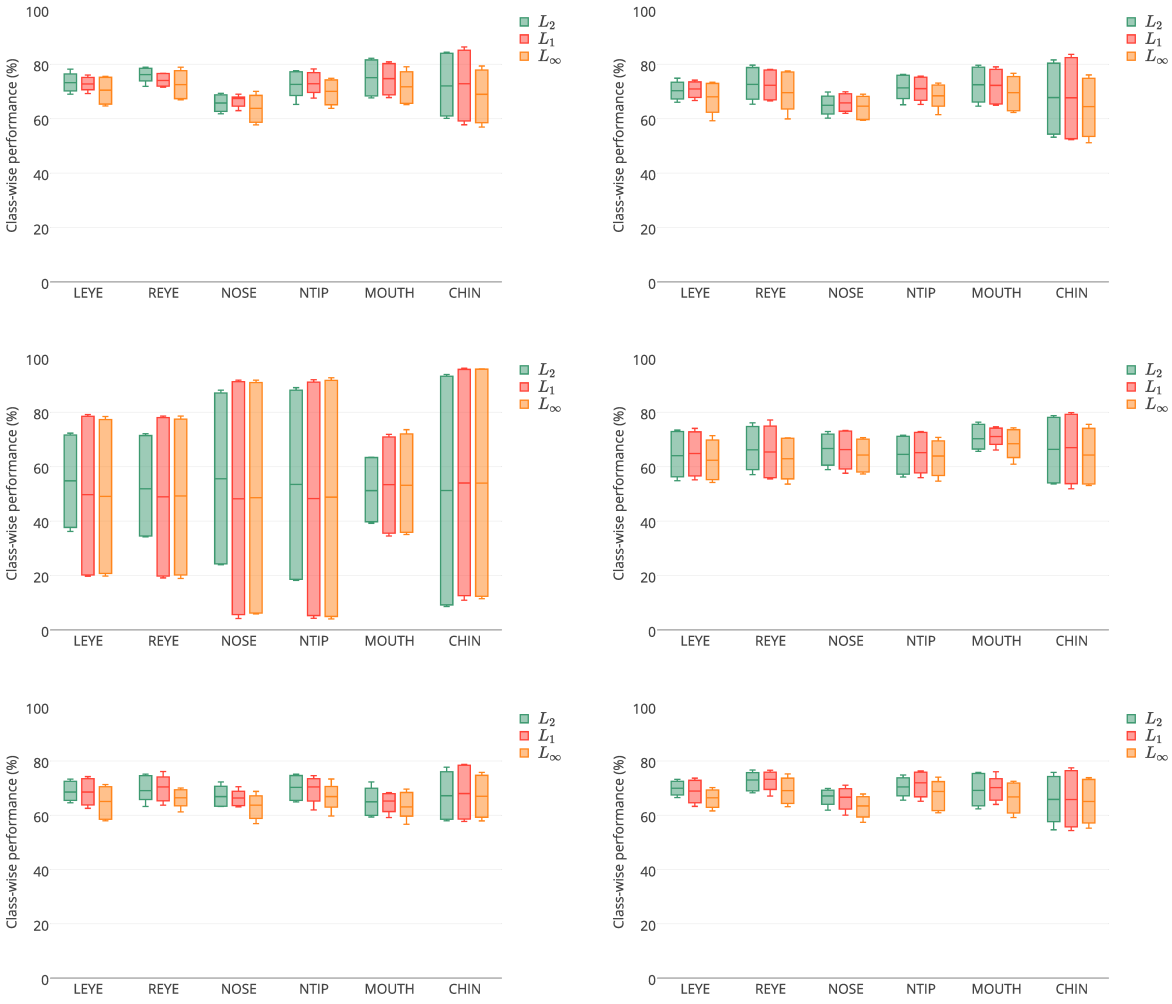


Figure 3.6: 1-NN gender classification on FRGC color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of each experiment. Top row (L to R): RGB, HSI. Middle row (L to R): YCbCr, YIQ. Bottom row (L to R): LCH, LUV.

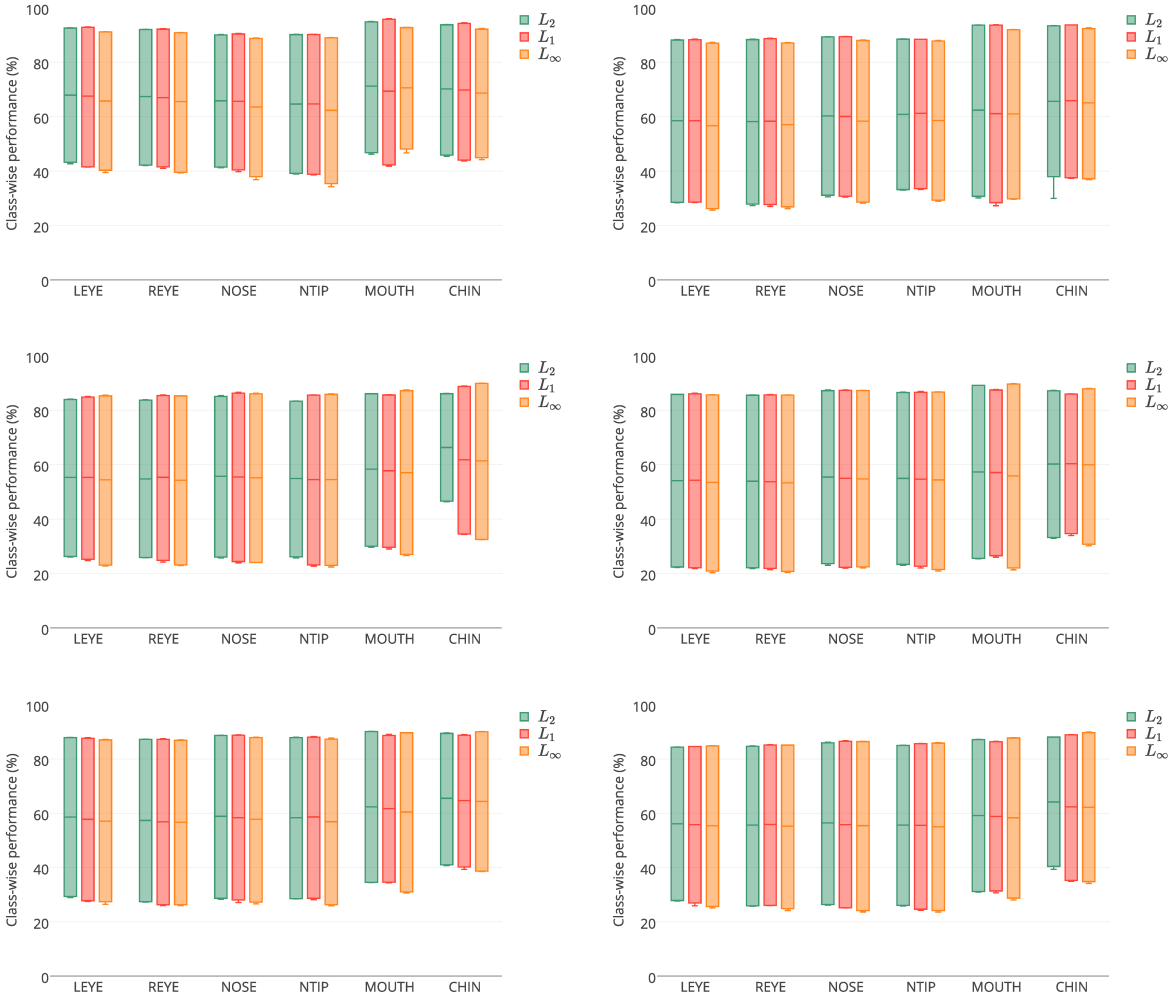


Figure 3.7: 1-NN gender classification on MORPH color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of each experiment. Top row (L to R): RGB, HSI. Middle row (L to R): YCbCr, YIQ. Bottom row (L to R): LCH, LUV.

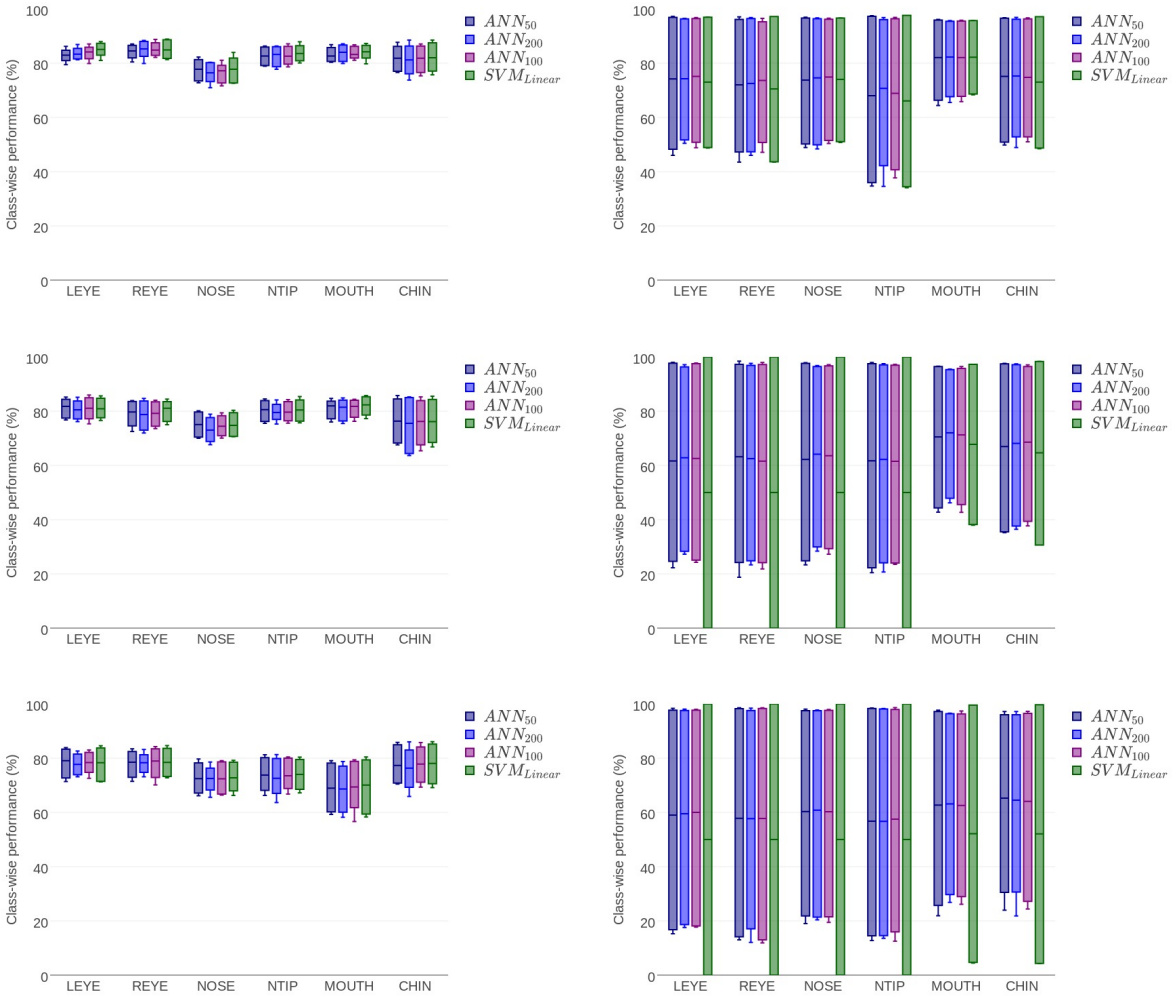


Figure 3.8: ANN and SVM gender classification on color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: RGB, HSI, LCH. L to R: FRGC, MORPH.

Figure 3.9: Easy and hard subjects in color gender classification on MORPH images. The subject on the far left was never misclassified. The other subjects were misclassified in over 95% of the color experiments. The *Male* in this section was mislabeled as *Female*.

NTIP is the only region that did not improve overall. The class performance of one class increased while the other decreased. That is actually the case for most of the LCH and HSI experiments as well. The ANN and SVM classifiers netted no overall improvement of performance in these spaces, with the exception of the MOUTH region in the HSI color space. The linear SVM results in the all the MORPH LCH and MORPH HSI upper face experiments where the box ranges from 0 to 100, show that these genders are not linearly separable. The poor performance also indicate that many of the features in these experiments are overlapping. MORPH results indicate that SVM and ANN classifiers would be most useful with RGB features or in HSI MOUTH experiments.

Over 80% of the MORPH images were misclassified in less than 25% of color experiments. One image was never misclassified in these experiments. No subjects were misclassified in all color gender experiments, but eight were misclassified in over 95% of the experiments. Images corresponding to these subjects can be seen in Figure 3.9. The classifiers were able to catch one error in the metadata. The *Male* image in the right grouping of the figure, misclassified over 95% of the time, is mistakenly labeled as *Female*. Taking that into account, that subject actually had a correct classification most of the time. These images suggest that subjects in the *Black Female* group are the hardest to predict gender on using color information.

Table 3.2 shows the results of baseline experiments on full face for the FRGC and MORPH datasets. Included are the results of gender classification using the VeriLook SDK. In the FRGC dataset, the eye regions, using the ANN classifiers, fall approximately 10% below the VeriLook

	FRGC			*MORPH*		
Classifier	Male	Female	Average	Male	Female	Average
ANN_{100}	99.95	0.13	50.04	96.07	61.68	78.88
SVM_{Linear}	92.39	89.50	90.94	96.84	58.82	77.83
L_2	79.48	69.22	74.35	87.45	25.62	56.54
VeriLook	92.72	98.45	95.59	99.14	67.75	83.45

Table 3.2: Gender performance using color and full face. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment. VeriLook refrained from predicting gender on 2.75% and 0.89% of male and female images respectively in the FRGC dataset. In the MORPH dataset it predicted no gender on 0.5% and 11.26% of male and female images.

gender classification accuracy. Following the same feature extraction but with face images, the gap is lessened to 5% with the best classifier, but the regions still perform less favorably than the whole face. In the MORPH dataset, the MOUTH region comes within 1-2% of the VeriLook performance. This indicates that, with images of a lower resolution, the color of the mouth region holds comparable gender information to the entire face.

3.5.2 Age

Figure 3.10 shows classification performance by age groups for the RGB space using 1-NN with Euclidean distance for classification. Other classification measures and color spaces give similar performance with exceptions noted below. The graphs show less variance on the top of the box suggesting that color in one of the classes is less impacted by age than the other. This would be the *Male* class in this instance. Again, the *Male* class does have the highest representation, which given any overlap, would provide that class an advantage.

There does not seem to be an age that is easiest to classify by gender overall. Younger females are easier to classify in the eye regions than older females. This is true in face experiments as well. In the eye regions again, there is no definite trend in the male class. For the nose regions, an erratic downward trend is noticed in the class-wise accuracies as age increases, suggesting that the color of the nose is not age invariant and behaves similarly for both males and females. The mouth region has a definite peak for both male and female performance between 26 and 45 yrs old. This suggests that gender classification based on color is more difficult for younger and older subjects than middle-aged subjects in the mouth region. For the chin region, there is a small upward trend for male classification, but the trend is down for the female class. This indicates that 21-35 year old

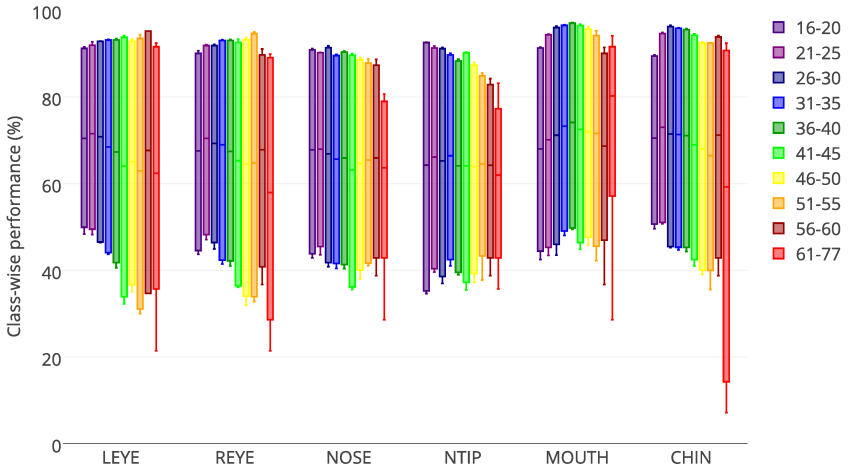


Figure 3.10: Gender performance by age group for features in the RGB space using 1-NN classification (L_2 distance).

males are easier to classify by the chin region than others.

It seems as though all regions are impacted by age either in one class or the other. The regions that are least impacted by age are the nose, nose tip, and mouth regions. The most impacted regions, those with the most variance, are the eye regions, not for the *Male* class, but definitely for the *Female* class. This same trend is seen in full face experiments.

Over all k-NN classification experiments, all the color spaces perform similarly. Some differences occur in the RGB, LCH, and HSI color spaces using the ANN and SVM classifiers. With the ANN classifier, the variance for females is more pronounced in the HSI space. In both ANN and SVM classification, more of an upward trend is noticed for females as age increases in the NTIP and MOUTH regions in the HSI and RGB color spaces. As females age, less color variance is noted in these regions resulting in better performance. The LCH and HSI color space features perform similar to 1-NN, except with the SVM classifier. In the eye and nose regions, the classifier does not get enough separation in the training set and ends up predicting all *Male* class.

3.6 Ethnicity

3.6.1 Reliability

Figure 3.11 shows the results of ethnic nearest neighbor classification on the FRGC dataset using the L_1, L_2, and L_∞ distance measures. The boxes in the ethnic results are much larger than the ones seen in the gender results. This indicates a wider range of class-wise accuracies. One or two classes may have good performance, as seen by the upper portion of the plot, but the other classes are not well recognized, extending the plot close to the zero line. The anomaly in these graphs is once again the performance of the YCbCr features, with outliers that fall well above the box plots. With 25 class-wise accuracies to plot, when one class performs very well and the others do not, the third quartile falls within the second-best group of accuracies. With a small IQR from the poor performance of the other classes, the best performances will be outliers on the graph.

For most of the experiments, the class with the best performance is the *White* class, followed by the *Asian* class. Since the *White* and *Asian* classes have the largest number of samples present in the experiment set, it makes sense that they would have the best class-wise accuracies. With the L_2 distance measure on YCbCr features, the best performance is on the *Indian* class, followed by the *White* class for the nose regions and the chin. As previously discussed, not much variance is found with the YCbCr features, and since the trend is only present in one of the three distance measures it very likely deals with the peculiarity of the data and not differences within the ethnic groups.

In most of the experiments, the class-wise accuracies of the *Hispanic*, *Black*, and *Indian* classes are less than 20%. In the LUV and YIQ and the eye and nose regions, the class-wise accuracy for the *Indian* class reaches up to 46% performance. The best accuracies for the *Black* class are found in the mouth region using the LCH features, but only reach 30%. The *Hispanic* class has a consistently low class-specific accuracy, rarely exceeding 15%, but it is less likely to have a 0% accuracy than the *Black* class. This indicates that there is a small subset of the *Hispanic* class which is more distinguishable than the rest, but also more generalized than just one person. The rest of the class, however, overlaps with another class. These results indicate that different regions might be better for ethnic classifications for different ethnic groups. With these three classes, the number of subjects was 15 or less, so any conclusions drawn from this information may not generalize to the larger problem.

Figure 3.12 shows the results of ethnic nearest neighbor classification on the MORPH dataset

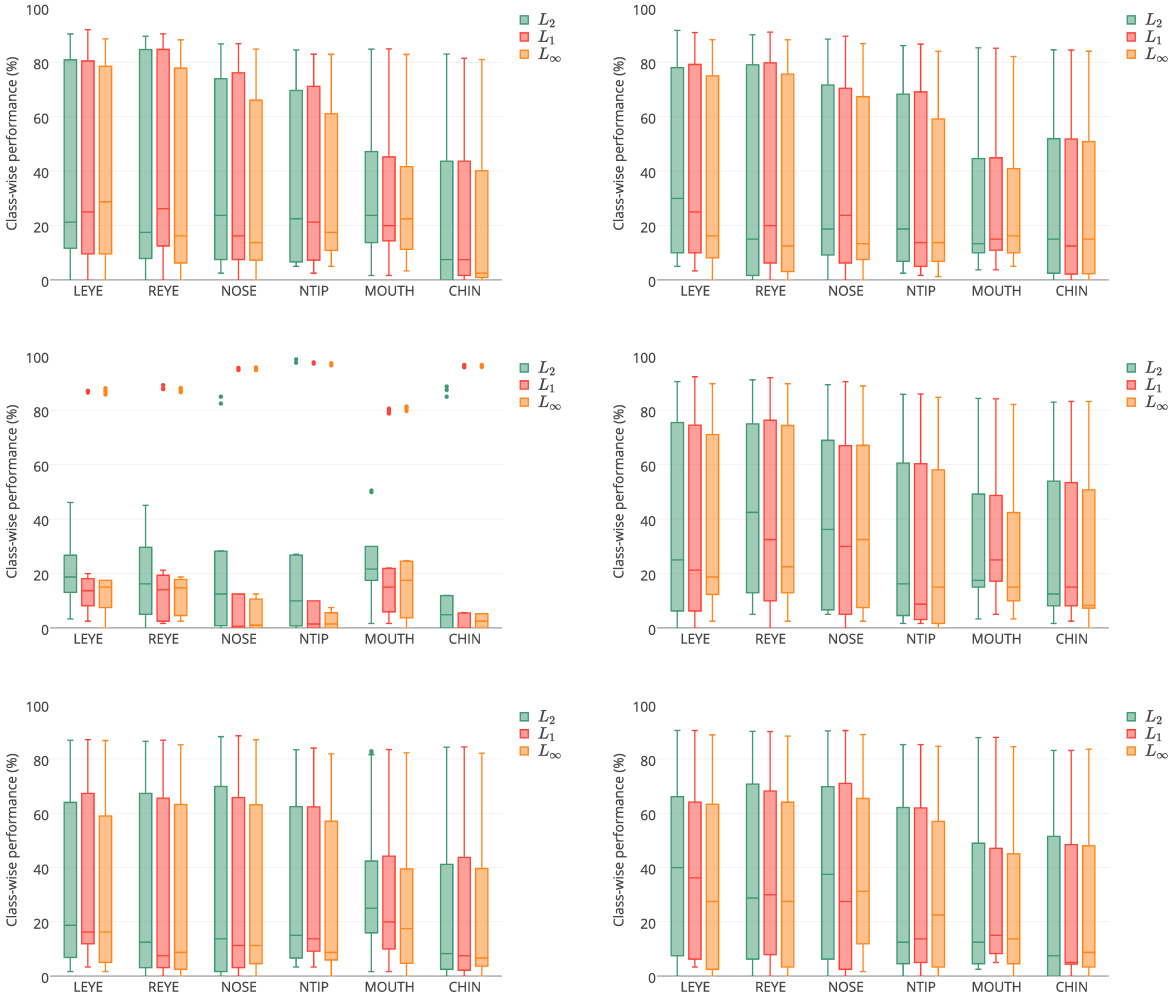


Figure 3.11: 1-NN ethnic classification on FRGC color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of each experiment. Top row (L to R): RGB, HSI. Middle row (L to R): YCbCr, YIQ. Bottom row (L to R): LCH, LUV.

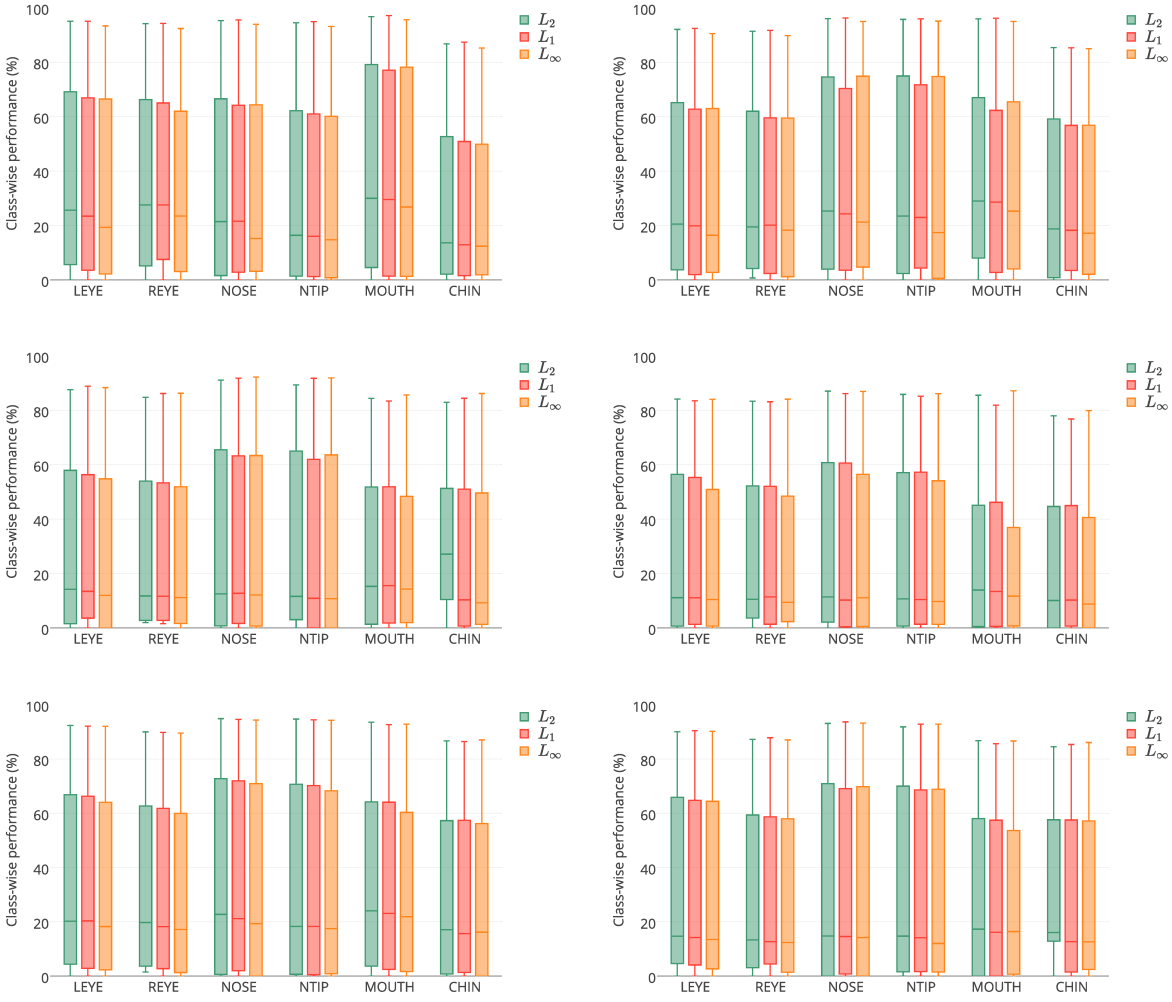


Figure 3.12: 1-NN ethnic classification on MORPH color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of each experiment. Top row (L to R): RGB, HSI. Middle row (L to R): YCbCr, YIQ. Bottom row (L to R): LCH, LUV.

using the L_1, L_2, and L_∞ distance measures. With this dataset, the best and worst performing classes change. The class with the best class-specific accuracies is the *Black* class, followed by the *White* and *Hispanic* classes. This is partially due to the number of samples for each of these classes. The remaining two classes have less than 100 subjects present in the experiment set.

The *Black* class has the highest class-specific accuracies of the MORPH ethnic experiments with accuracies falling between 85% to 95% on average. The best results are found using the RGB or HSI features for all of the regions. The best class-wise accuracies for this class are in the MOUTH region using the RGB and HSI features. Performance in this region ranges up to 97%. The next best region performance is around 95% with either a nose or eye region depending on the color space and the distance measure.

The class with the next best performance is the *White* class. The RGB and HSI color spaces produce the best results for this class, reaching 79% in the MOUTH region. This class has a larger variance than the other ethnic classes with the majority of class-wise accuracies falling between 50% and 70%. After the MOUTH region, the next best performing regions are the nose regions. The best performance on these regions is between 70% and 75% with the HSI and LCH features.

The highest class-specific accuracies for the *Hispanic* class can be found using the RGB, HSI, and LCH color space features for most regions. The best performance for this class in 1-NN experiments was approximately 30% using RGB and HSI features in the MOUTH region. On average, the class accuracy for *Hispanics* was between 10% and 25% in these experiments.

The best performance for the *Asian* class can be found using the RGB, LUV, and HSI spaces. This class has a fairly erratic performance indicating there is a lot of overlaps with another class and is highly dependent on the training data. Class-wise accuracies for the *Asian* class range between 0% and 10% for the most part. The features that most consistently achieved a non-zero class accuracy were found in the RGB space. The best region was different for each distance measure.

With only 13 subjects, classification on the *Native American* class is very poor, most often with no correct classifications. In less than 30 out of 108 gender experiments, color features managed to achieve at least one correct classification for this class over each of the five CV runs. These experiments were concentrated in the eye and nose regions with a few using the MOUTH region. Several color spaces were represented within these experiments as well.

Since the best ethnic performance is found for most of the classes in the RGB and HSI color spaces, these color spaces will be investigated using ANN and SVM classifiers. The hope is that with

Classifier	White	Asian	Hispanic	Black	Indian	Average
ANN_{100}	99.95	0.13	0.00	0.00	0.00	20.01
SVM_{Linear}	95.73	82.76	5.33	11.67	26.50	44.40
L_2	84.30	56.01	3.67	9.17	10.75	32.78

Table 3.3: Ethnic performance using color and full face, FRGC. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment.

Classifier	White	Asian	Hispanic	Black	Native American	Average
ANN_{100}	87.82	0.00	36.17	97.96	0.00	44.39
SVM_{Linear}	84.98	11.59	39.99	97.89	0.00	46.89
L_2	66.63	4.76	26.77	97.28	4.09	39.50

Table 3.4: Ethnic performance using color and full face, MORPH. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment.

a more sophisticated classifier, class-wise accuracy will increase. The global feature vectors for these spaces show values in approximately half of the bins. This is different from the YIQ, YCbCr, and LUV color spaces with values in only a third of the bins. It is similar, however, to the global features in the LCH color space. On the assumption that the less sparse the feature vector, the better the classification, the LCH color space will also be examined with the ANN and SVM classifiers.

Results of ethnicity classification using the RGB, HSI, and LCH color space features with the ANN and SVM classifiers are shown in Figure 3.13. Most of the experiments increased performance on the two most represented classes in the dataset, but did little to improve performance of the other classes. In most of the MORPH experiments, the class-wise accuracy for these classes decreased to 0%. This indicates that these classes do not have enough representation and that the samples that belong to them overlap with the larger classes.

Images from 30 subjects were never misclassified according to ethnicity. All of the subjects were males from the *Black* class. This is logical since this class has the largest representation and the largest classes have performed the best in the classifiers. Images from 20 subjects were misclassified by every run of every experiment. These subjects can be seen in Figure 3.14. With only 13 subjects present in the *Native American* class and 10 of them being misclassified in every experiment, there is a problem with this class. Either it is indistinguishable from another class based on color, or it is lacking the training samples needed to created an accurate representation of the class. The problem still exists for the *Asian* class, but not as badly with only 10 out of 50 subjects misclassified in all experiments.

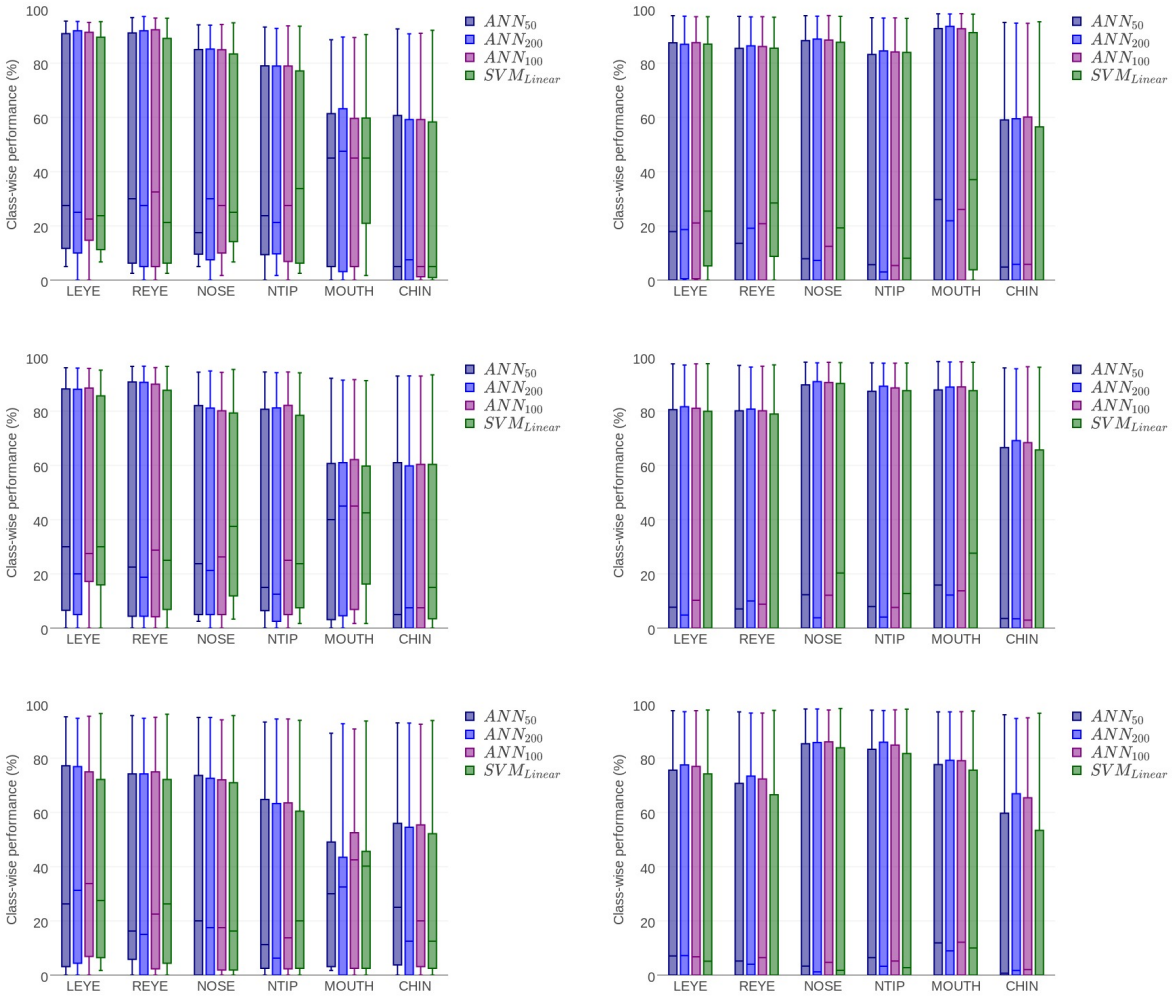


Figure 3.13: ANN and SVM ethnic classification on color features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: RGB, HSI, LCH. L to R: FRGC, MORPH.

Figure 3.14: Hard subjects in color ethnicity classification on MORPH images. The subjects to the left are from the *Asian* class. The subject to the right are from the *Native American* class.

Tables 3.3 and 3.4 show the results of ethnic classification on the full face using RGB color space features. In the FRGC dataset all regions but the CHIN region performed similarly on the best performance for each region with all the classifiers. The best CHIN performance was 10% below the face results. In the MORPH experiments, the MOUTH region compares favorably to the full face results with the eye regions slightly below. This indicates that the MOUTH and eye regions hold a comparable amount of color ethnicity information as the whole face.

3.6.2 Age

Figure 3.15 shows ethnic performance broken down over the age groups for MORPH experiments. The 41-45, 56-60 and 61+ age groups do not have subjects from each ethnic class in them, as seen in Table 3.1. The 41-45 age group has no *Asian* subjects in it. The 56-60 age group is missing *Asian* subjects as well as *Hispanic* subjects while the 61+ group only has subjects from the *Black* and *White* classes in it. If a class was not present in the age group, no value was plotted for the class-wise accuracy. This shows an improved performance for these age groups since the missing classes normally have poorer performance.

Performance on the best performing class does not vary much with age. There is a trend on most of these experiments for the next best performing classes. Excluding categories that do not have all age groups present, ethnic classification is more accurate on subjects under 30 years of age than older subjects in the second best performing class. The exceptions to this trend lie in the RGB features. Ethnic performance across the age groups varies less in this color space and does not show a definite trend. For all the others, the performance drops off as the subjects age until the spike in the 41-45 group. From there it depends on the region. In the CHIN region, ethnic performance on the 46-55 age groups increases again, but rarely more than the best performance on the younger subjects.

It seems as though all regions are impacted by age either in one class or another. The most heavily affected region for ethnic classification is the CHIN. Performance on all classes in this region show variance over age. In the other regions the best performing class is not as affected by age as it is in the CHIN region; however, the other classes do show some variance. The amount of variance per region is dependent on the color space. Features from the RGB space show less variance with age than the remaining color spaces. The most stable region with respect to age in that space is the MOUTH.

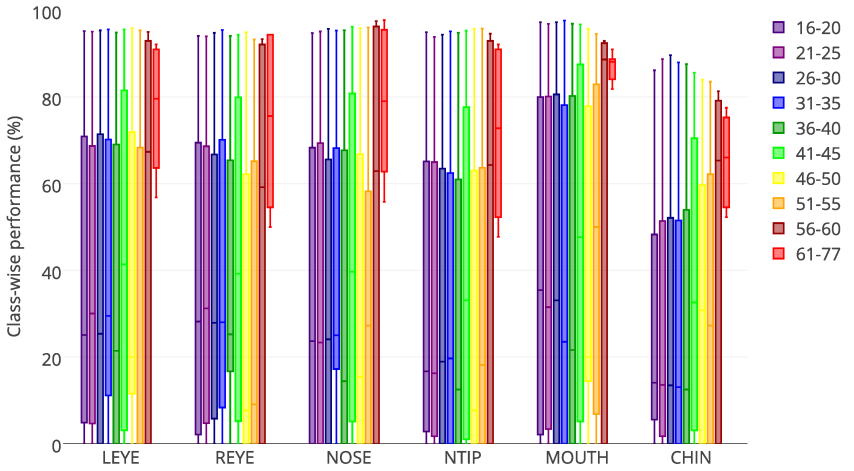


Figure 3.15: Ethnic performance by age group for features in the RGB space using 1-NN classification (L_2 distance).

3.7 Conclusions

Overall color is fairly stable around the face. The main color information comes from skin and is similar across the face. The region that is most likely to have different color information than the others is the CHIN region, very likely due to the presence of facial hair in some males.

Upon further analysis, coarse 2D histograms covering the entire space may generalize color information too much. The range of possible skin tones is very small compared to the entire color space represented by the color histogram. Future work will most likely use a finer histogram over a smaller portion of the color space. Colors detected outside this area could indicate make-up, face paint, glasses, or really anything that could be on the face and be a different color. This could help give a measure of confidence. If all detected colors fall within the colors of interest, the classification could proceed. If a certain percentage fell outside the colors of interest, a flag could be raised or a low-confidence value returned. Features from different color spaces could also be combined to improve color representation.

In high quality images, color information can be used to predict gender correctly over 80% of the time using the eyes, nose tip, and mouth regions. It is not as reliable for the lower-quality images, having between 75-85% average class-wise accuracy. In this dataset, the mouth provides the most reliable results, comparing favorably to face. The color of the eye regions performs fairly well which agrees with previous color experiments [46]. Based upon this, the conclusion is that color

information is fairly reliable for gender classification. Color information is not as reliable in ethnic classification as it is in gender classification for the experiments included within, but the mouth and eye regions compare favorably to to the face. This suggests that better representation is needed, both in training samples and feature extraction, since color can be major indicator for human-based ethnic classification. It also suggests a need for better ethnic labels.

Ethnicity classification is, by nature, a hard problem. Ethnicity is based on how an individual perceives either himself or others. This perception may not coincide with how another person would perceive the same individual. Most databases rely on self-reported ethnicity which gives no measure of consistency between ethnicities on the various subjects. Two people with the same mixed heritage might identify themselves with different ethnic groups. Add that to the problems of non-binary classification and less representation per class, and one can see how ethnicity results tend to be worse than those for gender classification.

All regions and colors are impacted in some way by age. The most affected regions for gender are the mouth and chin regions for both genders and the eye regions for females. The most impacted region in ethnic classification is CHIN. The color space that was least affected by age was the RGB color space. The least impacted region in gender is nose. The least impacted regions in ethnicity are the mouth and nose tip regions. This suggests that regions covering the nose are the most stable regions with respect to age in both gender and ethnic classifications.

Based upon the results and analysis from this chapter, features from the RGB color space and the MOUTH and NTIP regions will be used for both gender and ethnicity classification in Chapter 6 fusion experiments. The next type of features to be investigated is shape. The following chapter will cover results and discussion on partial-face shape experiments.

Chapter 4

Shape

Shape can mean several different things. It can mean the actual shape of an object, the outline and contours, or the distribution of smaller pieces within the object, such as the eyes, nose, and mouth. The relationship between the pieces gives an idea of the shape. Each of these can be represented differently. Common representations of shape include Active Shape Models (ASM), edges and lines, and the relationship between specified points.

4.1 Feature Extraction Methods

Face metrology has been used in previous works by Cao *et al.* to determine gender using points from the entire face [10]. Pairwise distances and angles between each possible pair of points were calculated. An example of some of the distances on the face can be seen in Figure 4.1. The features were ranked according to how well each one separated the classes in the training set, similar to d-prime. The most important features were chosen and used for gender classification. This method was adapted to work within the partial face schema. Only points found within a region were used. The only concession to full face was that all distances within each region were normalized by the interocular distance. The distance between two points was calculated using the Euclidean distance metric. The angle between two points was calculated by finding the inverse tangent between them. Table 4.1 shows the number of shape points used for each region and the length of the feature vectors. Since the number of points per region is small, all point pairs were considered important and no feature reduction was performed at this point.

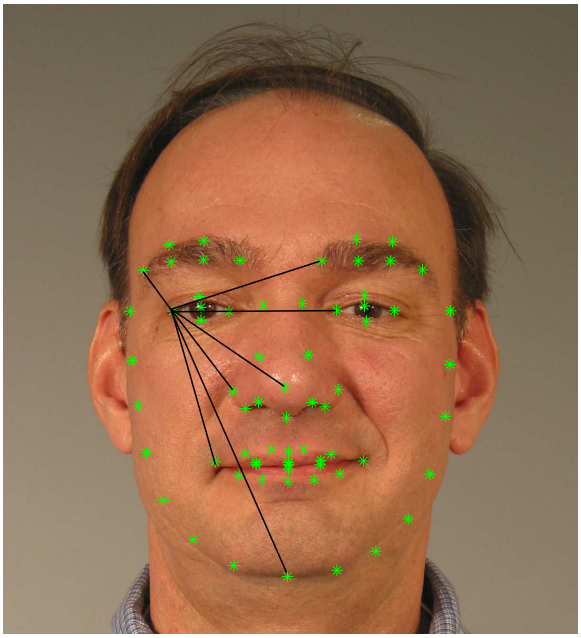


Figure 4.1: Example of shape feature calculation on a full face image.

Region	Points	Pairs	Total
FACE	68	2,278	4,556
EYE	5	10	20
NTIP	8	28	56
NOSE	10	45	90
MOUTH	19	171	342
CHIN	3	3	6

Table 4.1: Number of points used per region for shape features. Total feature vector length is $2 \times \binom{n}{2}$ where n is the number of points per region.

4.2 Experiment Setup

The experiments for this section will be performed on subsets of the FRGC and MORPH databases as previously described in Chapter 2. The points used in the feature extraction come from preprocessing performed on the color face images using the VeriLook SDK. Experiment sets remain the same from color experiments. Five runs of a stratified cross-validation experiment are run for each region and classifier combination. Experiments are performed with and without PCA feature reduction because of the small number of features present in each region. Eighty-four shape experiments are run for each demographic and dataset.

4.3 Analysis

Looking at the distance and angle features separately allows for several conclusions to be made. With respect to gender, the mean male subject has larger distances, and thus larger features in the chin and nose regions. The eyes and mouth regions are more similar between the genders. Even with this difference, the distance between points overlaps for the genders. In ethnicity, the *Black* class tends to have higher distances in the nose regions, indicating larger noses within this class. However, as with the genders, the difference between ethnic classes are still small and have a high overlap. The angle features show no trend with respect to either gender or ethnicity.

The shape of the nose changes as an individual ages, with distances between points increasing slightly as the age of a subject increases. This agrees with research findings summarized by Albert *et al.* [4] that the length and height of the nose increase with age. This trend is present in all demographic groups, both gender and ethnicity, although it is not as smooth in the *Female* class. The eyes, chin, and mouth regions show no consistent shape change with respect to age. Lip thickness likely changes in the mouth according to Albert *et al.* [4], but the age change is lost in the variability of the mouth due to changes in expression over the experiment set.

4.4 Gender

4.4.1 Reliability

Figure 4.2 shows the class-specific accuracies on the FRGC and MORPH datasets using 1-NN classification. The gap between class-wise performances is smaller in FRGC, but the actual performance is not much greater than chance. The higher resolution of the FRGC data allows for better point localization during the face annotation phase. Feature reduction with these shape features loses valuable information and actually makes performance on the *Female* class worse in some instances. The most notable, and best performing, instances are the NOSE and NTIP regions, likely due to the size difference between genders.

Figure 4.3 shows the results of gender classification using the ANN and SVM classifiers. Once again, feature reduction deteriorates performance. The nose regions seem to hold the most gender specific shape information for both MORPH and FRGC; however, the poor performance indicates that it is only a small amount. The shape features lack enough separation by gender to

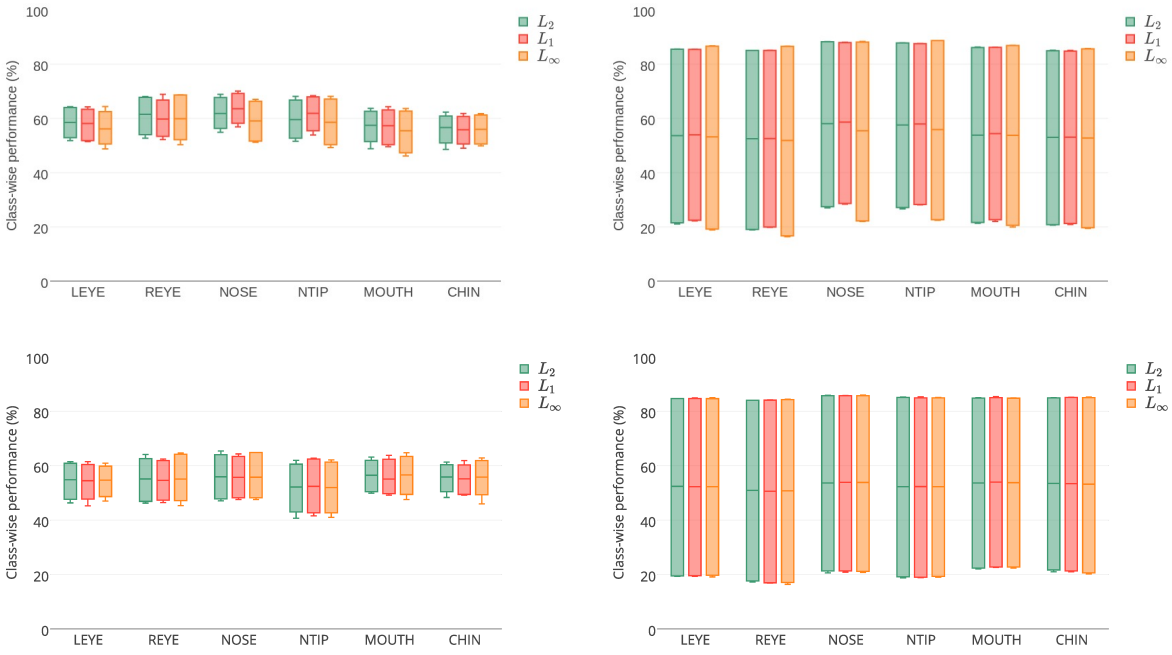


Figure 4.2: Gender nearest neighbor results on shape features. Values graphed are the class-wise accuracies over 5 runs of each experiment. (L to R): FRGC, MORPH. Top: no PCA feature reduction. Bottom: PCA feature reduction keeping 95% of the variance.

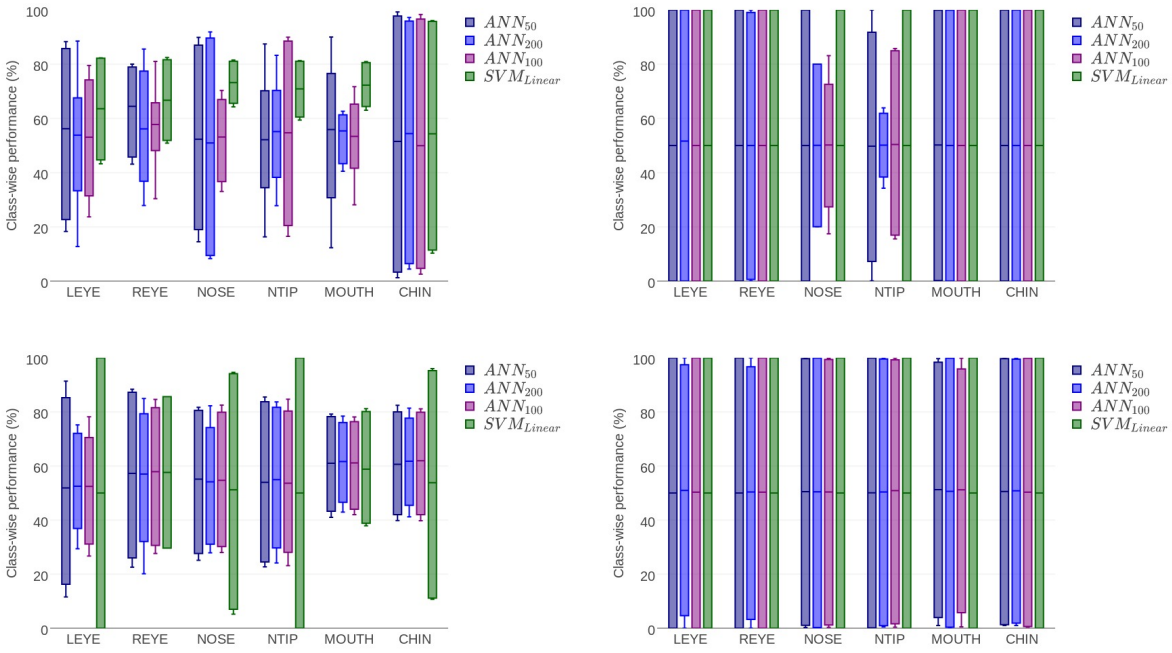


Figure 4.3: Gender ANN and SVM results on shape features. Values graphed are the class-wise accuracies over 5 runs of each experiment. (L to R): FRGC, MORPH. Top: no PCA feature reduction. Bottom: PCA feature reduction keeping 95% of the variance.

Figure 4.4: Hard subjects in shape gender classification on MORPH images. Images are of subjects classified incorrectly in over 98% of shape gender experiment runs.

	FRGC			MORPH		
Classifier	Male	Female	Average	Male	Female	Average
ANN_{100}	100.00	0.00	50.00	100.00	0.00	50.00
SVM_{Linear}	81.12	74.63	77.88	–	–	–
L_2	70.70	59.22	64.96	87.45	25.62	56.54
ANN_{100}	100.00	00.00	50.00	95.54	34.42	64.98
SVM_{Linear}	60.42	61.38	60.9	99.52	7.41	53.47
L_2	67.09	55.64	61.36	87.45	26.23	56.84
VeriLook	92.72	98.45	95.59	99.14	67.75	83.45

Table 4.2: Gender performance using shape features and full face. Top section is without PCA, while the bottom section of the table is results using PCA for feature reduction. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment. VeriLook refrained from predicting gender on 2.75% and 0.89% of male and female images respectively in the FRGC dataset. In the MORPH dataset it predicted no gender on 0.5% and 11.26% of male and female images.

train a reliable classifier in most instances based upon the 0-100% performance in the MORPH experiments. The FRGC experiments show a little more separation than MORPH, with correct predictions in both classes, for all regions.

No subjects were classified correctly in all shape experiments. The subjects and images that performed the worst can be seen in Figure 4.4 being misclassified in at least 98% of the shape experiments. The majority of the subjects are *Black* with one *Native American* and one *White*, but all the subjects are *Female*.

Table 4.2 shows the results of gender classification using shape features in the full face. In FRGC, the highest performance of an individual region is 73% average class-wise accuracy with the NOSE region and SVM classifier. This is below shape classification on full face and definitely below

VeriLook classification. In MORPH, the results are even less favorable dropping to 55% with the MOUTH and one of the nearest neighbor classifiers. The face shape results are also low for MORPH. This differs from what was seen in the work by Cao *et al.* [10]. Full face shape was able to achieve gender classification within 5-10% of appearance based methods. Therefore, the choosing of the best features is an important part of the method. This is the step where the purposed method deviated, either choosing all features or using PCA for feature reduction.

4.4.2 Age

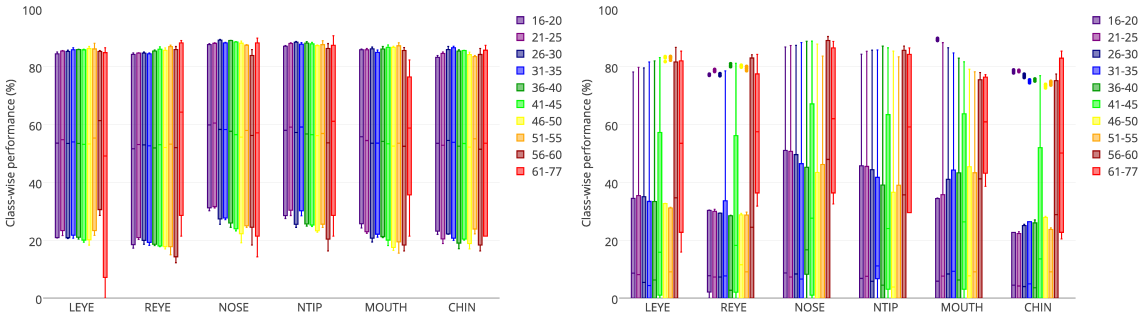
Gender class-wise performance partitioned by age groups can be seen in Figure 4.5(a). In most of the regions, the points used are stable, like the eye corners. Age should not influence the performance of those regions, unless the features themselves change with age. Points in other regions, like the mouth, are less rigid. It is possible that variability in these regions may be due to behavioral factors, as well as age. For instance, it is possible that younger subjects smile more often in the experiment set than older subjects.

No definite trend is noticed in the results with respect to age, except in the nose regions. The *Female* class shows a decrease in performance as age increases. As discussed in the analysis of shape features, the samples in the *Male* class, on average, have larger noses than those in the *Female* class; however, as an individual ages, the size of the nose increases. This means older female samples are more likely to overlap with the *Male* class and be classified incorrectly. The face results follow the same trend for PCA reduced experiments. Other regions show little or inconsistent changes in gender classification with respect to age.

4.5 Ethnicity

4.5.1 Reliability

Results of ethnicity classification using shape features can be seen in Figure 4.6. Once again, there is a slight performance gain when excluding the feature reduction step. Performance is still very low for both FRGC and MORPH. The NOSE and MOUTH regions seem to hold the most shape information related to ethnicity. For all experiments, one class performs well, which is the most well represented class, but the performances of the other classes suffer. This indicates that the



(a) (b)

Figure 4.5: Shape results by age for a) gender and b) ethnicity classification. Results shown are from the 1-NN experiments (L_2) with no feature reduction.

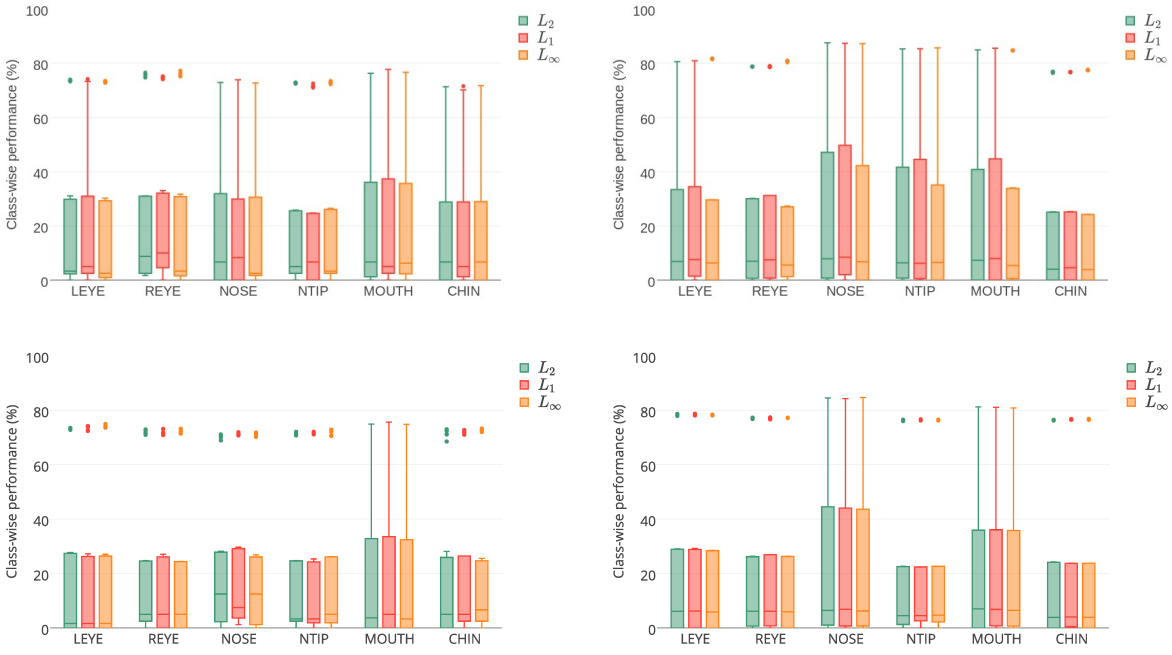


Figure 4.6: Ethnic nearest neighbor results on shape features. Values graphed are the class-wise accuracies over 5 runs of each experiment. (L to R): FRGC, MORPH. Top: no PCA feature reduction. Bottom: PCA feature reduction keeping 95% of the variance.

Figure 4.7: Easy and hard subjects in shape ethnic classification on MORPH images. Subject on the far left was classified correctly in all experiments. The subjects to the right are a subset of those misclassified in all shape ethnicity experiments, 18 out of 69. From top to bottom the real ethnic labels are *Hispanic*, *Native American*, and *Asian*. Subjects in the last column are *Female*; the rest are *Male*.

partial face shape data for ethnicity classes overlaps and is not very useful for ethnicity classification.

Figure 4.8 shows the results of ethnic classification using ANN and SVM classifiers. In most instances, better performance is achieved using a 1-NN classifier, indicating small clusters spread out over the classification space. One class does really well with the ANN and SVM classifiers, and the rest perform poorly, resulting in the outliers on the graphs. In FRGC this is the *White* class, in MORPH this is the *Black* class. Shape classification does improve on the next best class as the ANN gets larger.

One subject, *Black Male*, was correctly classified in all MORPH shape ethnicity experiments. Sixty-nine were misclassified in all of the experiments, some on multiple images. A subset of these subjects can be seen in Figure 4.7. The majority of these subjects are in the *Male* class, but *Female* subjects from each of the smaller classes are present as well. Half of the *Asian* class subjects are in this set, as well as 70% of the *Native American* class. The rest of the subjects are *Hispanic*, but are only responsible for approximately 6.5% of the subjects in that class. This supports the idea that the smaller classes overlap a lot with one another and the larger classes.

Tables 4.3 and 4.4 show the results of ethnic classification on shape features on the whole face. In FRGC, the MOUTH region with the SVM classifier is the only one that approaches the performance on the face. Results in the MORPH dataset improve slightly. The nose and mouth regions obtain average accuracies within 5% of the face results. While not the best representation,

Classifier	White	Asian	Hispanic	Black	Indian	Average
ANN_{100}	100.00	0.00	0.00	0.00	0.00	20.00
SVM_{Linear}	84.36	57.32	1.00	4.33	3.00	30.00
L_2	76.86	38.49	0.67	1.83	4.25	24.42
ANN_{100}	100.00	0.00	0.00	0.00	0.00	20.00
SVM_{Linear}	73.31	52.23	1.67	5.83	5.75	27.76
L_2	75.45	33.29	3.33	1.17	7.00	24.05

Table 4.3: Ethnic performance using shape features and full face, FRGC. Top section is without PCA, while the bottom section of the table is results using PCA for feature reduction. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment

Classifier	White	Asian	Hispanic	Black	Native American	Average
ANN_{100}	0.00	0.00	0.00	100.00	0.00	20.00
SVM_{Linear}	80.40	2.00	8.44	96.21	0.00	37.41
L_2	42.10	1.41	8.44	86.43	0.00	27.68
ANN_{100}	80.35	0.00	1.90	96.22	0.00	35.69
SVM_{Linear}	81.05	0.00	0.00	96.63	0.00	35.54
L_2	42.63	0.77	8.91	86.46	0.00	27.75

Table 4.4: Ethnic performance using shape features and full face, MORPH. Top section is without PCA, while the bottom section of the table is results using PCA for feature reduction. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment

shape features on the nose and mouth can reach comparable performance to similar shape features over the whole face.

4.5.2 Age

Ethnicity class-wise performance partitioned by age groups can be seen in Figure 4.5(b). Performance on the best class, *Black*, showed an upward trend in the eye regions as the age of subjects increased, suggesting that classification of ethnicity is easier on older subjects in the *Black* class. This is opposite to the trend found in the MOUTH and CHIN regions. The decrease in performance for the CHIN region could be due to the proclivity of the males in the class towards growing facial hair. This could possibly interfere with the accuracy of the facial annotation in that portion of the face and deteriorate performance. The nose regions perform similarly for all age groups in this class. With the *Black* class having larger measurements in these regions to begin with, an increase in the measurements would keep the separation between the classes.

In the *White* and *Hispanic* classes, the eyes seem fairly stable with respect to age while the nose regions show a downward trend in these classes. An increase in the nose measurements

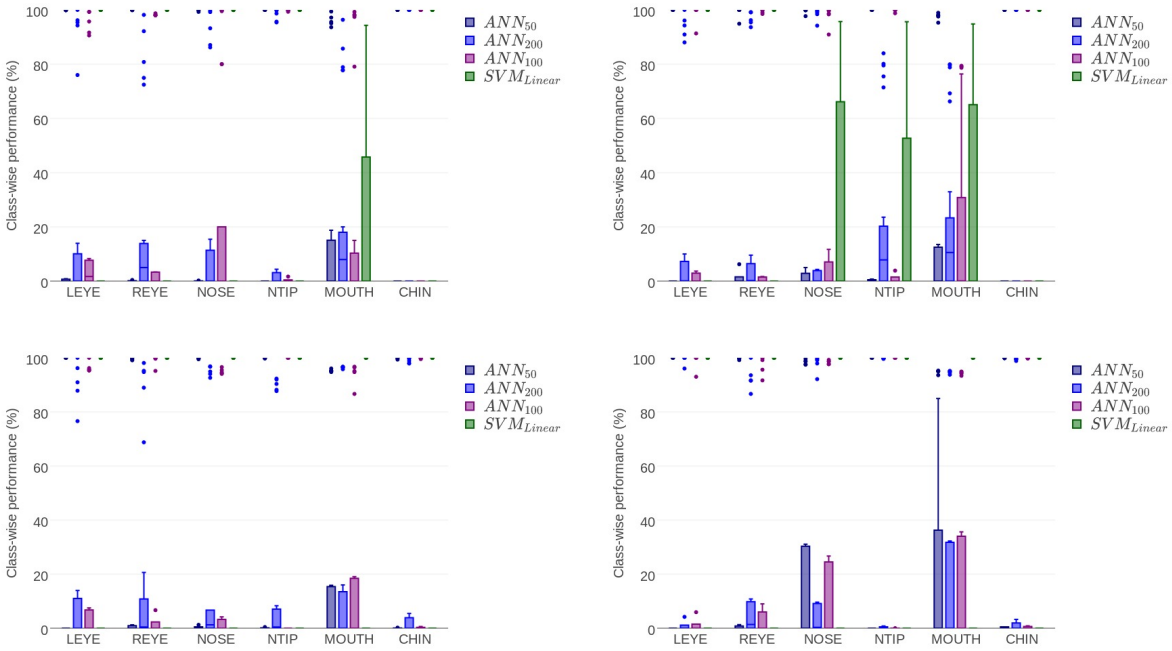


Figure 4.8: Ethnic ANN and SVM results on shape features. Values graphed are the class-wise accuracies over 5 runs of each experiment. (L to R): FRGC, MORPH. Top: no PCA feature reduction. Bottom: PCA feature reduction keeping 95% of the variance.

with these classes could begin to overlap with the *Black* class, decreasing performance. Performance on the CHIN region for both these classes declines as age increases, likely due to facial hair and landmark accuracy. The MOUTH region improves noticeably with age in the *White* class but is inconsistent with the *Hispanic* class. The performance of the MOUTH on both of these classes is still markedly below the corresponding performance on the *Black* class.

The *Asian* and *Native American* classes are the worst performing classes with the *Native American* class having very few correct classifications. In the *Asian* class, correct classifications more likely in the age groups under 40 years of age. From 16-25, correct classifications could be found in the eyes, nose, and mouth regions, but then the eyes and mouth became harder to classify, and just the achieved correct classification until subjects reach 40 years of age. This suggests that younger Asian individuals can be classified by eye shape, but the nose becomes a more important ethnic indicator as they age.

4.6 Conclusions

These particular partial face shape features do not work well with gender and ethnic classification. Performance is better with these features in higher quality images which allow for better localization of the points used. This indicates that there are some differences in shape between gender and ethnic classes. It is possible that these differences could be represented better with a more sophisticated shape representation. In future work, a different shape representation, possibly Active Shape Models, will be investigated for use in partial face.

The shape features perform better for the most part without feature reduction. The size of the original feature vectors in this category are much smaller than those used in color and texture experiments. Feature reduction on these small feature vectors resulted in a loss of too much information for both gender and ethnic classifications, resulting in lower classification performance than the original vectors. Baseline face experiments did not necessarily follow this trend. Some classifiers performed better after PCA reduction was performed.

The most stable regions with respect to age for both gender and ethnicity are the eye and chin regions. The overall shape of the eye does not change much with age, neither does the shape of the chin. The chin region has its own variance due to facial hair, but the small section of chin used for shape here, does not vary much with age. These regions might be the most stable, but they are

not the best performing regions for the shape features investigated.

The nose regions show the best performance for both gender and ethnicity. These regions are also impacted the most by age. Males, on average, have larger noses than females which give the best performance in gender classification. Between the ethnic classes there are also some size differences which allow for better classification. However, the nose grows slightly with age, which can negatively impact both gender and ethnic performance as subjects get older. For future work, the age of subjects should be taken into account when planning the training and testing sets.

Different types of classifiers performed better between gender and ethnicity. Nearest neighbor classification performed better for ethnicity classification than the ANN and SVM classifiers, indicating a large overlap between the classes that the more sophisticated classifiers were unable to classify correctly. The SVM classifiers performed well in the FRGC dataset with gender classification, but MORPH did not have enough separation to learn the *Female* class, always predicting *Male*. MORPH performed better with larger ANNs or nearest neighbor classification than SVM or ANN with the smallest hidden layer.

For further experiments in the application section, the NOSE region will be used for only ethnicity classification. Since shape region experiments performed better without feature reduction, PCA will not be performed on the nose shape features. The linear SVM classifier will be used to classify this region in the hopes of obtaining the highest performance. The final category of features investigated for the reliability and age questions is texture. The following chapter will discuss the results on texture classification.

Chapter 5

Texture

Local texture can be found in the face in terms of skin texture, wrinkles, imperfections in the skin, and facial hair, including, but not limited to, eyebrows, mustaches, and beards. The stableness of texture depends on what is causing the specific texture. Skin texture will most likely be stable on a day to day basis, but change over the years as wrinkles increase. Imperfections in the skin and facial hair can change more quickly. Three different local texture representations will be investigated in this chapter: Histograms of Oriented Gradient, Local Binary Patterns, and Local Phase Quantization. A brief description of the extraction method follows for each texture representation.

5.1 Feature Extraction Methods

5.1.1 Histograms of Oriented Gradient

Originally proposed by Trigg and Dalal [18] for the detection of human pedestrians, Histograms of Oriented Gradient (HOG) have been used for facial recognition purposes [20, 65]. The basic idea of HOG is that local shape and appearance, or texture, can be characterized by the distribution of the local image gradients. The image is divided into smaller regions, providing the localized area. The Prewitt convolution kernel is used to compute the image gradient. The gradient magnitude, G_M, and the gradient angle, G_A, are computed by

$$G_M = \sqrt{G_X^2 + G_Y^2} \text{ and } G_A = \text{atan2}(G_Y, G_X).$$

G_X and G_Y are the image gradients in the horizontal and vertical directions. The gradient orientations, or angles, for each pixel are used to select the histogram bin and increment by the gradient magnitude. For this work, the orientations are divided into 30° segments resulting in 12 bins and a feature vector of length 12 per patch.

5.1.2 Local Binary Patterns

Local Binary Patterns (LBP) were first introduced by Ojala *et al.* [53] to classify texture patterns. The LBP method looks at the neighborhood around each pixel in an image. The value of this feature representation is that it encodes different textures that can represent curved edges, spots, and even uniform areas. The texture for a specific pixel is represented by thresholding the intensity values of the neighboring pixels with the intensity value of the center pixel, given by the equation:

$$LBP_{P,R}(g_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \text{where } s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

P is the number of pixels in the neighborhood investigated along a circle of radius R which is centered at pixel c. g_c and g_p refer to the grayscale value for the given pixels. $LBP_{P,R}(g_c)$ represents the texture pattern at pixel g_c. The texture patterns are accumulated into a histogram. The granularity of the histogram is dependent on the choices for P and R. For the LBP features used in this work the values $P = 8$ and $R = 2$ were chosen. The "uniform" version of LBP was used which limits the patterns in the histogram to those with 2 or less changes between 0 and 1, resulting in a feature vector length of 59 per patch. Patterns that are not uniform are counted in the last bin of the histogram. LBP was chosen having been used successfully for soft biometric classification using full facial images [41, 75].

5.1.3 Local Phase Quantization

Local Phase Quantization (LPQ) was recently proposed by Ojansivu *et al.* [54] as a descriptor for texture which is robust to image blurring. Similar to HOG and LBP, the LPQ method looks at each pixel individually and accumulates the results into a histogram. The phase information is quantized and compiled into a histogram. By utilizing only the phase information, the method is also not affected by uniform illumination changes. It has been used successfully for facial recognition

[2, 11, 12].

The LPQ feature extraction method first performs a Discrete Fourier Transform (DFT) on the image, given by the equation,

$$F(u,x) = \sum_{y \in N_x} f(x-y) e^{-j2\pi u^\top y},$$

where u is the frequency, x is the pixel location within the image, N_x is a rectangular $M \times M$ neighborhood, and $f(x)$ is the intensity value of a pixel in the image. Four frequencies are considered in this algorithm. The real and imaginary components of the frequencies are used to create a transform matrix such that $F_x = W f_x$ where f_x is a vector containing the pixels in N_x, W is the transform matrix, and F_x is the transform coefficient vector.

After further processing of F_x, including decorrelation and whitening, the coefficient vector is quantized by thresholding the vector at 0. The LPQ code, corresponding to the histogram bin, is calculated by

$$b = \sum_{j=1}^{8} g_j 2^{j-1},$$

where g_j is the thresholded value in F_x. The feature vector for LPQ has 256 elements per patch.

5.2 Experiment Setup

The experiments for this section will be performed on subsets of the FRGC and MORPH databases, using a grayscale version of the image. Experiment sets remain the same from color and shape experiments and preprocessing is performed as described in Chapter 2. Features are extracted using the HOG, LBP, and LPQ methods. Gender and ethnicity classifications are performed on all texture features using all classifiers. The total number of partial-face experiments in this chapter is 126 for each dataset and demographic. Five runs of a stratified cross-validation experiment are run for each feature, region, and classifier combination. Baseline face experiments are performed using LBP texture features.

5.3 Analysis

Figures 5.1 show the average HOG features over each region for MORPH. Not much consistency exists across the regions. Even the eyes do not share similar texture indicating that texture around the eyes are independent. Issues may arise with the differences in texture between the regions. Texture may rely more on accurate localization of the facial features. If the region extraction is off, different texture may be present and not be classified correctly. These features also indicate that the mouth and chin region hold different types of textures than the other regions, see bins 4-7.

Figures 5.2 show the average LBP features over all the regions for MORPH. The last bin in the histogram counts all the miscellaneous textures that are not counted in the rest of the bins. The FRGC features have twice the amount of miscellaneous textures as MORPH. This is most likely due to the higher resolution of the FRGC images allowing texture to be extracted in greater detail. Differences between regions in the rest of the bins are much smaller due to the content being spread over a greater number of bins. Other than magnitude, the shapes of the feature vectors are similar over both datasets. The texture features for the eyes are more similar than HOG, but several sections of the histogram are different, indicating that LBP texture is somewhat independent as well.

The LPQ feature vectors are much larger than both HOG and LBP, and will not visualize well. With 256 bins in the histogram, the quantity in each bin is very small. Values are found in the majority of bins, unlike many of the color space features.

Figure 5.3 shows the differences between the regions for each feature, using the histogram intersection distance measure. Differences between the left eye and the other regions were similar to differences between the right eye and the rest of the face. The same is true for the nose and nose tip regions; therefore the graphs have been condensed to show pairs with the left eye, nose tip, mouth and chin regions. The nose and the nose tip are the most similar region still because of the overlap present between the two. The largest differences in texture exist between the eye and chin regions. These differences are larger for females on average, so it is not entirely due to facial hair. The *Black* class has the smallest region differences for each texture representation, indicating texture is more stable around the face for that particular ethnic group. The region differences with texture features are larger than those reported in Section 3.4 indicating that texture is not as stable around the face as color.

Region differences show less change with respect to age for the *Female* class than the *Male*

class. This indicates that texture across the face is more stable over the years for females than males. This could be due to make-up and other facial care products that are more widely used by females which smooth skin texture. It could also be due to the variation in facial hair found in both classes. The largest difference that also shows a large variance with respect to age for the *Male* class is the eye to mouth comparison. Younger male faces show a larger difference in texture between these two regions than older males. This trend holds true for more than just the *Male* class and this region comparison.

In the majority of the comparisons, if a variance is present with respect to age, the younger subjects have larger differences than older subjects, indicating that texture is not as stable around the face for younger subjects as it is for older subjects. The exception is present in some comparisons on the lower face. In the *Black* class, the distances between the nose and mouth or chin regions is larger for older subjects, and also in the *White* class in the mouth to chin comparison. In the *Black* class, the largest difference with high variance due to age is the difference between texture in the eye and mouth regions. Combined with the face that this class also has the most stable texture around the face, this suggests that wrinkles, especially in the eye region, show up later than in other classes. For the *White* class, the corresponding difference is found between the eye and nose regions. The *Hispanic* class has the most age variance between the same regions as both the *Black* and *White* classes. The other ethnic classes show no discernible pattern with respect to age.

5.4 Gender

5.4.1 Reliability

Figure 5.4 shows the results of 1-NN gender classification on the texture features for both FRGC and MORPH. FRGC shows there are differences between the genders based on texture with some performances above 80%. The LPQ features perform the best out of the three texture representations indicating that this feature representation encodes the most gender information. These features are able to distinguish gender correctly at least 70% of the time in all but the nose region. This region performed the worst in all three feature representations, indicating less gender texture information can be found in the nose. The CHIN region was able to distinguish between the genders the best. This is likely due to the presence of more facial hair, and thus different texture, in the CHIN region of the *Male* class. The HOG features performed the worst, which is a result of the

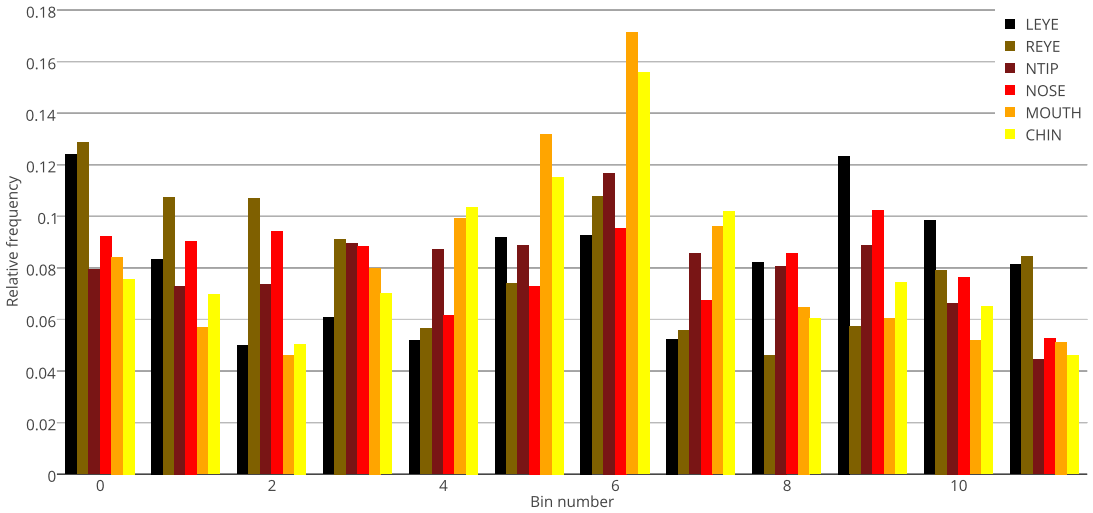


Figure 5.1: Average global feature vector for MORPH HOG features. These are the normalized histograms for each region. The x-axis corresponds to the bin in the histogram. The y-axis corresponds to the relative frequency of values found in each bin.

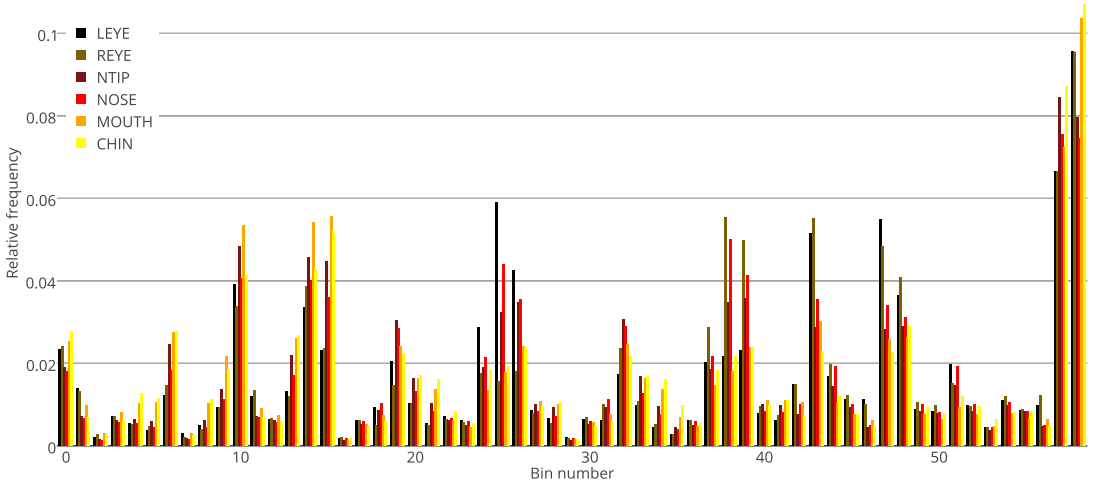


Figure 5.2: Average global feature vector for MORPH LBP features. These are the normalized histograms for each region. The x-axis corresponds to the bin in the histogram. The y-axis corresponds to the relative frequency of values found in each bin.

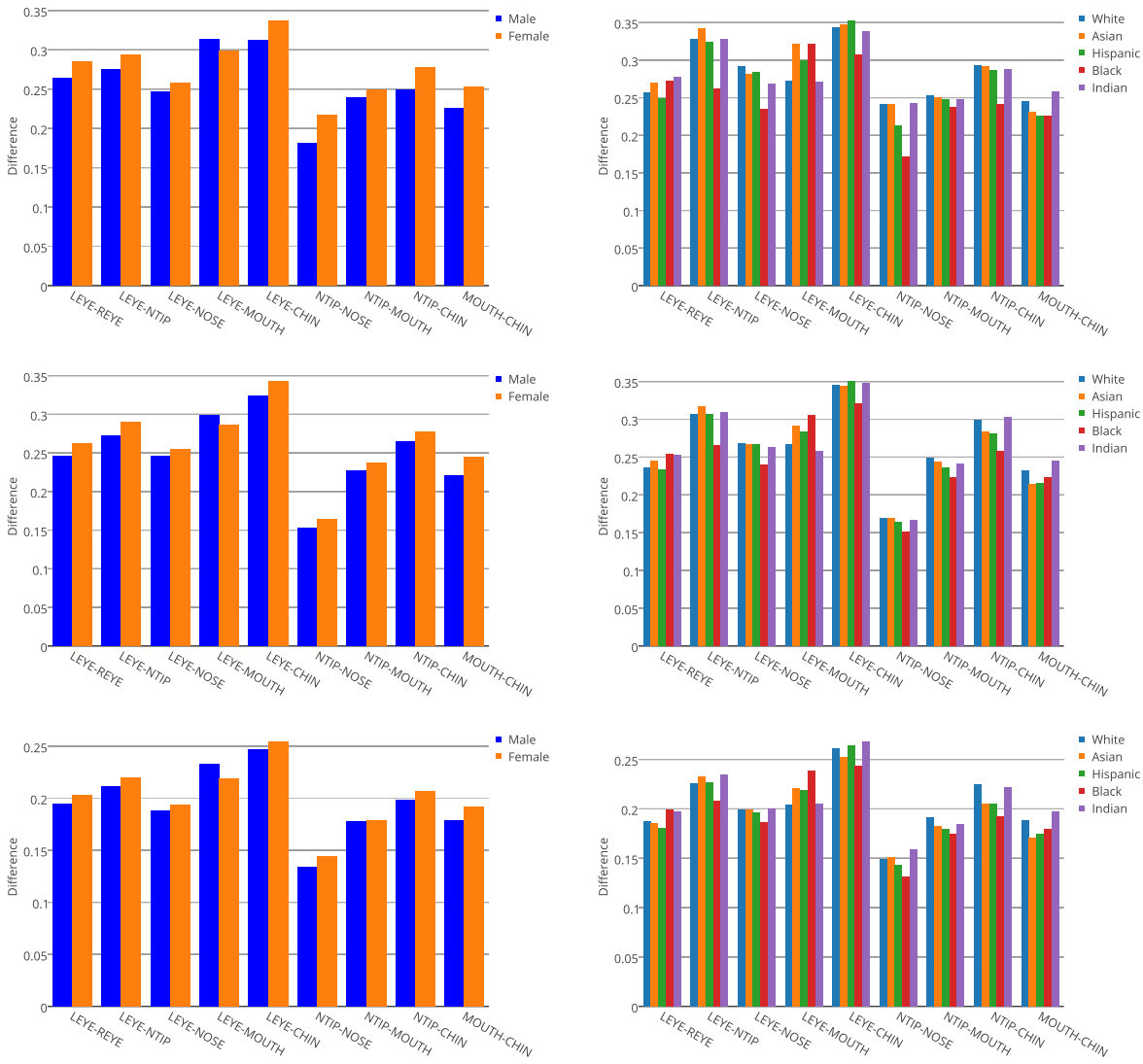


Figure 5.3: Mean difference between MORPH region-wide global texture vectors by demographic. Differences shown are for HOG, LBP, and LPQ (top to bottom) using the Histogram Intersection distance.

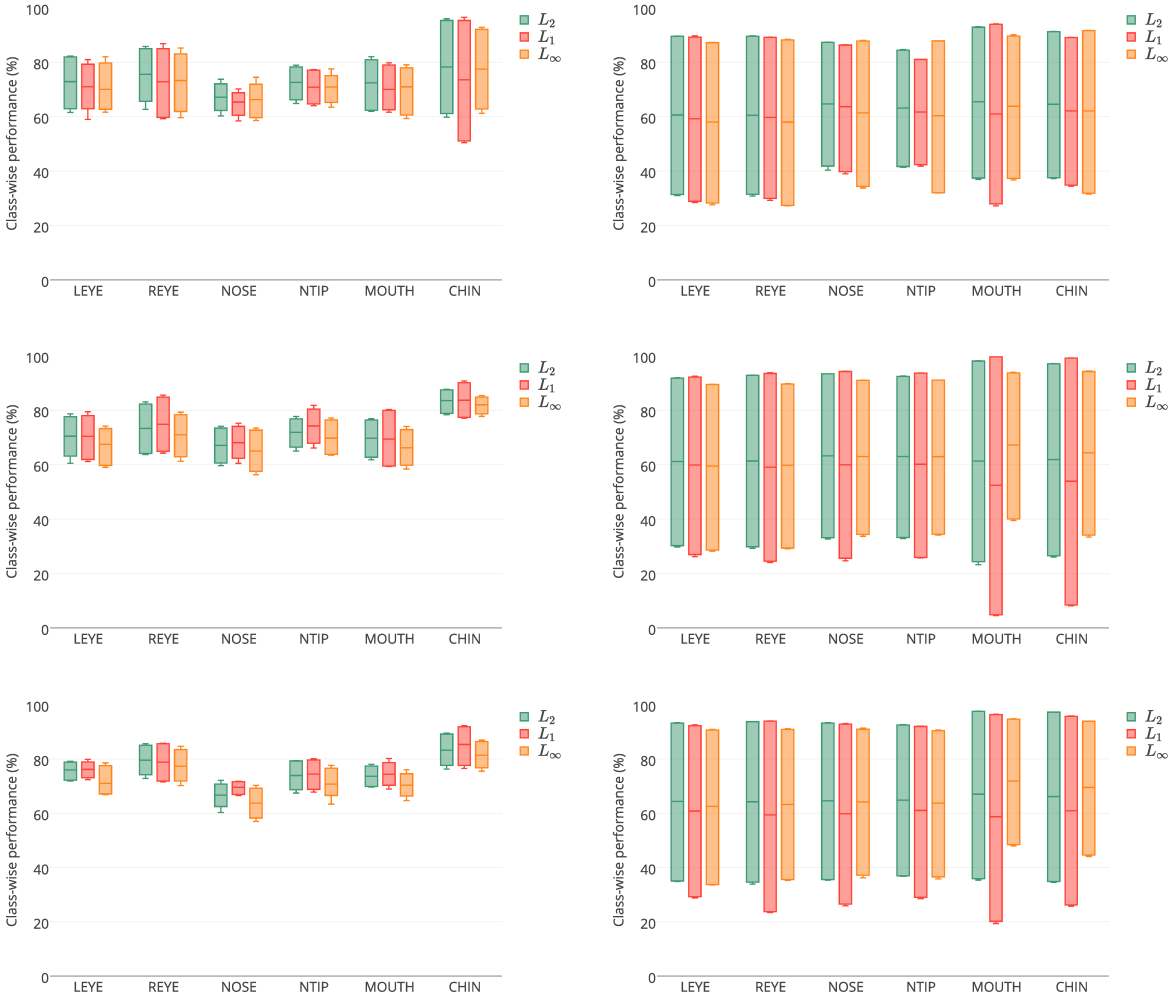


Figure 5.4: 1-NN gender classification on texture features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: HOG, LPQ, LPQ. L to R: FRGC, MORPH

texture not being consistent around the face. Small errors in localization result in different textures extracted, which leads to worse performance.

The performance on the MORPH dataset shows more overlap between the classes than is seen in the FRGC dataset. This is likely due to the lower quality of the MORPH dataset. The texture cannot be extracted as reliably because it is not present in the same detail as in FRGC. The CHIN region does not perform noticeably better like it does in the FRGC dataset. Performance on the *Male* class increases, but the performance of the *Female* class suffers for the L_1 distance measure. The performance of the HOG features in MORPH is much closer to the performance of the other two representations than it is in FRGC. This could be a positive side effect of the lower resolution. Since the texture is not extracted in as much detail, the HOG features are more similar and not as dependent on the precise localization of the region.

Figure 5.5 shows the results of gender classification on the texture features using the more complex classifiers, ANN and SVM. The results improve for both datasets, all regions, and all representations. This suggests that there is a lot of overlap in the feature space, but differences exist that set the classes apart, which the ANN and SVM classifiers are able to learn. The HOG features still perform the worst in FRGC, and the gap between HOG and the other two representations are more noticeable in MORPH with these classifiers. These results suggest that HOG is not the best texture feature for encoding gender information.

Figure 5.6 shows some of the easier and harder subjects to classify according to gender using texture features. The majority of the subjects classified correctly in all texture experiments were from the *Black* class. Some were from the *White* and *Hispanic* classes. All were from the *Male* class. The images that were misclassified in over 95% of texture gender experiments were all from subjects labeled *Black Female*. As previously seen in color classification results, one of the subjects labeled *Female* is not female. Texture classification also predicted the class correctly, but was marked incorrect since the subject was mislabeled.

In these results, the chin region still performs very well in the FRGC dataset, but it can also be seen that the mouth and nose regions perform better than the other regions while looking at the MORPH dataset. The lack of texture detail found in MORPH seems to shift the importance from the chin region to the mouth and nose regions. This indicates that the lower face regions hold more texture information specific to gender than the upper face regions. The eye regions also perform very well in FRGC, but the lack of texture detail in MORPH is more detrimental, putting their

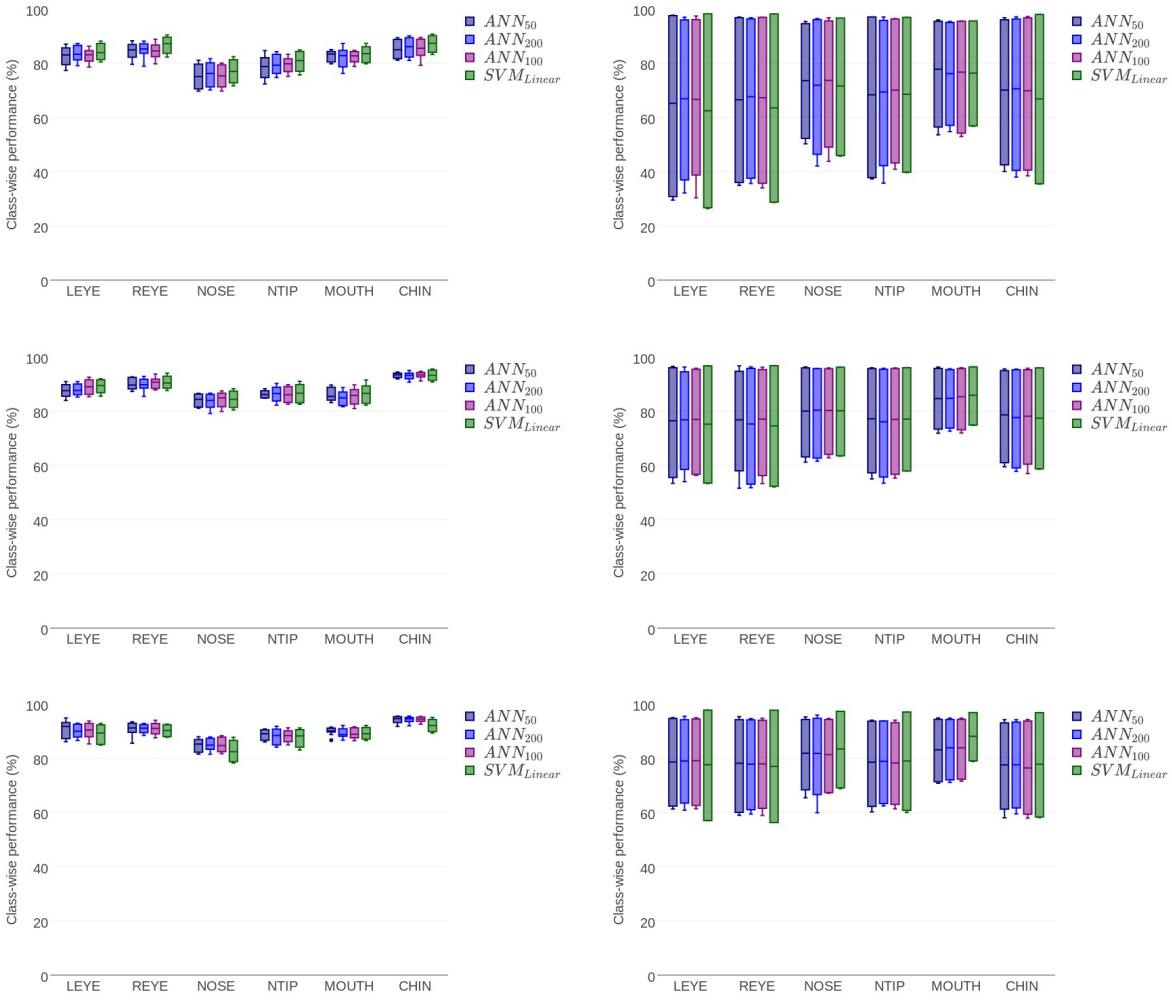


Figure 5.5: ANN and SVM gender classification on texture features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: HOG, LPQ, LPQ. L to R: FRGC, MORPH

Figure 5.6: Easy and hard subjects in texture gender classification on MORPH images. Subjects on the left are a subset of those correctly classified in all texture experiments. Top to bottom: *Black*, *White*, and *Hispanic*. Subjects on the right those who were misclassified in over 95% of experiments. The subject on the bottom was labeled incorrectly as *Female*.

	FRGC			*MORPH*		
Classifier	Male	Female	Average	Male	Female	Average
ANN_{100}	94.85	15.80	55.33	–	–	–
SVM_{Linear}	92.89	91.02	91.96	–	–	–
L_2	79.21	64.35	71.78	97.43	22.16	59.80
VeriLook	92.72	98.45	95.59	99.14	67.75	83.45

Table 5.1: Gender performance using LBP texture and full face. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment. VeriLook predicted no gender on 2.75% and 0.89% of male and female images respectively in the FRGC dataset. In the MORPH dataset it predicted no gender on 0.5% and 11.26% of male and female images.

performance below the nose regions. This still follows what was seen in in previous work, the nose and eye regions perform among the best in previous texture experiments [47].

Table 5.1 shows the results of the baseline face experiments using LBP texture and gender classification. Texture on the CHIN region was able to achieve similar performance to the commercial classification of the entire face in FRGC. The MOUTH region, as well as the NOSE region, was able to do the same in the MORPH dataset. Since these regions achieve similar performance to the commercial results and not just the classification of the full face features, the gender information encoded in the texture of the MOUTH and CHIN seems to be sufficient for partial face classification.

5.4.2 Age

The results of gender classification on MORPH partitioned according to age can be seen in Figure 5.7. The classifier used in the experiments shown is the ANN_{50} classifier. These graphs indicate that the *Male* class is impacted less by changes in age than the *Female* class; however the mouth and chin regions do show variance in both classes with respect to age. The nose regions show the least variance in classification accuracy in the LBP and LPQ features. The nose tip region shows less variance than the nose region, which makes sense, because the nose region overlaps the eye regions which show the most variance with respect to age. The impact of age is seen the most severely in the eye and chin regions for the *Female* class. Overall, the nose tip region is the least impacted by age, but the mouth region still performs well over all age groups, even with variance caused by age.

In full face experiments, little change is seen in the performance of the *Male* class according to age. The *Female* class shows a downward trend with respect to age, with the lowest performance on subjects 41 to 50 years of age. After that performance improves again.

5.5 Ethnicity

5.5.1 Reliability

The results of nearest neighbor ethnic classification on both datasets can be seen in Figure 5.8. Compared to gender classification, the results are much more spread out. In both datasets, one class performs rather well, with a high class-specific accuracy, but at least one or more of the classes also has a class-specific accuracy below 10%. MORPH most often has at least two classes below 10%, *Asian* and *Native American*, and one more under 20%, *Hispanic*. The best performing classes are still the most well-represented classes in the experiment set.

Figure 5.9 shows the results of ethnic classification of the texture features using the ANN and SVM classifiers. FRGC experiments improve overall, but still have a very low median class-wise accuracy. MORPH experiments improve on the best performing classes, but decline slightly on the less well-represented classes. The exceptions to these are in the LBP features using a SVM classifier. All but the CHIN region show a slight increase in the worst performing classes. In these experiments, MORPH results are likely to have the same two classes under 10%, but the other class, the *Hispanic*

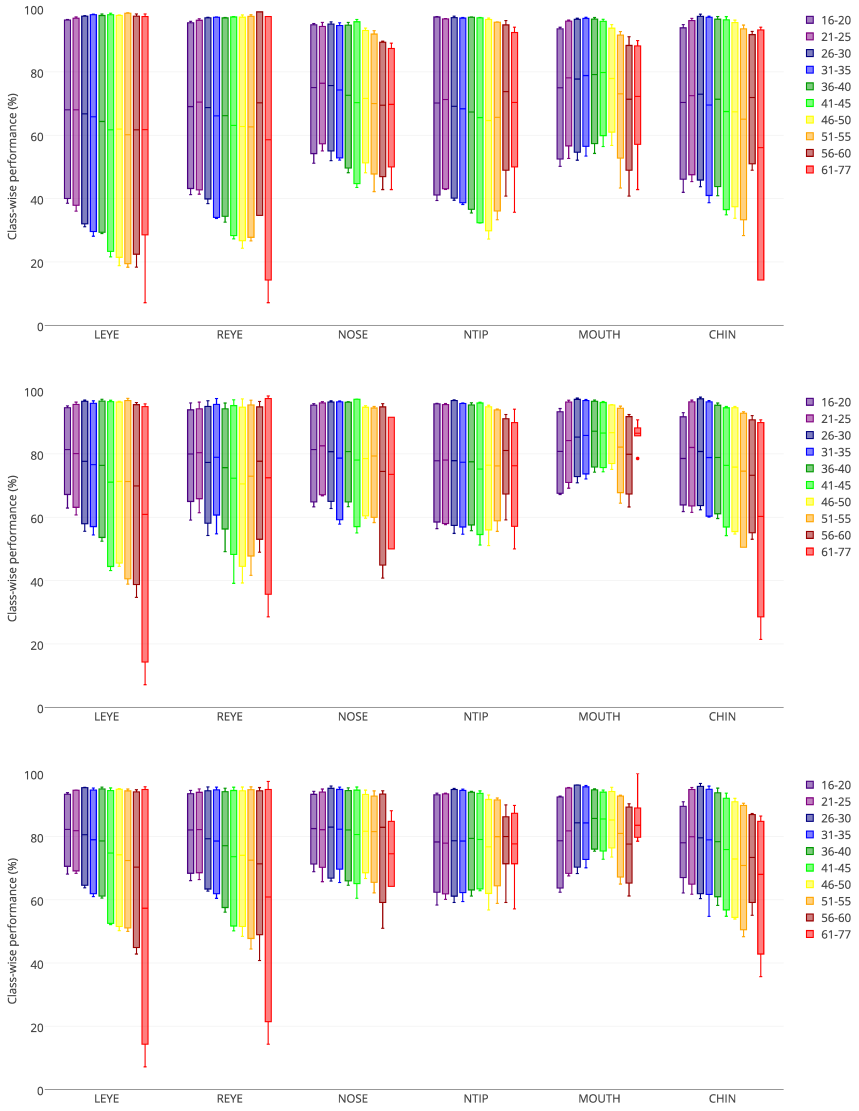


Figure 5.7: Gender performance on texture features partitioned by age. From top to bottom the graphs represent the HOG, LPB, and LPQ experiment results. The classifier used is ANN_{50} for each graph shown.

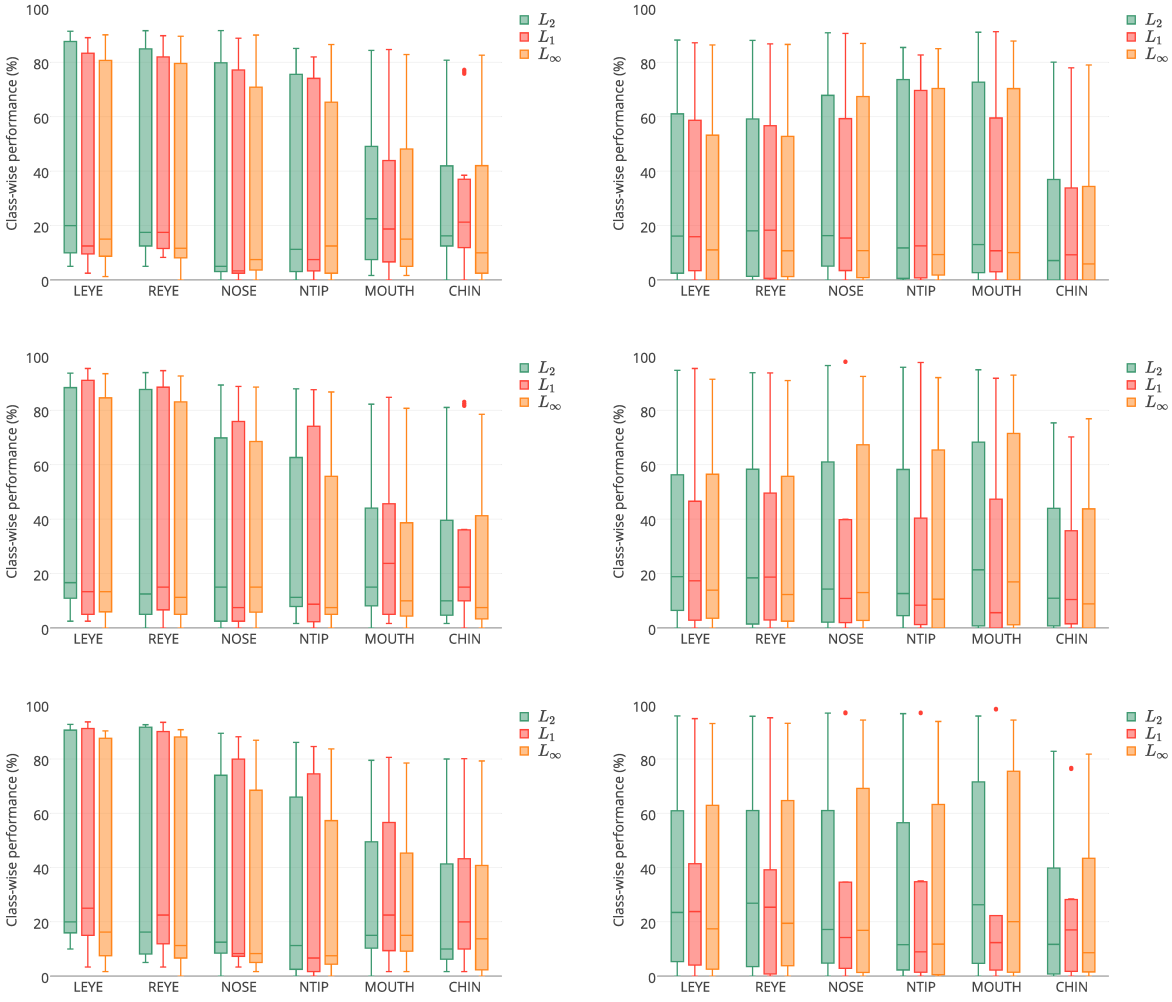


Figure 5.8: 1-NN ethnic classification on texture features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: HOG, LPQ, LPQ. L to R: FRGC, MORPH

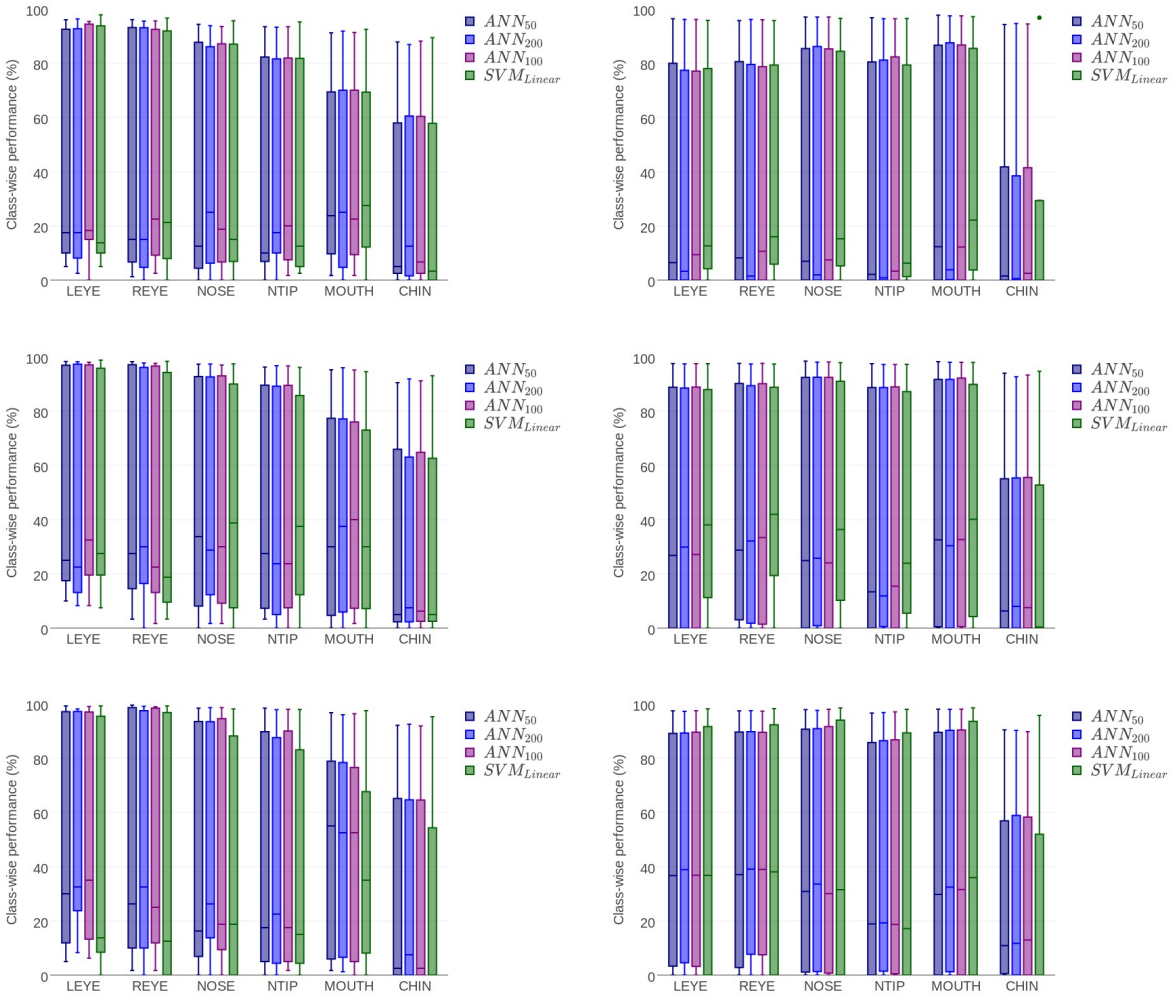


Figure 5.9: ANN and SVM ethnic classification on texture features over all regions. Values included in the box-and-whisker plot are class-specific accuracies from 5 runs of the experiment. Top to bottom: HOG, LPQ, LPQ. L to R: FRGC, MORPH

Figure 5.10: Easy and hard subjects in texture ethnic classification on MORPH images. Images to the left are a subset of those correctly classified in all experiments. Images to the right are from each subject having an image that was never classified correctly. Most are from the *Native American* class. The three images in the lower right-hand corner correspond to *Hispanic* subjects.

Classifier	White	Asian	Hispanic	Black	Indian	Average
ANN_{100}	99.54	12.84	0.33	0.33	0.00	22.61
SVM_{Linear}	97.28	92.01	2.33	13.00	14.25	43.78
L_2	84.04	53.10	1.00	3.50	8.75	30.08

Table 5.2: Ethnic performance using texture and full face, FRGC. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment.

class, improves so that it more often has a class-specific accuracy above 20%.

Images from 25 subjects were classified correctly in all texture ethnicity experiments. Subjects with images classified incorrectly for all experiments were from the *Native American* and *Asian* classes. Figure 5.10 shows images from some of these subjects. The hardest class to identify with texture was the *Native American* class. All 13 subjects from that class have at least one image that was never classified correctly. The easiest subjects to identify were *Black* subjects, specifically *Male*.

Table 5.2 and 5.3 show the results of ethnic classification on the FRGC and MORPH datasets respectively using LBP features and the full face. The FRGC eye and nose regions outperform the full face experiments by at least 5% on the average class-wise accuracy. The same regions do not

Classifier	White	Asian	Hispanic	Black	Native American	Average
ANN_{100}	92.96	5.32	48.62	98.25	0.00	49.03
L_2	50.87	2.38	18.15	95.56	0.00	33.39

Table 5.3: Ethnic performance using texture and full face, MORPH. Percentage values represent the class-specific accuracies and the average class-specific accuracy for each experiment.

outperform MORPH face experiments, but come within 3-5% of full face results. This indicates that some of the texture ethnic information is lost with the lower resolution, but that enough can be found for comparable performance to full face experiments.

5.5.2 Age

Ethnic results according to age can be seen in Figure 5.11. The results shown use the ANN_{50} classifier. The graphs for each feature look similar, so only the graph for the LPQ features is shown. Overall the mean class-specific accuracy shows a downward trend with increasing age. The jump in performance for the 41-45 and 61-77 classes is due to the face that not all ethnic classes were present in these age groups and were omitted from the graph.

The *White* and *Black* classes show fairly consistent performance across the age groups with very small declines present for some features and regions after 56 years of age. Class-specific accuracies for the *Hispanic* class peak around 46% in the 21-25 age group. Performance for that class drops below 25% in the 41-45 age group and does not recover. This suggest that younger Hispanic partial faces are easier to classify according to ethnicity than older ones. Classification of the *Asian* class is spotty, but most of the correct classifications fall within the first two age groups, 16-25. This also indicates that younger Asian partial faces are easier to classifier according to ethnicity than older faces in the same ethnic group.

The regions that seem the least affected by age are the nose regions. The most affected region is the chin. However, in each of the regions, the *Hispanic* class is negatively affected by age, no matter the performance on the other classes. The results suggest that age affects different ethnic groups differently on the various parts of the face.

Overall, the face shows the same trends as the partial face experiments. Younger subjects are easier to classify, especially for the *Asian* and *Hispanic* classes. The *Black* class is the least affected by age with performance only dropping 8% from its highest when subjects reach 56 years of age.

5.6 Conclusions

Results from this chapter suggest that texture from partial face images can be used successfully for gender and ethnic classification. Analysis shows that texture is not as stable around the

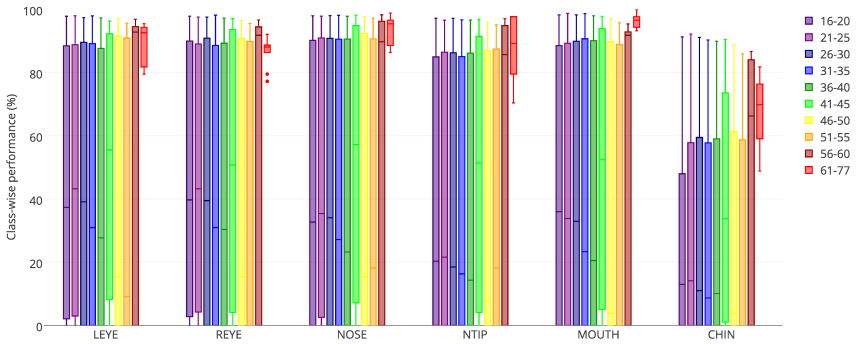


Figure 5.11: Ethnic performance on texture features partitioned by age. Results shown are from the LPQ experiments using the ANN_{50} classifier.

face as color. More texture gender information can be found in the lower regions of the face, while the upper face holds more ethnic texture information. The various regions of the face are affected differently by age. The texture information in the nose is least affected by age while the texture in the chin is the most affected. The LPQ and LBP features achieve higher performance than the HOG features.

The best results for gender classification were achieved using the chin region with the LBP and LPQ features and the ANN classifiers on high quality images and the mouth region with LPQ features and the linear SVM classifier with the MORPH images. The best average class-specific accuracy on FRGC images was 94% and the best on MORPH was 87%. The classifiers performed better on the *Male* class than the *Female* class indicating males have more distinct gender characteristics than females. The MOUTH and CHIN regions were able to achieve comparable results to full face classification using a commercial system.

The best results for ethnic classification were achieved using the left eye region with the LPQ features and the ANN classifiers on high quality images and the right eye region with LBP features and the linear SVM classifier with the MORPH images. The best average class-specific accuracy was 52% on FRGC and 49% on MORPH. The eye regions performed well in both the LPQ and LBP features. With the ANN classifiers, the eyes performed just 2-3% under the best performance on MORPH. The eye and nose regions were able to achieve comparable results to full face classification using similar features.

The effect of age seems to be present more by class than facial region. The *Black* and *White* classes were less impacted by age than the *Hispanic* class, but that might be due to the larger

number of samples present for each age group in these classes. Younger *Hispanics* are easier to classify by ethnicity than older individuals. The same is true for the *Asian* class. The performance of the *Male* class overall holds steady over the different age groups while performance of the *Female* class holds steady until the 41-45 age group. The mouth region improves slightly with age while the nose regions decline the least.

Since one of the fusion methods will use the scores from ANN classification, the $ANN_1 00$ classifier will be used for each region and feature combination. The REYE region with LPQ features will be used for ethnic classification and the MOUTH with LBP for gender classification as the best performing regions over all. The NOSE region with LPQ will also be used for both gender and ethnicity as a more age stable option. This concludes the analysis of reliability and the impact of age on partial-face images. The following chapter covers the fusion experiments using partial-face gender and ethnicity classifications to filter a face biometric experiment.

Chapter 6

Application

Using the conclusions from the previous chapters, the experiments in this section will combine classification results with a typical face recognition or verification experiment in an effort to improve performance. The features and regions that will be used were chosen from color, texture, and shape. The following experiments will be performed on the Pinellas dataset.

6.1 Experiment Setup

Experiments in this section require three sets of images: training, gallery, and probe sets. Table 6.1 shows the demographic breakdown of each of these sets. In combining classification results with the performance of the VeriLook SDK face experiment, the Pinellas images had to pass the quality check within VeriLook. Images from previous FRGC and MORPH experiments did not have to meet this criteria. VeriLook also recommends 50-75 pixels between the eye centers. The average interocular distance for the Pinellas images used was 110.80. A maximum of four images were selected for each subject. Subjects with 3-4 good images were used in the gallery and probe sets, while subjects with 1-2 good images were used in the training set. In the probe and gallery sets, the image with the highest age was placed in the probe set. Any subject with images labeled with conflicting gender or ethnicity information was excluded from the lists. In order to create a more balanced training set, many subjects and images were discarded from the candidate training list. A limit of 10,000 subjects per gender and ethnicity class (i.e. *White Male*) was set. One image per subject was used. This limit affected both gender classes in the *White* and *Black* ethnicities.

	Ethnicity				Gender		
	Training	*Probe*	*Gallery*		*Training*	*Probe*	*Gallery*
White	20,000(1)	62,025(1)	62,025(2)	Male	30,555(1)	71,383(1)	71,383(2)
Asian	1,562(1)	738(1)	738(2)	Female	21,620(1)	37,208(1)	37,208(2)
Hispanic	10,613(1)	7,059(1)	7,059(2)				
Black	20,000(1)	38,769(1)	38,769(2)				
Total	52,175(1)	108,591(1)	108,591(2)		52,175(1)	108,591(1)	108,591(2)

Table 6.1: Breakdown of subjects in the Pinellas experiment sets by gender and ethnicity. The number of images per subject in each class is given in parentheses.

Because of the relative small size of the *Native American* class, less than 200 subjects, and the poor performance of the class in the MORPH experiments, subjects of this class were excluded from these experiments. Out of 393,426 subjects with consistent gender and ethnicity labels, 52,175 were used for training and 108,591 for classification and the biometric experiment.

Prior to training and classification, PCA is performed on the texture and color features, keeping 95% of the variance, similar to the previous experiments. The PCA projection for the classification experiments is learned on the training set.

Two types of fusion will be investigated, using two levels, decision and score. In the first method, which can be described as a weighted sum decision fusion, the results of multiple classifications can be used in combination with the match scores from the recognition and verification experiments. These classifications will determine if the match score is used or discarded. Each of the probe classification results, p_i, are provided to the experiment along with a weight, m_i for each classification set and a threshold, where $i = 1 \ldots C$ and C is the number of classifications used. A particular match score, comparing probe j to gallery k is used when the following evaluates to true: $\sum m_i(p_i(j) == g_i(k)) > threshold$. The value $p_i(j)$ is the predicted class of the j^{th} value in the probe using the i^{th} classification. Similarly, the value $g_i(k)$ is the given class of the k^{th} value in the gallery using the i^{th} classification. Basically, this is a weighted sum of which classifications match. If enough match, the value is above the threshold, and the score is used. This equation can be used to perform 'AND' and 'OR' decisions as well as equally weighted 'j out of k'. The second method is based on the scores generated by the classifier, one for each class. The maximum score is the class given as the decision, but it is possible that other scores are very close. If any of the scores are within a given threshold of the maximum score, the corresponding class is added to the predicted list. The match score will be used if the class of the gallery entry is found within the predicted list of the probe. Multiple classifiers may be used in this method as well.

	Ethnicity			Gender		
Label	*Region*	*Feature*	*Classifier*	*Region*	*Feature*	*Classifier*
1	MOUTH	LCH(RGB)	ANN_{100}	MOUTH	LCH(RGB)	ANN_{100}
2	NTIP	LCH(RGB)	ANN_{50}	NTIP	LCH(RGB)	ANN_{100}
3	REYE	LPQ	ANN_{100}	MOUTH	LBP	ANN_{100}
4	NOSE	LPQ	ANN_{100}	NOSE	LPQ	ANN_{100}
5	NOSE	Shape	SVM_{Linear}	–	–	–

Table 6.2: Region/Feature/Classifier combinations chosen for application experiments.

Table 6.2 lists the region/feature/classifier combinations that will be used in the experiments for this chapter. Two combinations were chosen for both gender and ethnicity classification for texture and color features. Shape features performed very poorly for gender classification, and so will only be used for ethnicity classification in this chapter. The MOUTH region was chosen for color as the best performing region for both gender and ethnicity. The NTIP region was chosen for being more stable with respect to age. The LCH features from the RGB color space perform the best over all color spaces investigated. The best texture region/feature combinations were the REYE/LPQ and MOUTH/LBP for ethnicity and gender classification respectively. The NOSE region with LPQ was chosen for its performance across the age groups. All except the shape classifier will use an ANN classifier to facilitate the score level fusion method. The ANN classification was close to linear SVM in the best cases, just a little lower. The shape classifier will use a linear SVM because all ANN classification was poor, even in ethnicity classification.

Results in this chapter will be reported following previous methods for classification, including confusion matrices and box-plots of class-specific accuracies. Verification and recognition results will be reported using DET and CMC plots as well as EER and Rank-1 values.

6.2 Analysis

6.2.1 Classification Results

Both the gallery and probe sets were classified according to the region/feature/classifier combinations given in Table 6.2. The results from classification of the gallery set will be reported here to give an idea of how well the combinations performed on Pinellas data. Table 6.3 shows the results for the combinations involving color, while Table 6.4 shows results using texture combinations. There are two main differences between MORPH and Pinellas ethnic classifications. Pinellas experiments

	White	Asian	Hispanic	Black
White	45.58	0.95	49.24	4.22
Asian	12.13	4.20	72.76	10.91
Hispanic	5.06	0.34	88.95	5.65
Black	1.53	0.11	12.00	86.36

(a) 1e, average class-wise accuracy: 56.27%, overall accuracy: 62.68%.

	White	Asian	Hispanic	Black
White	33.23	1.89	55.05	9.83
Asian	3.66	14.16	59.55	22.63
Hispanic	2.94	0.96	86.99	9.11
Black	1.85	0.51	24.08	73.56

(b) 2e, average class-wise accuracy: 51.99%, overall accuracy 51.00%.

	Male	Female
Male	81.45	18.55
Female	24.01	75.99

(c) 1g, average class-wise accuracy: 78.72%, overall accuracy: 79.58%.

	Male	Female
Male	84.94	15.06
Female	35.10	64.90

(d) 2g, average class-wise accuracy: 74.92%, overall accuracy: 78.07%.

Table 6.3: Confusion matrices from classifying the gallery set on the chosen color region/ feature/ classifier combinations. The row is the true class and the column is the predicted class. All entries are percentages. Class-specific accuracies are highlighted.

only utilize four ethnic classes, leaving out the *Native American* class. Since this class averaged close to 0% in MORPH experiments, the Pinellas average class-wise accuracies will show an increase without that class. The other difference is in the class-specific accuracies for *White* and *Hispanic*. The performance on the *White* class is reduced by almost half while performance on the *Hispanic* class doubles. The biggest factor for the *Hispanic* class is likely the larger number of samples in the training set. While not equal to the number samples in the *White* and *Black* classes, the *Hispanic* training set is much closer to their proportions in Pinellas than MORPH. The decrease in performance in the *White* class is most likely related due to overlap between the classes since a large percentage of the *Hispanic* class were misclassified as *White* in MORPH experiments.

The results of ethnic classification on the nose shape features can be seen in Table 6.5. The average class-wise accuracy for this experiment is low, but it has a higher overall accuracy than other ethnicity experiments. This is due to the higher performance on the *White* class, which accounts for approximately 57% of the probe and gallery sets.

Gender classification on both texture and color combinations is similar to that found in the MORPH experiments. Performance on the *Female* class showed little change, but the performance on the *Male* class dropped approximately 10%. *Males* are still correctly classified more often than *Females*, just not as much.

The texture combinations still show the best performance when looking at average class-wise accuracies. Color and shape are sometimes able to gain higher overall accuracies based on the distribution of classes in the set, but texture performs better for the majority of the classes.

	White	Asian	Hispanic	Black
White	50.13	0.32	43.34	6.21
Asian	8.20	17.28	61.31	13.21
Hispanic	4.46	0.39	89.12	6.03
Black	1.93	0.10	14.00	83.97

(a) 3e, average class-wise accuracy: 60.12%, overall accuracy: 64.52%.

	White	Asian	Hispanic	Black
White	42.50	0.50	50.19	6.81
Asian	5.49	12.87	57.45	24.19
Hispanic	3.58	0.40	86.53	9.49
Black	1.36	0.15	14.80	83.68

(b) 4e, average class-wise accuracy: 56.40%, overall accuracy 59.86%.

	Male	Female
Male	87.64	12.36
Female	23.68	76.32

(c) 3g, average class-wise accuracy: 81.98%, overall accuracy: 83.76%.

	Male	Female
Male	85.73	14.27
Female	21.94	78.06

(d) 4g, average class-wise accuracy: 81.89%, overall accuracy: 83.10%.

Table 6.4: Confusion matrices from classifying the gallery set on the chosen texture region/ feature/ classifier combinations. The row is the true class and the column is the predicted class. All entries are percentages. Class-specific accuracies are highlighted.

	White	Asian	Hispanic	Black
White	74.70	0.00	5.78	19.52
Asian	44.38	0.00	8.81	46.82
Hispanic	42.51	0.00	11.63	45.86
Black	10.06	0.00	3.40	86.53

Table 6.5: Confusion matrices from classifying the gallery set on the chosen shape combination, 5e. The row is the true class and the column is the predicted class. All entries are percentages. Class-specific accuracies are highlighted. Average class-wise accuracy: 43.22%, overall accuracy: 74.32%.

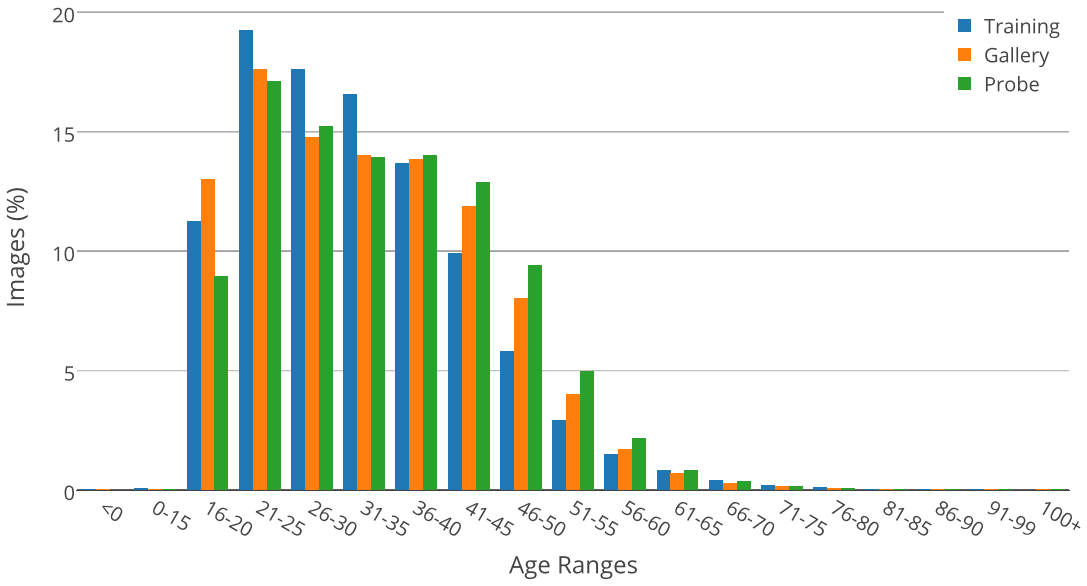


Figure 6.1: Age distribution of Pinellas experiment sets. Values graphed represent the percentage of images in each set found in the age ranges listed.

6.2.2 Age

Figure 6.1 shows the distribution of ages across the three image sets used for Pinellas experiments. While compiling information on age, it was discovered that six images in the gallery set and one in the training set are labeled with negative ages. Each one is from a different subject and those in the gallery set have two corresponding images with a maximum age gap of 2 years between them. The images were stored with date of birth and date of image capture which were used to calculate the age. In some instances the correct date of birth was entered, but a default image capture date instead of the actual date was used, January 1, 1900 for example. These cases resulted in a negative age. Figure 6.2 shows images from the gallery which had negative ages.

Images with calculated ages less than sixteen were also found. While it is possible that the images labeled with an age of fifteen are legitimate, it is highly unlikely that subjects age two to nine years old are present in a database compiled from booking photographs. The image lists show 19 images of two to nine year-olds, 24 of ten to fourteen year-olds, and 50 of fifteen year-olds. Figure 6.3 shows some examples of images that are associated with ages fifteen and under. At the other end of the scale, the image lists show 15 images of subjects greater than one-hundred years old. This is

Figure 6.2: Images from Pinellas with negative calculated ages.

Figure 6.3: Pinellas images with calculated ages below 16. Calculated ages of the images on the top row from L to R: 2 and 9; and on the bottom row: 13 and 15.

possible; however, it is very unlikely that a subject would be in the database at both twenty-five and one-hundred and five years of age, which is the case for one probe/gallery combination. Figure 6.4 shows images from the probe and gallery sets with ages greater than one-hundred years old. These obvious errors are due to mistakes in the metadata which can be caused by mistyped or default information. It is possible that other errors in age occur in the middle ages, but these are not as obvious since those ages are well represented and have a high probability of naturally occurring in the database.

Each of the image sets, training, gallery, and probe, contain approximately the same distribution of ages. The training and gallery sets contain more younger subjects, but are within 5% of the amounts in the probe set. The creation of the probe and gallery sets put the younger images in

Figure 6.4: Pinellas images with calculated ages over 100. Top row are images in the gallery set. Bottom row consists of corresponding probe images (also with an age over 100).

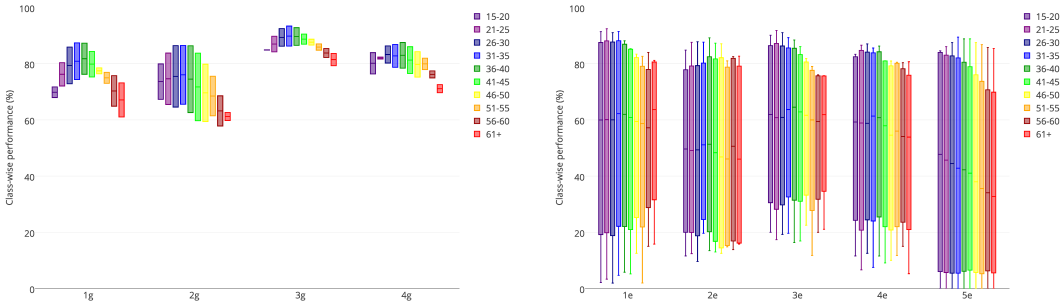


Figure 6.5: Performance of chosen region / feature combinations on the probe set according to age. Results are from gender (left) and ethnicity (right) classification.

the gallery set which accounts for the discrepancy between those lists. The selection of the training set did not account for age.

Figure 6.5 shows the performance of each of the region/feature/classifier combinations according to age for both gender and ethnicity classification. Gender classification shows more of an impact with respect to age than ethnicity, with the easiest ages to classify falling between 30 and 45. For most of the ethnic classes, younger subjects are easier to classify according to ethnicity than older ones. The decline is the most pronounced in the *Hispanic* class. These findings agree with previous observations on the MORPH experiments.

6.2.3 Rank-1 Score vs Genuine Score

The match scores of the straight face experiment directly impact how much improvement can be seen in the fusion experiments. The range of similarity scores for the straight whole-face experiment was 0 to 10080. Figure 6.6 shows an enlarged section of the distribution of scores

between impostor and genuine comparisons. The area of overlap between impostor and genuine scores is of interest in this section.

Several aspects were investigated for individuals that had a genuine match not at Rank-1 for the straight whole-face experiment. The average score difference between the Rank-1 match and the highest genuine match was 51.51, with a standard deviation of 102.31, which is not much considering the range of scores present. Over 95% of the subjects had a difference of less than 200 between the genuine and Rank-1 match. 174 of them actually had a difference of zero. These instances were ties that were unaccounted for in the experiment. The system marked the lowest score as a match and did not update the match unless a score was strictly less than the match it held.

The average genuine score for probes with an incorrect Rank-1 match was 78.58 with a median of 54 and a standard deviation of 68.44. The average Rank-1 match score was 130.09 with a median of 88 and a standard deviation of 133.40. The average rank of the genuine match was 1,777, but with a median of 3, it seems highly skewed. The standard deviation was 12,368.48. The mean genuine score over all the probes was 199.36, with an average Rank-1 score of 207.18. This supports what can be seen in Figure 6.6, that the lower the genuine score, the more likely it is for the probe to have an impostor score as a Rank-1 match.

6.2.4 Ground Truth Fusion

A baseline experiment was run with no soft biometric classifications used to filter the results. This experiment, performed with the VeriLook system, obtained a Rank-1 performance of 84.83% and an EER of 3.27%. The straight facial experiment was also fused with the ground truth information for gender and ethnicity, both together and separately. This is the best performance possible if the classifiers performed perfectly. These graphs can be seen in Figure 6.7. For Rank-1 performance there is little change; the fused experiments increase the Rank-1 performance up to 85.29% for the experiment where gender and ethnicity both match. The biggest increases occur when ethnicity information is utilized. The CMC curves reflect the same trends, with the fused experiments having a steeper beginning than the straight experiment. The EERs increase with the fusion of both (AND) gender and ethnicity, increasing the most to 4.02%. This feels counter-intuitive. If obviously wrong scores are being excluded and all the genuine scores are being kept, it seems as if the performance should get better. Table 6.6 gives the Rank-1 performance and EER for each experiment as well as the total number of genuine and impostor scores used in the calculations.

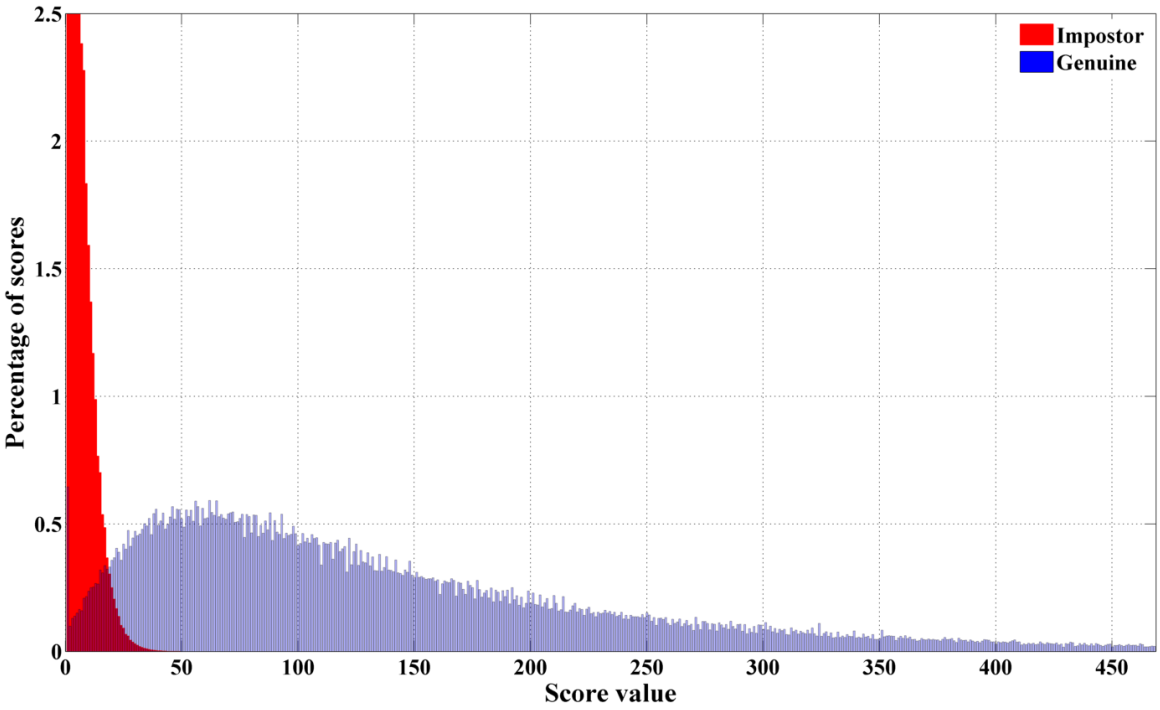


Figure 6.6: Distribution between impostor and genuine scores for straight whole face experiment.

In verification experiments, the FAR and FRR are used to create the graphs and calculate the EER. For these ground truth experiments, the FRR will stay constant since all genuine scores were included. The FAR is what changes. In calculating the FAR, the number of impostor scores above the threshold is divided by the total number of impostor scores. In this way the EER is dependent on the number of impostor scores. In Table 6.6, as the number of impostor scores used increases, the EER decreases. Table 6.7 shows the average number of comparisons per probe for each of the baseline experiments. The greatest reduction, almost 75%, occurs when both gender and ethnicity need to match, but a 24% reduction still occurs when either gender or ethnic class matches are included.

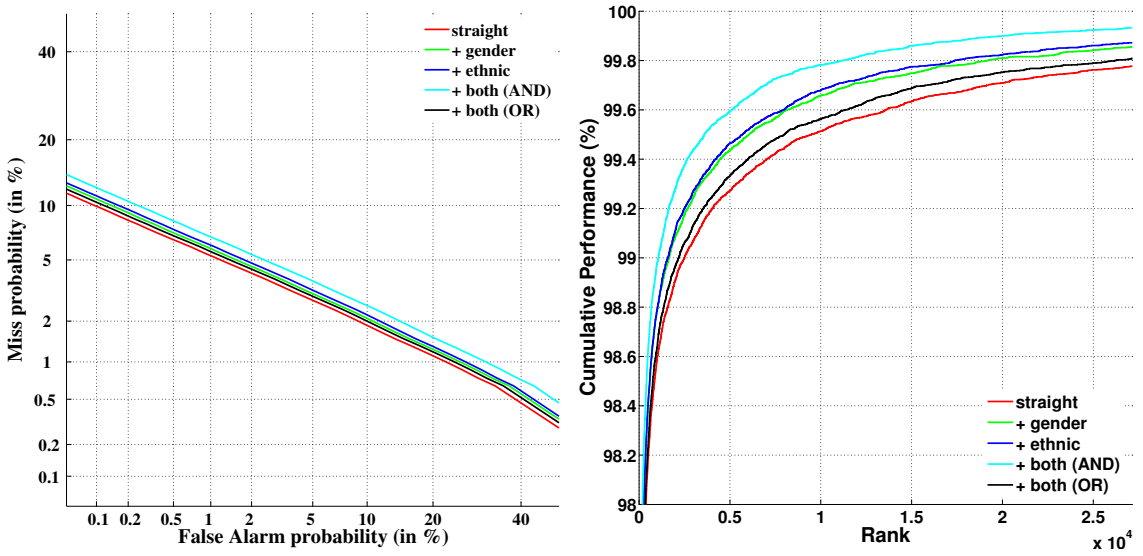
Figure 6.6 shows the distribution between genuine and impostor scores for the facial verification and identification experiments. The majority of the impostor scores are grouped with scores below 50, while the genuine scores are more spread out. 70% of the impostor scores are actually zero. Less than 1% of the genuine scores are also zero. It is likely that the genuine scores of zero contribute to the minimal increase in Rank-1 performance when using soft biometric information. Even if the candidate matches are narrowed to those matching in both gender and ethnicity, there is a good probability that one of the impostor scores will be greater than zero.

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
Straight	23,583,793,380	217,182	84.82	**3.27**
Gender	12,959,718,724	217,182	84.99	3.65
Ethnicity	10,800,803,040	217,182	85.12	3.68
Both (AND)	5,906,125,240	217,182	**85.29**	4.02
Both (OR)	17,854,396,524	217,182	84.85	3.49

Table 6.6: Baseline performance details on facial recognition fusion experiments. Highest Rank-1 performance and lowest EER are in bold.

Method	*Impostor (Reduction)*	*Genuine*
Straight	217,180.00 (0%)	2.00
Gender	119,344.37 (45%)	2.00
Ethnicity	99,463.15 (54%)	2.00
Both (AND)	54,388.72 (75%)	2.00
Both (OR)	164,418.75 (24%)	2.00

Table 6.7: Average comparisons per probe on baseline experiments. Percentage by which the number of impostor scores are reduced as compared to the straight experiment is given in parenthesis.



(a) Verification

(b) Identification

Figure 6.7: Baseline performance on identification and verification experiments. Baseline experiments include the straight whole face experiment and fusion with ground truth gender and/or ethnicity. DET is on the left and CMC is on the right.

A large percentage of the probe/gallery pairs, 91.6%, have an age gap of 0-4 years, with 52.2% of the pairs actually having a gap of less than one year. Only 7.3% of the pairs have an age gap between 5 and 9 years. Of the rest, 1% of probe/gallery pairs have an age gap from 10-14 years, while only 0.1% have an age gap of 15 years or more. Some of these large age gaps stem from issues in the metadata mentioned earlier, such as a 25 year old with an image also appearing with an age of over 100. While the quality of the images is not as high of quality as found in FRGC, the images in these experiments had to meet the quality standards of the VeriLook SDK algorithm. The close age gap for most of the images and the multiple gallery entries, as well as the quality standards, work together to make the problem easier than it could have been. It is possible that, by making the original problem harder and changing any of those factors, more room for improvement using fusion can be seen.

6.3 Results

There is always a risk in filtering the gallery before matching that the true match will be excluded. That is why it is important to use multiple classifications to minimize the possibility of the genuine matches getting filtered out. However, the more classifications used and/or the more lenient the rule, the more gallery entries will be included. What good is it to do all the work for classification and yet still end up using the entire gallery due to a lenient filtering rule? It would be less work to just do all the gallery comparisons. Mixing the types of classification will help minimize this. For example, there are two gender classifications for the probe and each predicts a different gender, an 'OR' classification rule will not rule out any of the gallery entries. But if an ethnicity classification is added and the rule changed to '2 out of 3 (equal weights)', some gallery entries would be excluded based on the ethnicity classification. Still, having multiple gender classifications will decrease the number of matches used, only on probes with differing classifications will all gallery entries be included. The following experiments will investigate using different numbers and types of classification, both feature and soft biometric trait. The legends in the following graphs will use the labels from Table 6.2 with an 'e' or a 'g' following to indicate ethnicity or gender for the classification type. An indication of the rule used will also be included, such as 'AND' or 'OR'.

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
1g	12,707,142,872	171,434	67.46	3.40
2g	13,069,745,244	165,812	65.02	**3.39**
1g2g OR	16,174,486,048	198,884	77.85	3.40
1e	6,471,957,338	125,026	48.25	3.80
2e	5,518,189,458	99,636	38.57	3.63
1e2e OR	7,961,731,784	138,652	53.59	3.68
1g2g1e2e OR	18,694,632,076	209,012	**81.58**	3.40

Table 6.8: Weighted sum fusion with color information. Highest Rank-1 performance and lowest EER are in bold.

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
5e	10,451,692,626	159,468	61.92	3.88
1e5e OR	12,374,469,974	190,666	74.49	3.50
2e5e OR	12,229,813,908	182,340	71.28	**3.46**
3e5e OR	12,420,118,146	190,800	**74.53**	3.48
4e5e OR	12,321,065,622	187,332	73.16	3.48

Table 6.9: Weighted sum fusion with shape and other information. Highest Rank-1 performance and lowest EER are in bold.

6.3.1 Weighted Sum Decision Fusion

Table 6.8 shows the results of filtering the gallery results using classifications based on color, combinations 1 and 2, and the weighted sum decision rule. None of the experiments using just color classifications were able to get too close to the results of the straight experiment. This indicates that color alone is not what is needed to improve performance; however, since the color regions were not the best performing regions from the MORPH dataset, this is not unexpected.

It is interesting to note that the performance in the verification experiments (EER) does not change as dramatically as the performance in the recognition experiments (Rank-1). The CMC curve is affected more by the exclusion of the genuine matches than the verification measurements. The verification measurements include information about the number of impostor attempts, which can outweigh the genuine attempt information.

Results using ethnic shape classification combined with other ethnic classifications can be seen in Table 6.9. Shape ethnicity alone outperforms the color ethnicity experiments, but still does not achieve comparable performance to the straight facial recognition experiment. Performance can be traced back to good performance on the most prevalent class in the experiment set, *White*.

Table 6.10 shows the results of the weighted sum fusion using texture classifications. Results

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
3g	12,843,138,128	192,678	75.54	3.54
4g	12,716,977,840	178,866	70.30	3.46
3g4g OR	15,415,754,464	210,044	82.19	3.49
3e	6,893,394,536	129,712	50.30	3.67
4e	6,451,954,140	120,486	46.66	3.66
3e4e OR	8,819,949,264	154,140	59.71	3.63
3g4g3e4e OR	18,436,420,630	215,202	**84.09**	**3.44**

Table 6.10: Weighted sum fusion with texture information. Highest Rank-1 performance and lowest EER are in bold.

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
1g3g OR	15,522,684,680	207,358	81.17	3.47
1g4g OR	15,742,372,896	203,056	79.57	3.44
2g3g OR	16,197,310,236	209,090	81.87	3.44
2g4g OR	15,871,225,112	199,762	78.25	3.42
1g2g3g OR	17,498,432,304	212,966	83.33	3.42
1g2g4g OR	17,439,113,114	209,562	81.98	3.40
1g3g4g OR	16,962,513,126	213,370	83.46	3.44
2g3g4g OR	17,278,916,008	213,618	83.57	3.42
1g2g3g4g OR	18,227,744,400	215,076	**84.12**	3.40

Table 6.11: Weighted sum fusion with mixed color and texture features. Highest Rank-1 performance and lowest EER are in bold.

from these classifications are higher than those using color and shape, although shape outperformed the single texture ethnic experiments. Using both gender and ethnicity texture classifications, these experiments were able to come within 1% of the Rank-1 performance of the straight experiment. These results further support the conclusion that the most gender and ethnicity information can be found in texture features.

The results in Table 6.11 show results using a mixture of color and texture classifications, either all gender or all ethnicity. Since the ethnic classifiers do not perform as well as the gender classifiers, fusion experiments using only ethnic classifications are unable to match the performance of the experiments utilizing gender classifications and are so omitted from the table. All four texture and color gender classifiers together are able to achieve a higher Rank-1 performance than any of the experiments using just color or texture, but it is still below the Rank-1 performance of the straight experiment.

Taking the top performing classifiers from gender and ethnicity, 3g, 3g4g, 5e, and 3e5e, gender and ethnicity classifications from different feature types were used together. The results can

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
3g5e OR	17,526,217,284	209,666	81.93	3.45
3g3e5e OR	18,434,508,896	213,634	83.48	3.44
3g4g5e OR	18,997,060,648	214,982	84.00	3.42
3g4g3e5e OR	19,678,094,918	216,108	**84.43**	**3.40**

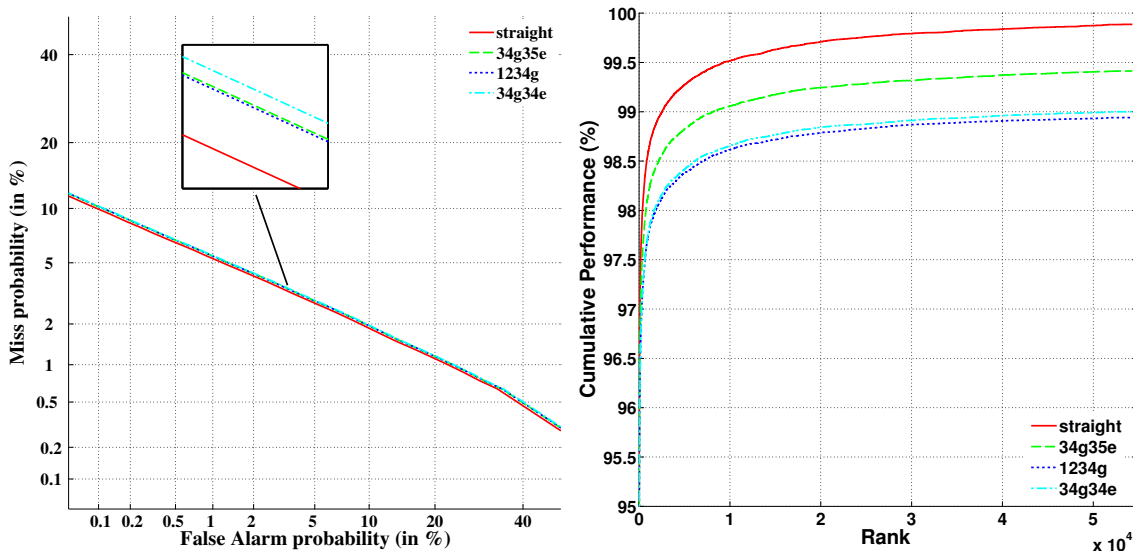
Table 6.12: Weighted sum fusion with mixed texture and shape classifications, best single and double taken for gender and ethnicity. Highest Rank-1 performance and lowest EER are in bold.

be found in Table 6.12. The best result seen so far can be found in these experiments utilizing 3g4g and 3e5e. Replacing 4e with 5e from the all texture combination results in an improvement of less than 0.5%.

One possible way to improve the results would be to compare the predicted class of the probe with the predicted class of the gallery. This would allow someone who is consistently misclassified by the classifier to have the genuine match score included. Using the predicted gallery class provides a 5-7% improvement in Rank-1 when using color gender classifications and an improvement of 20-30% when using color ethnicity. When using both gender and ethnicity, the improvement shrinks to only 3%. Texture shows an improvement of 2-3% on gender and approximately 20% on ethnicity, but less than 1% when utilizing both. Shape experiments improve 8-10% when using the predicted gallery classes. The results using the best classifiers improve less than 0.5%. The best Rank-1 results when using predicted gallery is still 3g4g3e5e, 84.70%, followed by all gender classifiers, 84.69%, and all texture, 84.66%. All these results are still 0.1% below the straight facial recognition system. The related CMC and DET plots can be seen in Figure 6.8.

6.3.2 Score Fusion

Experiments in this section have a slight advantage over those in the previous section. By using the scores returned by the classifier instead of just the decision, the experiment has a little more information with which to make the decision if the probe and gallery entry might be a match. Table 6.13 shows the results of a score-level fusion on color classifications to decide if match scores are used in the recognition experiment. Rank-1 scores are 2% higher using color gender and 5% higher with color ethnicity than the corresponding weighted sum fusion experiments. However, using both color gender and ethnicity classifications, the Rank-1 performance still falls 2% below the straight face experiment. Table 6.14 shows the results using texture scores. Gender results improve less



(a) Verification (b) Identification

Figure 6.8: Best weighted sum fusion experiments compared to the straight experiment. DET is on the left and CMC is on the right.

than 2% from weighted sum, while ethnicity shows an improvement of approximately 10%. Using both gender and ethnicity improves less than 1% from the other fusion method, but the Rank-1 performance is within 1% of the straight face experiment. There are no shape experiments in this section, because this fusion requires scores from an ANN classifier, and the shape classifier used a SVM classifier.

Experiments mixing both gender and ethnicity classification using both color and texture information allow for a small increase as seen in Table 6.15. At this point the number of impostor

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
1g	13,193,346,780	175,740	69.13	3.40
2g	14,258,951,034	176,196	69.06	3.39
1g2g OR	17,003,923,860	203,298	79.57	3.38
1e	7,716,306,302	137,362	53.01	3.72
2e	6,746,179,502	113,470	43.89	3.62
1e2e OR	9,634,780,128	152,988	59.13	3.59
1g2g1e2e (G AND E)	6,934,431,200	144,566	56.03	3.76
1g2g1e2e (G OR E)	19,704,272,788	211,720	82.67	3.37
All (G OR E) predicted gallery	22,587,493,941	217,153	**84.81**	**3.30**

Table 6.13: Score fusion with color information. Highest Rank-1 performance and lowest EER are in bold.

Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
3g	14,303,998,630	192,596	75.58	3.43
4g	14,422,806,168	192,442	75.51	3.45
3g4g OR	17,163,555,578	211,572	82.81	3.42
3e	9,047,946,474	155,040	60.26	3.56
4e	8,455,913,224	143,936	55.78	3.61
3e4e OR	11,400,418,864	177,956	69.16	3.70
3g4g3e4e (G AND E)	8,314,459,576	173,258	67.49	3.72
3g4g3e4e (G OR E)	20,249,514,866	216,270	84.49	3.37
All (G OR E) predicted gallery	22,383,990,928	216,920	**84.83**	**3.30**

Table 6.14: Score fusion with texture information. Highest Rank-1 performance and lowest EER are in bold.

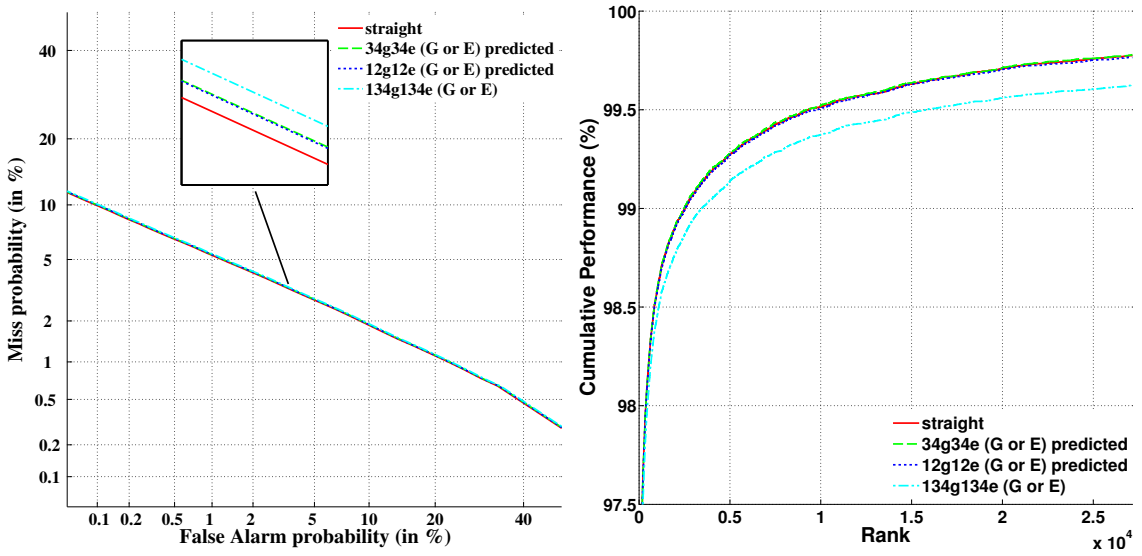
Method	*Impostor Attempts*	*Genuine Attempts*	*Rank-1 (%)*	*EER (%)*
3g4g1e3e4e	20,586,393,362	216,570	84.60	3.36
1g3g4g1e3e4e	21,032,866,890	216,826	**84.70**	**3.35**

Table 6.15: Score fusion with mixed gender, ethnicity, color and texture classifications. Highest Rank-1 performance and lowest EER are in bold.

scores and genuine scores are nearing the numbers for the straight face experiment. As with weighted sum fusion, it is possible that matching to the predicted gallery would increase the Rank-1 performance by allowing more genuine scores to be included. Using all color combinations, the Rank-1 performance is just 0.01% under the performance of the straight experiment, but less than 5% of the impostor scores are actually ruled out from the experiment. Using all texture combinations, the Rank-1 performance improves to just 0.01% over the performance of the straight experiment. Again, approximately 5% of the impostor scores are excluded. The CMC and DET graphs for these experiments can be seen in Figure 6.9. The EERs are not as good for the fusion experiments, and the performance on the CMC is very similar. So the improvement is very small and possibly not worth the cost of four partial face gender and ethnicity classifications.

6.4 Conclusions

While the gender and ethnicity information is useful in decreasing the number of impostor scores, in the proposed fusion experiments, multiple classifications on both the probe and gallery sets are needed to achieve a very slight improvement, 0.01% in Rank-1 and only a 5% reduction in comparisons. If the classifiers were perfect, the increase in Rank-1 in this particular scenario



(a) Verification (b) Identification

Figure 6.9: Best score fusion experiments. DET is on the left and CMC is on the right.

would be slightly larger, 0.47%, but the reduction of comparisons is much greater. Any of the region/feature/classifier combinations by themselves were unable to improve Rank-1 recognition. The small improvement in performance is not worth the cost multiple partial-face classifications in this instance, but this may not be true in all cases. The score-level fusion allowed for better results with the same or fewer classifications than the weighted sum decision fusion.

It is possible that some performance could be gained by making several changes to the proposed fusion methods. The first change was implemented by comparing the probe classification to the gallery classification instead of the ground truth information for the gallery subject. This resulted in a small gain in Rank-1 performance. Another possibility is to compare the probe classification with both the gallery classification and the ground truth and combine those in a weighted sum. The weights could be individually learned in an authentication experiment, so that, for an individual who is generally misclassified more weight is given to the classification comparison and for an individual whose classification is haphazard, the weights could give equal consideration to both. For a verification system, the weights would most likely be learned off-line from a training set or set as a system parameter. The final change would be to start with a partial face experiment as opposed to a whole face experiment. This would be more for the identification experiments than for verification experiments. With less information to make the decision of who an individual is, the

soft biometric information might be able to leverage a larger gain in performance.

Improving the classification accuracy would also help increase fusion experiment performance. In looking at the confusion matrices and misclassification rates for this section and the others, a better representation of the *Asian* and *White* subjects is needed. Class-wise accuracy on these classes is lower compared to the others. Key-point features as opposed to local-appearance based features might be a better fit for these classes. The color and texture features used in this work were extracted using local-appearance-based features. This would also play into using partial-face experiments. If feature extraction could be done without facial alignment, similar to the work by Liao and Jain [42], it would open up a lot more options for biometric applications. The system would not need the whole face and it would not need specific points to reliably extract a certain part of the face. This would allow a more covert acquisition of facial images and help deal with occlusion and image quality. Another method to improve classification accuracy is to incorporate boosting into the training algorithm so that the classifier spends more time learning how to classify difficult subjects and less time on individuals that are easy to classify [21].

This concludes the discussion of fusion experiments performed using partial-face classifications to filter a face experiment. The final chapter will summarize the conclusions made from all experiments and discuss directions for future work.

Chapter 7

Conclusions and Future Work

7.1 Reliability

The first question considered in this work was which parts of the face hold reliable gender and ethnicity information for machine-based classification. To the human eye, different parts of the face can have a feminine or masculine quality which can agree or disagree with a person's identity. The same can occur with ethnicity. Experiments were performed on two datasets, for both gender and ethnicity classification. Regions of the face considered were the left eye, right eye, nose tip, nose, mouth and chin.

From 360 partial-face color experiments on the FRGC and MORPH datasets, the conclusion was made that the mouth and eye regions provide the most gender information. The mouth region achieved similar performance to full face indicating the potential for as much gender information as the whole face. From the same number of ethnicity classifications, the mouth and eye regions also provide the most ethnicity information, again, comparable to face. The best region depends on the particular ethnic group. For most it was the mouth region, but *Asian* class was erratic. The best color space for feature extraction was the RGB color space. The difference between this space and the others was that RGB included color information in all three channels and did not compress it into two. This agrees with what was seen in previous work, that the eye regions are useful for gender and ethnicity classification and that the RG channels from the RGB color space work well.

Partial-face shape experiments for FRGC and MORPH numbered 168 for each trait, gender and ethnicity. Results suggest that partial-face shape is not reliable for either gender or ethnicity.

This is possibly due to feature representation and selection as opposed shape encoding no information. Full face results in previous work were able to achieve similar gender classification performance to texture features. Differences were seen between the gender and ethnic classes, but the overlap between them was too large for classifiers to adequately separate them. The best classifications used all features with no reduction. This is another indication that shape shows promise. It is possible that a better representation or different feature selection could perform better.

In 252 partial-face texture gender experiments over the FRGC and MORPH datasets, it was determined that texture features encode the most gender information in the lower face regions, namely the mouth and the chin. These regions achieved performance comparable to the gender classification of a commercial face system. In the corresponding ethnic experiments, it was determined that the texture of the upper face, namely the eye regions, holds the most ethnicity information. The eye and nose regions were able to achieve similar performance to the texture classification of the entire face. Previous work has shown most regions of the face to be useful for gender classification using either texture or pixels, so these results agree with some and contradict others. Previous ethnic research has shown that the eye region holds discriminating ethnic information, which is confirmed in texture experiments.

Over all 780 gender and 780 ethnicity experiments, texture features encode the most reliable gender and ethnicity information. Not all texture features are equal; however. The best texture representations for this work were LBP and LPQ. The color features encode more soft biometric information than HOG features, but less than LBP and LPQ, indicating that it may be a useful feature when texture cannot be accurately extracted. The best regions for gender classification are the mouth. For ethnic classification, it is the eyes followed by the nose. The eyes perform well in gender classification as well. Region size does not seem to matter although the resolution of the original image does.

Ethnicity is a harder problem than gender classification. Not only does it have more possible classes, the classes are open to interpretation by individual. An individual's self-perception of ethnicity does not always agree with anyone else's or even society's perception of their ethnic identity. This makes classifications of this nature more difficult.

Within these experiments, no clear conclusion could be drawn in regards to the best classification method. Both SVM and ANN classifiers had good performance and bad performance. It seemed to depend more on the feature than on the actual classifier. In most cases, both of these

classifiers outperformed the k-NN classifier.

7.2 Age

The second question considered in this work was the effect of age on the machine classification of gender and ethnicity using partial-face images. The human face changes as it ages, with skin becoming more coarse and gaining wrinkles. MORPH experiments were analyzed here to see if any region/feature combinations were less susceptible to changes in age than others.

From the experiments in Chapters 3, 4, and 5, the conclusion was made that no region or feature remained unaffected by changes in age. The regions including the nose were determined to be the the most stable for gender classification and fairly stable for ethnic classification. The chin region was stable for both as well, but had the lowest performance in ethnicity classification. Ethnicity classification was less impacted by age than gender classification, which follows the trends seen in *previous work* classification involving the whole face. The impact of age on specific features was dependent on the region. No clear conclusion could be drawn on the invariance of specific features to age. Color features did seem less susceptible to changes in age than texture features, but were still negative impacted by age.

Different classes were affected differently by age. The *Female* class was more impacted by age than the *Male* class, especially in the eye regions. Younger subjects were easier to classify in most cases, both gender and ethnicity.

7.3 Application

The third question considered in this work was the extent to which partial-face gender and ethnicity classification could be used to improve the performance in a biometric application. Experiments were performed on one dataset not previously used in this work. Choices on the proposed methods were made based on the experiments in Chapters 3, 4, and 5. Two types of fusion were investigated, one each at the score and decision levels.

While the partial-face gender and ethnicity classifications were useful in decreasing the number of impostor scores, the proposed fusion methods gained little to no improvement in Rank-1 performance as compared to the straight face experiment. An improvement of 0.01% in Rank-1

performance using score fusion only excluded 5% of impostor scores. This contradicts the idea that the soft biometric information used must be independent of the biometric modality, but follows the trend that the more dependent the traits and the modality are, the smaller the improvement in biometric performance. Multiple partial-face classifications of both gender and ethnicity were needed to gain the small improvement in Rank-1. The complexity and number of classifiers most likely outweighs the reduction in comparisons achieved. Other experiments excluded more impostor scores, but this resulted in the exclusion of more genuine matches and decreased Rank-1 performance. Using ground truth information only netted a small gain in Rank-1 performance, 0.47% and resulted in higher EERs.

As it stands, partial-face gender and ethnicity did not improve the performance of the chosen whole-face biometric application much. The conclusion was made that more improvement is possible, just not with the given whole-face application. Several changes were suggested to improve upon the methods given in this work, including using a partial-face recognition system, a single gallery experiment, basing the fusion on the comparison of the predicted probe class to the predicted gallery class, and changing how the soft biometric information is incorporated into the experiment. Other suggestions were based on improving the classification performance such as using key-point based features instead of local-appearance and incorporating boosting algorithms into the training of the classifiers.

7.4 Future Work

Future work directly tied to improving the proposed methods in this work can be divided into two sections: improving the soft biometric classification and improving the performance of the fusion of soft biometric classifications with a biometric application. To improve the classification performance, the use of boosting algorithms during training may prove useful. Also, both classification methods used in this work are machine-learning based, so trying a statistical classification method may provide more insight into the problem. Features used for classification were confined to local-appearance based texture and color features. Key-point features as well as shape features might encode more reliable information for classification. To improve fusion performance, the base experiment can be changed to a partial-face experiment and the fusion method changed from filtering to a feature to be compared.

There are several other topics that would be helpful to pursue but are not directly tied to improving the proposed methods in this work. In the field of biometrics, there is literature on the 'biometric menagerie' or 'Doddington's Zoo' [73]. This classification refers to a subject's likelihood of contributing to the FAR and FRR in an authentication problem. The categories are Sheep, Lambs, Goats, and Wolves. Subjects who are well separated from the others and rarely get rejected are labeled as sheep, whereas subjects who are very hard to recognize and contribute to the FRR are classified as goats. Lambs are subjects who overlap significantly with other users and contribute to the FAR. Subjects classified as wolves are able to pass for other subjects and contribute to the FAR as well. It would be interesting to apply this theory to soft biometric classification. The categories would not be quite the same, since it would be a classification problem instead of an authentication problem, but it would be interesting to look at what characteristics of a particular demographic make it stand apart from the others or what characteristics that individuals within that demographic have that are similar to another demographic. Categorizations of individuals would could help determine weights in an updated fusion scheme.

Intrinsic to a soft biometric zoo problem, as well as to the soft biometric classification problem, is the issue of the demographic labels themselves. Ethnic and gender identities are no longer as black and white as they once were as cultures blend together and technologies advance. For instance, someone's self-perceived ethnic identity could be different from how others perceive their ethnic identity, which is different from their ethnic heritage, which might not fit into any labels that a soft biometric system may have learned to classify. It would be interesting to do a study on the perception of ethnic and gender identity to see if the results impacted the zoo classifications or provided insight as to which labels soft biometric systems should learn.

An evaluation of partial-face soft biometric classification under varying image quality would also be useful. Knowing the minimum quality requirements for a specified performance would provide a guide for researchers to incorporate the soft biometric classifier into their research. Quality could include resolution, image blur, lighting, occlusion, and distortion.

Glossary

k-NN *k-Nearest Neighbor*, simplistic machine learning based classifier.. 25, 36, 49, 112

ANN *Artificial Neural Network* , machine learning based classifier. 11, 27, 28, 36, 111

AR face database collected by Aleix Martinez and Robert Benavente. 7, 8

BioID face database. 7, 8

BioSecure iris database. 7

CAS-PEAL *Chinese Academy of Sciences-Pose, Expression, Accessories, and Lighting*, face database. 7–9

CASIA *Chinese Academy of Sciences Institute of Automation*, iris database. 7, 8

CMC *Cumulative Match Curve*, performance measure used in recognition/identification problems. 30, 31, 93, 99, 103

CMU-PIER iris database collected at Carnegie Mellon University. 7, 8

confusion matrix performance analysis used in classification problems. 28

CV *cross-validation*, experiment technique that does not use specific training and testing sets. 30

DET *Detection Error Trade-off*, performance measure used in verification/authentication problems. 29, 93

EER *Equal Error Rate*, place in a ROC or DET graph where FRR and FAR are equal. 29, 31, 93, 99, 100, 113

FAR *False Accept Rate*, percentage of false matches accepted at a given threshold in a verification/authentication problem. 29, 30, 100, 114

FERET *Face Recognition Technology*, face database. 6–9

FRGC *Face Recognition Grand Challenge*, face database. 7, 8, 15, 16, 18, 19, 23, 36, 60, 74, 102

FRR *False Reject Rate*, percentage of true matches rejected at a given threshold in a verification/authentication problem. 29, 100, 114

HOG *Histograms of Oriented Gradient*, texture feature representation. 10, 11, 72, 89, 111

LBP *Local Binary Patterns*, texture feature representation. 10, 11, 73, 89, 111

LCH *Local Color Histograms*, color feature representation. 11, 32

LDA *Linear Discriminant Analysis*, statistical analysis method for feature reduction which seeks to maximize between-class scatter while minimizing within-class scatter. 25

LPQ *Local Phase Quantization*, texture feature representation. 11, 73, 74, 89, 111

MBGC *Multiple Biometric Grand Challenge*, face and iris database. 7, 8

MORPH *Craniofacial Longitudinal MORPHological Face*, face database. 8, 16–19, 23, 36, 60, 74

PCA *Principal Component Analysis*, statistical analysis method typically used in biometrics for feature reduction. 24, 25, 92

Pinellas face database. 8, 18, 19, 23, 91

ROC *Receiver Operating Characteristic*, performance measure used in verification/authentication problems. 29–31

Softopia Japan face database available at: `http://www.hoip.softopia.pref.gifu.jp/`. 7, 8

SUMS *Standford University Medical Student*, face database. 7, 8

SVM *Support Vector Machine*, machine learning based classifier. 11, 27, 36, 111

UBIRIS iris database (noisy, visible wavelength). 7, 8

UND face database collected by the University of Notre Dame. 7

UPOL iris database collected by researchers at Palacký University Olomouc. 7, 8

XM2VTS *Extended Multi-Modal Verification for Teleservices and Security*, face database available at: `http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/`. 6–9

Bibliography

[1] H. Abdi, D. Valentin, B. Edelman, and A. J. O'Toole. More about the difference between men and women: evidence from linear neural networks and the principal-component approach. *Perception*, 24(5):539 – 562, 1995.

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using local phase quantization. In *Proc. 19th Int. Conf. Pattern Recognition ICPR 2008*, pages 1–4, 2008.

[3] R. Akbari and S. Mozaffari. Performance enhancement of PCA-based face recognition system via gender classification method. In *Proc. 6th Iranian Machine Vision and Image Processing (MVIP)*, pages 1–6, 2010.

[4] A. M. Albert, K. Ricanek, Jr., and E. Patterson. A review of the literature on the aging adult skull and face: Implications for forensic science research and applications. *Forensic Science International*, 172(1):1 – 9, 2007.

[5] Y. Andreu and R. Mollineda. The role of face parts in gender recognition. In A. Campilho and M. Kamel, editors, *Image Analysis and Recognition*, volume 5112 of *Lecture Notes in Computer Science*, pages 945–954. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-69812-8_94.

[6] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. SpringerVerlag, 2003.

[7] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. The common biometrics. In *Guide to Biometrics*, Springer Professional Computing, pages 31–49. Springer New York, 2004.

[8] S. Buchala, N. Davey, R. Frank, T. Gale, M. Loomes, and W. Kanargard. Gender classification of face images: The role of global and feature-based information. In N. Pal, N. Kasabov, R. Mudi, S. Pal, and S. Parui, editors, *Neural Information Processing*, volume 3316 of *Lecture Notes in Computer Science*, pages 763–768. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-30499-9_117.

[9] S. Buchala, N. Davey, R. J. Frank, and T. M. Gale. Dimensionality reduction of face images for gender classification. In *Proc. 2nd Int Intelligent Systems IEEE Conf*, volume 1, pages 88–93, 2004.

[10] D. Cao, C. Chen, M. Piccirilli, D. Adjeroh, T. Bourlai, and A. Ross. Can facial metrology predict gender? In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8, 2011.

[11] C. H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikäinen. (Multiscale) Local Phase Quantisation histogram discriminant analysis with score normalisation for robust face recognition. In *Proc. IEEE 12th Int Computer Vision Workshops (ICCV Workshops) Conf*, pages 633–640, 2009.

[12] C. H. Chan, M. Tahir, J. Kittler, and M. Pietikäinen. Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1164–1177, 2013.

[13] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] S. R. Coleman and R. Grover. The anatomy of the aging face: Volume loss and changes in 3-dimensional topography. *Aesthetic Surgery Journal*, 26(1, Supplement):S4 – S9, 2006.

[15] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *Computer Vision - ECCV'98*, volume 1407 of *Lecture Notes in Computer Science*, pages 484–498. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0054760.

[16] G. W. Cottrell and J. Metcalfe. Empath: Face, emotion, and gender recognition using holons. In *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, 1991.

[17] A. A. Dahl. Heterochromia iridis symptoms, causes, treatments. Online: `http://www.medicinenet.com/heterochromia_iridis/page2.htm`, 2013. Accessed November 2 2014.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, 2005.

[19] M. Demirkus, K. Garg, and S. Guler. Automated person categorization for video surveillance using soft biometrics. In *Proceedings of SPIE*, volume 7667, pages 76670P–76670P–12, 2010.

[20] O. Déniz, G. Bueno, J. Salido, and F. D. la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598 – 1603, 2011.

[21] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001. Chapter 9.

[22] B. Edelman, D. Valentin, and H. Abdi. Sex classification of face areas: how well can a linear neural network predict human performance. *Journal of Biological System*, 6(3):241–264, 1998.

[23] Fast artification neural network library (FANN). Online: `http://leenissen.dk/fann/wp/`. Accessed May 8, 2014.

[24] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Proceedings of the 1990 conference on Advances in neural information processing systems 3*, NIPS-3, pages 572–577, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

[25] G. Guennebaud, B. Jacob, et al. Eigen v3. Online: `http://eigen.tuxfamily.org`, 2010. Accessed May 12, 2014.

[26] G. Guo, C. R. Dyer, Y. Fu, and T. S. Huang. Is gender recognition affected by age? In *Proc. IEEE 12th Int Computer Vision Workshops (ICCV Workshops) Conf*, pages 2032–2039, 2009.

[27] G. Guo and G. Mu. A study of large-scale ethnicity estimation with gender and age variations. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 79–86, 2010.

[28] S. Gutta, J. R. J. Huang, P. Jonathon, and H. 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.

[29] S. Gutta, H. Wechsler, and P. J. Phillips. Gender and ethnic classification of face images. In *Proc. Third IEEE Int Automatic Face and Gesture Recognition Conf*, pages 194–199, 1998.

[30] S. Hosoi, E. Takikawa, and M. Kawade. Ethnicity estimation with facial images. In *Proc. Sixth IEEE Int Automatic Face and Gesture Recognition Conf*, pages 195–200, 2004.

[31] Y. Hu, J. Yan, and P. Shi. A fusion-based method for 3D facial gender classification. In *Proc. 2nd Int Computer and Automation Engineering (ICCAE) Conf*, volume 5, pages 369–372, 2010.

[32] A. Jain, J. Huang, and S. Fang. Gender identification using frontal facial images. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, page 4 pp., July 2005.

[33] A. Jain, K. Nandakumar, X. Lu, and U. Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In D. Maltoni and A. Jain, editors, *Biometric Authentication*, volume 3087 of *Lecture Notes in Computer Science*, pages 259–269. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25976-3_24.

[34] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *International conference on Biometric Authentication*, pages 731–738, 2004.

[35] T. Kawano, K. Kato, and K. Yamamoto. A comparison of the gender differentiation capability between facial parts. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 350 – 353 Vol.1, Aug 2004.

[36] T. Kawano, K. Kato, and K. Yamamoto. An analysis of the gender and age differentiation using facial parts. In *Proc. IEEE Int Systems, Man and Cybernetics Conf*, volume 4, pages 3432–3436, 2005.

[37] A. Lapedriza, M. Marin-Jimenez, and J. Vitria. Gender recognition in non controlled environments. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 834–837, 2006.

[38] A. Lapedriza, D. Masip, and J. Vitria. Are external face features useful for automatic face classification? In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, page 151, June 2005.

[39] G. Li, G. Sun, and X. Zhang. Robust face recognition in low resolution and blurred image using joint information in space and frequency. In J. Park, A. Zomaya, S.-S. Yeo, and S. Sahni, editors, *Network and Parallel Computing*, volume 7513 of *Lecture Notes in Computer Science*, pages 616–624. Springer Berlin Heidelberg, 2012.

[40] Y. Li, M. Savvides, and T. Chen. Investigating useful and distinguishing features around the eyelash region. In *Proc. 37th IEEE Applied Imagery Pattern Recognition Workshop AIPR '08*, pages 1–6, 2008.

[41] H.-C. Lian and B.-L. Lu. Multi-view gender classification using multi-resolution local binary patterns and support vector machines. *International Journal of Neural Systems*, 17(6):479–487, 2007.

[42] S. Liao, A. Jain, and S. Li. Partial face recognition: Alignment-free approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1193–1205, May 2013.

[43] L. Lu and P. Shi. A novel fusion-based method for expression-invariant gender classification. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2009*, pages 1065–1068, 2009.

[44] L. Lu, Z. Xu, and P. Shi. Gender classification of facial images based on multiple facial regions. In *Proc. WRI World Congress Computer Science and Information Engineering*, volume 6, pages 48–52, 2009.

[45] X. Lu and A. K. Jain. Ethnicity identification from face images. *Proceedings of SPIE*, 5404:114–123, 2004.

[46] J. R. Lyle, P. E. Miller, S. J. Pundlik, and D. L. Woodard. Soft biometric classification using local appearance periocular region features. *Pattern Recognition*, 45(11):3877 – 3885, 2012.

[47] F. S. Manesh, M. Ghahramani, and Y. P. Tan. Facial part displacement effect on template-based gender and ethnicity classification. In *Proc. 11th Int Control Automation Robotics & Vision (ICARCV) Conf*, pages 1644–1649, 2010.

[48] J. Merkow, B. Jou, and M. Savvides. An exploration of gender identification using only the periocular region. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–5, Sept 2010.

[49] A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(11):1927–1943, 2007.

[50] *Minority Report*. Dir. Steven Spielberg. Twentieth Century Fox Film Corporation, 2002. Film.

[51] MORPH — IISIS. Online: `http://www.faceaginggroup.com/morph/`. Accessed October 17, 2013.

[52] S. Nissen. Neural networks made simple. *Software Developer's Journal*, 2005. Online: `http://fann.sourceforge.net/fann_en.pdf`. Accessed May 12, 2014.

[53] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[54] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Proceedings of the 3rd international conference on Image and Signal Processing*, ICISP '08, pages 236–243, Berlin, Heidelberg, 2008. Springer-Verlag.

[55] O. Özbudak, M. Kirci, Y. Çakir, and E. O. Güneş. Effects of the facial and racial features on gender classification. In *Proc. MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conf*, pages 26–29, 2010.

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

[57] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 947–954 vol. 1, June 2005.

[58] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, and W. Worek. Preliminary face recognition grand challenge results. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 15–24, 2006.

[59] X. Qiu, Z. Sun, and T. Tan. Global texture analysis of iris images for ethnic classification. In D. Zhang and A. Jain, editors, *Advances in Biometrics*, volume 3832 of *Lecture Notes in Computer Science*, pages 411–418. Springer Berlin / Heidelberg, 2005. 10.1007/11608288_55.

[60] X. Qiu, Z. Sun, and T. Tan. Learning appearance primitives of iris images for ethnic classification. In *Proc. IEEE Int. Conf. Image Processing ICIP 2007*, volume 2, 2007.

[61] A. W. Rawls and K. Ricanek, Jr. MORPH: development and optimization of a longitudinal age progression database. In *Proceedings of the 2009 joint COST 2101 and 2102 international conference on Biometric ID management and multimodal communication*, BioID_MultiComm'09, pages 17–24, Berlin, Heidelberg, 2009. Springer-Verlag.

[62] K. Ricanek and B. Barbour. What are soft biometrics and how can they be used? *Computer*, 44(9):106–108, 2011.

[63] K. Ricanek and T. Tesafaye. MORPH: a longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345, 2006.

[64] Y. Saatci and C. Town. Cascaded classification of gender and facial expression using active appearance models. In *Proc. 7th Int. Conf. Automatic Face and Gesture Recognition FGR 2006*, pages 393–398, 2006.

[65] W. Schwartz, H. Guo, and L. Davis. A robust and scalable approach to face identification. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision - ECCV 2010*, volume 6316 of *Lecture Notes in Computer Science*, pages 476–489. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-15567-3_35.

[66] C. Shan. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4):431 – 437, 2012. Intelligent Multimedia Interactivity.

[67] S. Tamura, H. Kawai, and H. Mitsumoto. Male/female identification from 8x6 very low resolution face images by neural network. *Pattern Recognition*, 29(2):331 – 335, 1996.

[68] V. Thomas, N. V. Chawla, K. W. Bowyer, and P. J. Flynn. Learning to predict gender from iris images. In *Proc. First IEEE Int. Conf. Biometrics: Theory, Applications, and Systems BTAS 2007*, pages 1–5, 2007.

[69] M. Toews and T. Arbel. Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(9):1567 –1581, sept. 2009.

[70] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, 1991.

[71] Y. Wang, K. Ricanek, C. Chen, and Y. Chang. Gender classification from infants to seniors. In *Proc. Fourth IEEE Int Biometrics: Theory Applications and Systems (BTAS) Conf*, pages 1–6, 2010.

[72] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition and gender determination. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pages 92–97, June 1995.

[73] M. Wittman, P. Davis, and P. Flynn. Empirical studies of the existence of the biometric menagerie in the FRGC 2.0 color image corpus. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, pages 33–33, June 2006.

[74] D. Woodard, S. Pundlik, J. Lyle, and P. Miller. Periocular region appearance cues for biometric identification. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 162–169, 2010.

[75] Z. Yang and H. Ai. Demographic classification with local binary patterns. In S.-W. Lee and S. Li, editors, *Advances in Biometrics*, volume 4642 of *Lecture Notes in Computer Science*, pages 464–473. Springer Berlin / Heidelberg, 2007. 10.1007/978-3-540-74549-5_49.

[76] R. Zewail, A. Elsafi, M. Saeb, and N. Hamdy. Soft and hard biometrics fusion for improved identity verification. In *Proc. 47th Midwest Symp. Circuits and Systems MWSCAS '04*, volume 1, 2004.