RAPID DETERMINATION OF AGE CLASSIFICATION

BY

TIMOTHY H BALL

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2011
DOCTOR OF PHILOSOPHY DISSERTATION

OF

TIMOTHY H BALL

APPROVED:

Dissertation Committee:

Major Professor Victor Fay-Wolfe

Timothy Henry

Qing Yang

Lisa C. DiPppo

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2011
ABSTRACT

Law enforcement agencies that employ Computer Forensics teams often find themselves involved in cases of child pornography. Officers who investigate these cases are required to sort through thousands and sometimes tens of thousands of digital images in order to find the evidence they need. This manual search is time consuming and not an optimal use of the investigator’s time. In order to alleviate some of this burden, law enforcement has expressed an interest in an automated tool that can identify potential contraband images.

This dissertation outlines the research and implementation of a method that will aid in the identification of images of interest to computer forensics investigators. This project uses open source libraries to find images with faces, and from those faces, extract the main facial features. To achieve the highest rate of success in finding the face and the facial features, it is necessary for the subject’s full face to be clearly present. It is desirable for there to be very little horizontal and vertical rotation of the head, however, the feature detection method allows for some slight rotation. Calculations are performed on the position of the features resulting in an attribute set which is used by machine learning to determine if the face is that of a child or an adult. This study shows that the ideal separation of these classes is with children 12 years of age and under and adults 18 and over.

The resulting methodology described in this paper achieves an accuracy rate of 70% separation of images into the child and adult classes, and it processes images at a rate of approximately one second per image.
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................. ii
TABLE OF CONTENTS ............................................................................................... iii
LIST OF TABLES .......................................................................................................... vii
LIST OF FIGURES ....................................................................................................... ix

CHAPTER 1 .................................................................................................................. 1
1.1 Problem Statement ............................................................................................... 1
1.2 Goals .................................................................................................................. 1
1.3 Background ......................................................................................................... 2
1.3.1 Conception ...................................................................................................... 2
1.3.1.1 Art ........................................................................................................... 3
1.3.2 Current Solutions ............................................................................................ 5
1.3.2.1 Porn Scanners ....................................................................................... 6
1.3.2.2 Hashing ................................................................................................ 7
1.3.2.3 Project Redlight ................................................................................... 9

CHAPTER 2 ................................................................................................................ 10
2.1 Facial Aging Surveys .......................................................................................... 10
2.2 Facial Recognition ............................................................................................. 12
2.3 Medical Research of Facial Aging ..................................................................... 13
2.4 Related Work ...................................................................................................... 13
2.4.1 Kwon and Lobo ........................................................................................... 14
2.4.2 Modeling and Template Work ..................................................................... 18
2.5 Understanding Haar Classifiers ...................................................................... 25
2.5.1 Alfred Haar .................................................................................................. 25
2.5.2 Wavelets ...................................................................................................... 26
4.6 Resizing the Image ................................................................................................. 50
4.7 Modifying the Classifier Settings ........................................................................ 50
4.8 Picture Comparisons ............................................................................................ 51
4.9 Fixing the Nose .................................................................................................... 52
4.10 Working With The Eyes ...................................................................................... 53
  4.10.1 Images Without Two Eyes ............................................................................. 53
  4.10.2 Face Division Revisited .................................................................................. 55
  4.10.3 Choosing the Classifiers Part 2 ....................................................................... 56
  4.10.4 A Word About Eye Distance .......................................................................... 57
........................................................................................................................................ 59
4.11 Plot of Feature Points .......................................................................................... 59
4.12 New Adult Database ............................................................................................ 60
4.13 Finding the Best Attribute Set ............................................................................ 61
4.14 Results of Attribute Set Tests ............................................................................ 62
  4.14.1 Attribute Set Trials ......................................................................................... 64
4.15 Optimal Set of Attributes ................................................................................... 77
  4.15.1 Alternate Sets of Attributes .......................................................................... 77
4.16 Determining the Age Cutoff ................................................................................ 79
  4.16.1 Image Database Divisions .............................................................................. 79
  4.16.2 Results of DB Division Tests .......................................................................... 79
  4.16.3 Combination Test Sets .................................................................................. 81
  4.16.4 Results of Combination Tests ........................................................................ 82
4.17 Gender and Ethnicity ........................................................................................... 84
4.18 Independent Database Tests ................................................................................. 85
4.19 Performance Analysis .......................................................................................... 87
4.20 Summary of the Results ...................................................................................... 89
4.20.1 Attribute Set ........................................................................................................... 89
  4.20.1.1 Optimal Attribute Set ......................................................................................... 89
  4.20.1.2 Alternate Attribute Sets ...................................................................................... 90
  4.20.2 Accuracy ................................................................................................................ 91
  4.20.3 Speed ...................................................................................................................... 92
CHAPTER 5 ......................................................................................................................... 94
  5.1 Conclusions from Test Results .................................................................................... 94
  5.2 Limitations .................................................................................................................. 94
  5.3 Future Work ............................................................................................................... 96
  5.4 Contributions ............................................................................................................. 96
BIBLIOGRAPHY .................................................................................................................. 100
LIST OF TABLES

Table 1: Accuracy results achieved in [13] ................................................................. 18
Table 2: Haar Face Classifiers ..................................................................................... 45
Table 3: Haar Eye Classifiers ....................................................................................... 46
Table 4: Face Division Accuracy .................................................................................. 49
Table 5: Image Resizing Accuracy ................................................................................ 50
Table 6: Accuracy after Modifying the Classifier Settings .............................................. 51
Table 7: Accuracy with Two Eyes .................................................................................. 54
Table 8: Accuracy after Dividing the Eye Region ........................................................... 56
Table 10: Results of testing individual attributes ........................................................... 64
Table 11: Attributes with the highest overall accuracy rate ......................................... 65
Table 12: Adding the third highest overall accuracy ...................................................... 65
Table 13: Adding the fourth highest overall accuracy ................................................... 65
Table 14: Adding the fifth highest overall accuracy ...................................................... 65
Table 15: Adding the sixth highest overall accuracy ..................................................... 66
Table 16: Attributes with highest accuracy for each class ............................................ 66
Table 17: Adding the next two highest accuracy rates for each class ......................... 67
Table 18: Adding two more attributes to the list ........................................................ 67
Table 19: Continuing the trend of adding two attributes .............................................. 67
Table 20: Final attempt in this pattern ......................................................................... 68
Table 21: Highest overall accuracy combined with highest for each class .................. 68
Table 22: Adding the next highest overall accuracy ..................................................... 68
Table 23: Adding the next highest for adult and child classes ..................................... 69
Table 24: Adding the next highest overall accuracy ..................................................... 69
Table 25: Removing eye to nose/nose to bottom attribute ........................................... 70
Table 26: Measurements with highest overall accuracy .............................................. 70
Table 27: Adding the measurement with the third highest accuracy rate ..................... 71
Table 28: Adding the measurement with the fourth highest accuracy ......................... 71
Table 29: Adding the measurement with the fifth highest accuracy rate ....................... 71
Table 30: Adding the measurement with the sixth highest accuracy rate ....................... 71
Table 31: Adding the measurement with the seventh highest accuracy rate .................. 72
Table 32: Adding the final measurement ..................................................................... 72
Table 33: Measurements with highest accuracy rate for each class ............................. 72
Table 34: Adding next two measurements .................................................................. 73
Table 35: Adding final two measurements .................................................................. 73
Table 36: Combination method for choosing measurement attributes ........................ 73
Table 37: Adding the eye distance measurement .......................................................... 74
Table 38: Adding the nose width measurement resulted in correct classification of more child images ................................................................. 74
Table 39: Adding mouth to bottom measurement ......................................................... 75
Table 40: Measurement-only attribute set with training and test split ......................... 76
Table 41: Combined attribute set with training and test split ....................................... 76
Table 42: Attribute set returning a high percentage of child images ............................ 78
Table 43: Another, less desirable, option for returning a high percentage of child images. This option has many more false positives.......................................................... 78
Table 44: Attribute set that minimizes false positives ................................................. 78
Table 45: Results of comparison between infants and adults ........................................ 80
Table 46: Results of comparison between toddlers and adults ..................................... 80
Table 47: Results of comparison between early childhood and adults ......................... 80
Table 48: Results of comparison between late childhood and adults ........................... 80
Table 49: Results of comparison between preteen and adults .................................... 81
Table 50: Results of comparison between early teen and adults ................................ 81
Table 51: Results of comparison between late teen and adults ................................... 81
Table 52: Comparing infants, toddlers to adults ......................................................... 82
Table 53: Comparing infants, toddlers, early childhood to adults ............................... 82
Table 54: Comparing infants, toddlers, early/late childhood to adults ....................... 82
Table 55: Comparing infants, toddlers, early/late childhood, and preteen to adults ...... 83
Table 56: Comparing infants, toddlers, early/late childhood, preteen, early teen to adults ................................................................................................................. 83
Table 57: Comparing all children under 18 to adult .................................................... 83
Table 58: Testing all children under 18 using the 12 and under model ......................... 83
Table 59: Accuracy of gender ..................................................................................... 84
Table 61: Accuracy of images of Africans ................................................................. 84
Table 62: Accuracy of images of Asians ..................................................................... 84
Table 63: Accuracy of images of Caucasians ............................................................ 85
LIST OF FIGURES

Illustration 1: Guiding lines used to draw a child's face ........................................... 5
Illustration 2: Guiding lines used to draw an adult's face ............................................ 5
Illustration 3: An example of an MD5 hash calculator ................................................... 8
Illustration 4: Samples of age progression prediction from [23] ................................. 11
Illustration 5: Proportions used to classify images in [13] ........................................... 15
Illustration 6: 6-stage process for finding the face region. [13] ................................. 16
Illustration 7: 6-stage process for pinpointing the eye location. [13] ......................... 17
Illustration 8: Facial calculations used in [17] ............................................................. 22
Illustration 9: A representation of the Haar wavelet and the function that defines it .. 26
Illustration 10: Digital music file represented in wave format .................................... 27
Illustration 11: A sample wavelet .................................................................................. 27
Illustration 12: Shading regions of the face [21] ......................................................... 28
Illustration 13: An example of a summed area table entry at point (x,y) ................. 30
Illustration 14: The standard set of basic Haar features .............................................. 31
Illustration 15: Haar features on a face. Images from [30] ............................................ 31
Illustration 16: A graphical representation of a summed area table lookup. ............ 32
Illustration 17: A cascade of classifiers for a region of an image ............................. 33
Illustration 18: Machine learning used to separate data into classes ...................... 35
Illustration 19: Detecting the face and the facial features .......................................... 37
Illustration 20: Facial proportions used for classifying images .................................. 43
Illustration 21: Samples of images where a face was either partially or incorrectly identified ................................................................. 45
Illustration 22: The plot on the left is the 15-stage classifier. On the right is the 20-stage classifier ................................................................. 47
Illustration 23: The 20-stage classifier was much more accurate. ............................ 47
Illustration 24: Dividing the face into search regions for the eyes, nose, and mouth. 49
Illustration 25: Progression of fixes to the detection of the facial features .......... 52
Illustration 26: Sample of pictures before and after implementing the algorithm to find the best nose position ......................................................... 53
Illustration 27: Further dividing the eye area into left and right search regions ....... 55
Illustration 28: Point plots before and after the eye region division and the new eye classifiers. The image on the right shows a much better grouping of the eyes in general. ........................................................................................................... 57
Illustration 29: Before and after images of the eye fixes. The set in the upper right shows a very subtle but more accurate positioning of the eyes. ............................. 57
Illustration 30: Example of how the framing of the face affects the eye distance ..... 59
Illustration 31: Upper row left to right: infant, toddler, early childhood, late childhood. Bottom row left to right: preteen, early teen, late teen, adult................................. 60
CHAPTER 1

INTRODUCTION

1.1 Problem Statement

When a computer forensics team takes on a case involving child pornography, the tasks are many and varied. From obtaining a search warrant, to seizing evidence, to prosecuting any illegal activity, agents are constantly being pulled in many directions during the course of the investigation. One of the most time consuming tasks is searching the seized computers and digital media for contraband images. This media can contain hundreds of thousands of images and not all are even necessarily images of people. Images can include operating system icons and many different types of graphics downloaded and cached by websites.

Computer forensic investigators are few in number and their caseloads can become quite large, so any significant reduction in time spent searching for photographic evidence would be a huge benefit to them. Currently, there is no good solution for finding images of child pornography on suspect media. Because of limited resources and manpower, law enforcement has expressed a need for an automated solution that can quickly determine if these types of images are present.

1.2 Goals

The Digital Forensics Group at the University of Rhode Island first began addressing law enforcement needs by attempting to identify images that contained any
type of pornography in general. This was a very worthwhile endeavor, and though it was automated and provided some reduction in data, it did not quite give law enforcement the full picture. At the time, it was a necessary and welcome solution, but we needed to refine that search to the images that were of greatest interest to the investigating officer. To achieve this, we started looking at ways to enhance what we had already done. In talking to local and federal law enforcements agencies, we were able to determine that what they really wanted was an automated way to quickly search through large quantities of images and find those that were mostly likely to be of interest to them. Their concern was not so much for accuracy as it was for speed. For finding even a few contraband images is enough to prosecute in many cases.

1.3 Background

The following sections describe the high level goals and concepts that were discussed for this project. Our digital forensics group had some general ideas as to what we wanted to achieve. Each section describes how these ideas evolved into more specific concepts of how I might proceed with the research. This background section also summarizes some of the techniques currently used by law enforcement in child exploitation cases.

1.3.1 Conception

In the beginning, our goal was to find a way to combine the technology used to find pornography with something new that would identify children. Combining these two technologies would theoretically give us the image set that would be of most use to forensics investigators. We wanted to do some sort of age classification and
separate the images into classes of child and adult. To do so, we would need to find a way to make some sort of calculations that would give us the desired results. This became my goal – to find a way to quickly determine the age classification of a human subject.

### 1.3.1.1 Art

As we were discussing ideas for accomplishing age classification, I was reminded of an art class I took in grade school in which we were taught how to draw human faces. We were taught that the human face has certain symmetry to it, and that this symmetry is what makes a face look human. As most artists do, we were shown how to use guiding lines to start our sketches of the face. We were also taught that there was a difference in the placement of these guiding lines when drawing a child and when drawing an adult. The human face maintains symmetry as we age, but the proportions change.

There are several significant differences in the way child and adult faces are drawn. The first, and most obvious, is the initial shape of the head. Both heads are sketched with an oval, however, the child's head is more rounded and the adult's head is more oblong [9]. In each case, the next step is to draw a horizontal line through the middle of the oval. This is the line that divides the forehead mass from the facial feature mass. This line is also the guide line for the eyes since the eyes are approximately in the middle of the face.

One important fact about the human face is that the eyes grow very little as we age. The eyeballs themselves remain almost exactly the same size, and any significant change happens very early in age. An infant's eyes look so large, because when
compared to the head, they are in fact proportionally larger. Another interesting fact about the eyes is that a human's eyes are one eye-length apart. So the smaller the eyes are, the closer together they are. [9]

For adults, the eyes are drawn just above this center line, and for children they are drawn either on or below the line. The remaining features - nose and mouth - are evenly spaced in the remaining area under the eyes. Since the child's face is more rounded, there is less area for the mouth and nose, hence the tighter grouping of the facial features of children. The adult's face is more oblong to account for the elongated jawline. The nose and mouth are drawn closer to the chin than the eyes to account for the lengthening of the bridge of the nose. These differences in proportions are what make a face look older or younger, and the symmetry is what makes it look human [9]. This concept is also what led to the initial idea of using facial proportions to make an age classification.

Illustrations 1 and 2 show an example of the guiding lines that are used for drawing the faces of both adults and children.
In [9], Hogarth describes in detail the various parts of the human face and head and how these parts change as a human ages. He also covers the proportions of the human face. He describes how, as the human ages, the nose gets longer, the jawline drops, the mouth region becomes larger, and the chin thicker and fuller. The age progression images used in the book are from a profile view, so while they are informative, the images themselves were not useful in creating frontal face proportions for the purposes of this research. However, the details of how the face changes helped to support ideas garnered from the art class.

1.3.2 Current Solutions

Before coming to a final decision on how to solve this problem presented by law enforcement, it was necessary to investigate what the state of the art is and what the current solutions are. The following two subsections describe the two most
common methods for detecting child pornography. Each section will discuss the pros and cons of each of these methods.

1.3.2.1 Porn Scanners

One of the most common technologies used for finding child pornography is referred to as a "porn scanner". In essence, this is a program that searches for all pornographic images. Porn scanners are widely used in the public domain by businesses, schools, parents, etc. They are fairly good at detecting pornographic images, and in their common usage, blocking these images. Most of these scanners use some combination of features such as limb detection, skin tone detection, and face detection to classify images as pornographic or not.

Another solution in this same vein is skin tone detection. Skin tone detection is one of the pieces of a porn scanner; however, there are many forensics tools that use this as the sole means of identifying pornographic images. It is not as effective as the porn scanners at identifying pornography, it is fast, but it also produces many more false positives than a porn scanner. Any object at all, whether human or not, that has a skin tone color, would be identified as a positive result.

There are two common types of porn scanners on the market today – web-based and disk-based. Web-based applications can be considered porn blockers since they typically block pornographic images from appearing in the browser. The two leading web-based filters are iShield [10] and Safe Eyes [24]. They are similar in functionality, but slightly different in methodology. Where Safe Eyes uses mostly skin tone detection, iShield uses a combination of skin tone, edge detection, and other features along with decision making to determine if an image is pornographic.
Disk-based porn scanners perform a slightly different function in that they scan physical media for pornographic images. One common use of these programs is clearing a disk of unwanted images; however, they are also used in several forensics tools to identify pornographic images during an investigation. The technologies used in these programs are very similar to those used in the web-based applications. For instance, Snitch [25] and X-Ways Forensics [32] use skin tone detection to find pornographic images. SurfRecon [27] and ADF Solutions [1] use hash values to detect pornographic images. The use of hash values in detecting pornography is described in more detail in the next section.

The biggest drawback with these applications is that they result in a very high number of false positives for law enforcement criteria. False positives are all images that are not child pornography, including adult pornographic images. Therefore, these types of solutions are not ideal for data reduction in child exploitation cases.

1.3.2.2 Hashing

The hashing method uses the individual bits in a data file to calculate a hash value for that file. There are a few common algorithms for calculating the hash value of a data file. SHA1 and MD5 are two of the more widely used hashing methods in law enforcement, and both employ the same basic concept. The main difference between the two is that MD5 produces a 128-bit value, or digest, and SHA1 produces a 160-bit digest [12].
Regardless of the hash method chosen, there are certain conditions that must be met in order for a hash function to be effective. The hash function has to be complex enough to make collisions nearly impossible. A collision, in terms of hash values, happens when two different data files produce the same hash value. At the same time, the hash value has to be easy to compute. It should also not be possible to reproduce the original file content from the hash value itself. And lastly, any change at all to the data file should result in a significant change to the hash value [12].

There are many uses for hash values, but in the context of the fight against child pornography hash values are used to store information about known contraband images. Hash values for images previously recovered are calculated and stored in very large databases. One such database is maintained by the National Center for Missing and Exploited Children [19].

In order to find contraband images, hash values are taken of all image files on the suspect digital media and are then compared to a database of hash values of known child pornography images. Calculating the hash value of a file is relatively fast, as is comparing the value to the database. Therefore, this method of detecting known contraband images is very fast and is not prone to a large number of false findings. However, the hash method will not find images that are not currently in the database. Any new images will not be detected. Also, the hash value of an illegal image can be

Illustration 3: An example of an MD5 hash calculator

Regardless of the hash method chosen, there are certain conditions that must be met in order for a hash function to be effective. The hash function has to be complex enough to make collisions nearly impossible. A collision, in terms of hash values, happens when two different data files produce the same hash value. At the same time, the hash value has to be easy to compute. It should also not be possible to reproduce the original file content from the hash value itself. And lastly, any change at all to the data file should result in a significant change to the hash value [12].

There are many uses for hash values, but in the context of the fight against child pornography hash values are used to store information about known contraband images. Hash values for images previously recovered are calculated and stored in very large databases. One such database is maintained by the National Center for Missing and Exploited Children [19].

In order to find contraband images, hash values are taken of all image files on the suspect digital media and are then compared to a database of hash values of known child pornography images. Calculating the hash value of a file is relatively fast, as is comparing the value to the database. Therefore, this method of detecting known contraband images is very fast and is not prone to a large number of false findings. However, the hash method will not find images that are not currently in the database. Any new images will not be detected. Also, the hash value of an illegal image can be
changed by simply changing one bit in the data file making this method very easy to circumvent.

1.3.2.3 Project Redlight

Project Redlight [29] is URI’s pornography detection tool, and is an ongoing project in the URI Digital Forensics Center. Currently, Redlight detects pornography through a method similar to that employed by iShield described in a previous section. However, Redlight uses fewer image attributes for training the machine learning making it much faster at processing images. The intent of this project is that it will be integrated into Redlight such that the combination of pornography detection and child detection will result in a tool that will detect child pornography. However, neither the creators of Redlight nor I will test these projects on actual child pornography images.
CHAPTER 2

REVIEW OF LITERATURE

Once the initial idea of using facial proportions to determine age was verified as viable by the concepts of art, the next step was to find further research that corroborated this idea. The following sections describe the research areas that supported the proposed solution and led to a more comprehensive understanding of the problem in general and to possible solutions.

2.1 Facial Aging Surveys

I researched multiple papers in an attempt to further understand the changes that take place as we age vis-a-vis the face and head. There were two studies in particular that provided the most support to using facial proportions for age classification and can be found in [14] and [23]. Both of these papers provided a survey of facial aging and current work being done in the fields of facial age estimation and facial age progression.

Facial age estimation is the art of attempting to determine the approximate age of a subject as closely as possible. On the other hand, facial age progression attempts to predict what a person will look like at a particular age, typically an older age. Both areas of research are interesting, however the studies of facial age estimation, to a certain extent, are more in line with this project. That being said, the field of facial age progression offers insight into how the face changes as a person ages, so it too is
relevant. Facial age progression has some practical applications as well. In [23], the idea is presented that facial age progression can be used in cases of missing children. The appearance of a child who has been missing for a number of years can be estimated using these techniques and helpful in identifying the missing child.

The first known project that attempted facial age estimation was performed by Kwon and Lobo in [13]. Since the release of this study, others have attempted to replicate and improve on their methods, but mostly by using modeling techniques.

There are two basic methods for processing images of faces, and Kwon and Lobo presented the first such method. They used the facial features of eyes, nose,
mouth, and chin to create a set of proportions that would describe the general appearance of the face be used to estimate the age. They also used facial wrinkles in determining the classification. The work they did in this area is closely related to this project and will be described in more detail in the Related Work section. After their study, most other researchers went with the modeling technique for age estimation.

A typical modeling solution usually starts by building a collection of facial models at varying ages. After the models are built, a newly unclassified face is compared to the patterns and classified according to the model to which it most closely matches. In most instances, the modeling techniques require more time and processing than would be desirable for a law enforcement application. The focus of these methods is on accuracy over speed, so again this is contrary to the requirements and goals of this project. Several of these projects are presented in the Related Work section along with the specific references for each.

2.2 Facial Recognition

Another area that has seen a growing amount of research in the past decade is facial recognition. In [33], Zhao provides a survey of the work being performed in this field. Facial recognition is loosely related to facial age estimation in that they both attempt to do some type of characterization of the face, either through modeling or feature comparisons. However, in facial recognition programs, the launching point is an image of a face. There is prior knowledge of a face. With facial detection, this is not the case. An image could contain any type of object, and it is up to the application to determine if a face is present.
2.3 Medical Research of Facial Aging

One of the leading experts in the field of child pornography is Dr. Carole Jenny. In cases involving child exploitation, she is one of the most sought after expert witnesses when the case goes to trial. In [11], she describes the process she uses to determine if the subject appearing in a digital image is a child or adult. For the purposes of this project, most of the methods she employs can't be used since they deal largely with areas of the anatomy that would be inappropriate and illegal to analyze. However, in her paper she describes the changes that occur in the face as a child ages. Most notable among these are the changes that take place with respect to the nose, mouth, and jawline. She notes that the nose lengthens, the mouth becomes fuller, and the jawline and chin drop. This is a confirmation of the techniques being used when drawing faces of children and adults.

2.4 Related Work

There are two basic methods for processing images when working with faces. The first method is using templates or modeling in order to achieve some goal in processing the face, and the main goal is to achieve the highest accuracy possible. Typically this method is used for facial recognition techniques [33]; however, recently these methods have been used in facial aging [14]. The second basic method for working with faces is object detection. This technique is mostly used for facial detection applications, and the main focus is on the speed of detection [30, 34].
Likewise, there are two general fields of study in the area of facial aging - age determination and age progression [14]. Kwon and Lobo are recognized as being the first to study age determination in faces in digital images. Their work was based mostly on modeling the face and using proportions to determine the age classification of the human subject.

2.4.1 Kwon and Lobo

As mentioned previously, Kwon and Lobo are credited with having done the first work in facial age estimation in [13]. They showed that making an approximate age determination from facial features is possible. Up to this point, no one had done any work in the area of determining the age of a face in an image. Using measurements of facial features and calculating proportions, they were able to separate images of faces into three classes: infant, adult, and senior.
In order to achieve this class separation, their particular solution included some human interaction to pinpoint the middle of the face. From there, they used computer modeling and repeated edge detection to find the contours and features of the face. They also used facial wrinkles to help in the age determination. The requirements for human interaction, the amount of time for modeling the face, and images of such high resolution as to pick out facial wrinkles are not ideal for this project. Despite that, knowing that facial features can be used in making an age classification is confirmation of the working premise of this project.

*Illustration 5: Proportions used to classify images in [13]*
Illustration 6: 6-stage process for finding the face region. [13]
The other disadvantage of this method is that they were only able to distinguish between infants and adults. That separation does not meet law enforcement criteria since the subjects of child exploitation can range up to pre-teen ages.

One other point of interest is that they used the distance between the eyes as one of their main determining factors. As mentioned before, this distance changes very little as we age, especially after the first few years of a child's life. This may be one reason why their separation of ages was limited to infants and adults. As encouraging as it was to know that age classification based on facial feature proportions is possible, this solution was not fast enough, was not fully automated, and didn't offer a desirable age classification from a law enforcement point of view.

*Illustration 7: 6-stage process for pinpointing the eye location.* [13]
As the above table shows, Kwon and Lobo used a threshold to make their classifications. It is also interesting to note that in many cases, as the accuracy for adults increases, the accuracy for infants decreases. This is a point that will come up again during my own testing and warrants its own discussion at that time.

### 2.4.2 Modeling and Template Work

Following Kwon and Lobo's work, there have been a series of studies done using the template and modeling methods for increasing the accuracy of age classification. This section provides a brief synopsis of this work and the achievements of the various studies.

After Kwon and Lobo completed their work, others began to attempt to improve on it by using different techniques in modeling the face or building templates of the face. The first such example is from 2002 and was performed by Lanitis et al.
In this work, there were several goals, one of which was attempting to "guess" the age of a human subject. To achieve this goal, the authors used a modeling technique in which they cropped faces of individuals at varying ages. They cropped the images in such a way as to obtain a consistent placement of the 50+ landmarks that they pinpointed on the face. Using these landmarks, they then created a function of how the landmarks transformed as the subject aged. They created this model for multiple test subjects until a standard set of functions was calculated for each age. They then created the same model for an unseen face and attempted to match it to a model of an existing subject. Once a match was found, they applied the formula for the existing model to the new model and calculated the age. Using this method, they were able to calculate the age of a subject to within 6 years of the actual age. This means that a subject could be classified as either 6 years older or younger - a 12 year difference.

The next example is a project by Geng et al. in 2007 [6]. In this study they generated templates, or patterns, of faces at various ages similar to the previous work by Lanitis et al. Once these patterns had been built for each age group, the next step was to build a pattern of a new face and attempt to match it to one of the existing patterns. Using this technique, they would classify the new image based on the closest pattern match. The difference here is that there was no calculation used to classify the face; it was simply matched to the closest pattern. However, creating the patterns themselves required significant mathematical calculations as well as cropping the images to obtain a standard appearance of each face used in creating the pattern. Across all ages, they were able to achieve an age approximation with an error of again
6 years. They then divided their image test set into age groups of 0 - 5, 6 - 30, and 31 - 69. For the first age group, they were able to calculate the age to within 1 year of the correct age. The second group was calculated to within 4.5 years of the correct age. Finally, for the last group, they calculated the age to within 8 years. However, it should be noted that for each test, the age group of the unseen image was already known and was tested against the corresponding group. This is an important distinction, since in the case of the project proposed here, there is no prior knowledge of the age of any subject.

In 2008, Fu and Huang [5] proposed a very similar solution using patterns they referred to as manifold structures. The difference here is that the manifold structure is constructed first by using facial detection software to obtain the face region. However, even though this part of the process is automated, the image of the face is then cropped and centered in such a way as to present a standard positioning of the features. The image then goes through multiple linear regressions using a quadratic function in order to obtain the manifold. The results of this particular method of pattern matching are consistent with the previous results of around 5 to 6 years of error in age estimation.

Also in 2008, Suo et al. [26] proposed a hierarchical modeling method using varying resolutions to obtain the features that are then compiled together to build the models for classification. A low level resolution is used to obtain the general appearance of the facial features. A middle resolution is used to obtain a more refined positioning of the features. Finally, a high resolution is needed for facial wrinkles, skin textures, and hair lines. All of these levels constitute the hierarchical model of the face. The models are built using the landmarks on the face extracted at each of the
three levels. However, there is prior knowledge of approximately where these landmarks appear. The age range of the test subjects is 9 to 89 years old. The end result of this study is again consistent with the other modeling techniques and achieves an error rate of 5 years, or a 10 year span. The test subjects were then divided into age groups of under 20, and then a grouping every 10 years. Using this division, the authors were able to obtain an error as low as 4 years in some cases, however, there was again prior knowledge of the approximate age of a test subject and the subject was tested against the corresponding age range. Not only that, but the first age group included children and adults, so it would not work for a law enforcement tool.

In 2009, Ni et al. [20] made an improvement to the above techniques by automating the image cropping process. They also expanded the image database by performing several searches using the Google search engine and different search parameters. They then attempted to filter out images that weren’t ideal for their study by performing multiple passes of the images with different facial detection software. Once they built the database, their actual modeling technique was very similar to previous works using multi-instance regression. Given the fact that the images were not hand-cropped, and therefore had less precision, the outcome of this project was a larger margin of error, most significantly in the 0 – 9 year age range. In this age range, the authors were only able to achieve a 10.98 year margin of error. Given this high margin of error, an 8-year-old child could conceivably be classified as either a newborn or an adult. Additionally, the authors again used the classification technique of limiting the age estimation to the age range in which the subject of the image
belongs. The improvements in this project were detecting faces and automating the cropping, but there was still a reliance on having knowledge of the age of the subject.

In [17], Nasir Memon of New York University gives a high level presentation of a project that includes several slides pertaining to the distinguishing features of a face that would aid in the classification of children and adults. These features included the position of the eyes, nose, and mouth. The method also seems to require the exact location of the ears, eyebrows, and chin.

In a related news article [8], Memon claimed that by using the distance between eyes and the nose and other facial features, he can achieve about 70% accuracy in identifying children. However, without any details as to how the features are detected or what measurements and proportions are used, it is difficult to draw any meaningful conclusions or comparisons. There is also the matter of the specifications of the database that was used to derive the accuracy rate. As will be demonstrated later in this paper, the contents of the image database can have a profound effect on the outcome of the tests.

*Illustration 8: Facial calculations used in [17]*
In almost all cases, in the above mentioned research, there are several problems which must be addressed in order to meet the requirements set forth by law enforcement. First, even though most of these projects claim to be automatic, they are not necessarily automated. This is an important distinction since most of these methods required some sort of human interaction, whether from cropping the images or pinpointing certain regions of the face. Moreover, the goal of the work described in these projects attempted to improve on the accuracy of each previous work. The speed of the classification was not a determining factor in choosing the method in which the age classification was determined. Further, in many cases, the work described above required images of a relatively high resolution in order to obtain some, if not all, of the information needed to construct the models.

Another problem that needs to be addressed is that in several of the above studies, the age divisions did not correspond to a specific child and adult division. In order to achieve an age classification that is meaningful to a child exploitation investigation, the age classification must be able to make this differentiation. While some of the above methods did in fact address this division, each of the methods suffered from at least one shortcoming that would not completely satisfy law enforcement criteria.

In general, these methods of determining age estimation are based on images in which a face is known to exist. That is to say, there is a prerequisite that an image contains a face. For law enforcement purposes and requirements, this is not an ideal situation since much of the digital media seized in an investigation will contain many images of objects that are not necessarily human. Therefore, some type of object
detection must be performed in order to determine if there is a face in any digital image. Also, in several of these projects, the authors knew the approximate age of the subject and attempted to fit that image into the appropriate, pre-determined age category. In other words, they used faces in a particular age range to train, and then used faces in that same age range to “guess” the age of the test subjects. Again, this is not ideal for law enforcement since there will be no prior knowledge of the age of the subject of an image.

From the time Kwon and Lobo first attempted to determine an age classification of a face in a digital image, all of the work that has followed has been evolutionary to some extent. Most studies have tried to improve on one or two areas of that work in order to improve the accuracy of estimating the age of human face. This project seeks to be not evolutionary, but revolutionary by exploring new methods for age classification. To date, there has not been any work done in facial age determination using the object detection method. Since the main goal of this project is speed over accuracy, and given that there has already been work done in modeling and templates, using object detection was the most obvious choice for achieving the speed that was required. Not only that, but by using object detection, this project provides new and innovative research and techniques in the field of facial age determination. The following sections describe in detail the progression of the object detection methodology and technology and how it supports the goals of this project.
2.5 Understanding Haar Classifiers

Early on in my research into different methods of face detection technologies, I discovered that some interesting work was being done with Haar classifiers. In an attempt to more fully understand this field, I researched the technology and theory of Haar classifiers from the origins to their current use in facial detection. The following sections describe Haar classifiers in detail and how they relate to this work.

2.5.1 Alfred Haar

In order to understand Haar classifiers, it is necessary to go back to the very beginning and look at the work done by the Hungarian mathematician Alfred Haar himself. Alfred Haar is credited with creating the first wavelet, and wavelets are the basis for the classifiers that are used in detecting objects in digital images.

Not only was Alfred Haar's wavelet the first wavelet, it was also the simplest possible wavelet. Even though it is considered the simplest wavelet, it does have practical applications. The one of most interest for this project, and for all the following research and development in object detection using the Haar wavelet is the fact that it is very good at finding sudden changes in signals [7]. The significance of this will be explained in later sections.

Below is a representation of the Haar wavelet and the function that describes it. Images are courtesy of the Wikimedia Commons.
To understand the role that wavelets play in the Haar Classifiers, it is important to have at least a rudimentary understanding of wavelets themselves. I devised the following example to not only explain in basic terms what a wavelet is, but also to illustrate how they will be useful in classifying objects in digital images.

The image below shows an analog wave representation of a digital music file.

\[
\psi(t) = \begin{cases} 
1 & 0 \leq t < 1/2, \\
-1 & 1/2 \leq t < 1, \\
0 & \text{otherwise.}
\end{cases}
\]

Illustration 9: A representation of the Haar wavelet and the function that defines it

2.5.2 Wavelets

To understand the role that wavelets play in the Haar Classifiers, it is important to have at least a rudimentary understanding of wavelets themselves. I devised the following example to not only explain in basic terms what a wavelet is, but also to illustrate how they will be useful in classifying objects in digital images.

The image below shows an analog wave representation of a digital music file.
Now assume we want to determine if this music file contains a C# note. We will say that hypothetically the C# note is represented by the wavelet pictured here and is defined by some cosine function as its basis function.

Once the wavelet has been defined, it is then a matter of scanning the music file for an instance of that note, looking for a match to that defined curve. This is one example of how a wavelet is defined by a basis function and how it could be used in a real world application.
2.5.3 Wavelets as Classifiers

One of the first instances of using the wavelet concept for object classifiers comes from [21] that is a project by Papageorgiou in which they use differences in intensities to identify faces in images. The wavelets are defined by the changes in the pixel intensities in key locations on the face. These basis functions are calculated using conjoining dark and light regions of the face.

They used a series of positive images, images of faces, to train the classifier to look for contrasting areas of dark and light that are common to all faces - regions such as the eyes, nostrils, and lips. They then used negative images, or images not containing a face, to complete the training of the classifier. Using this technique, they were able to consistently find the images that contain faces.

There are a couple of very significant developments that came from this study. First, it was possible to train a classifier to find a particular object in an image. And second, given a random image, it was possible to determine if the object being sought was in that image. Up to this point, much of the research in face processing and facial recognition work relied upon the fact that the image contained a face. With this new
technique, it was possible to determine whether or not there was a face in the image at all.

2.5.4 Viola-Jones

In [30], Viola and Jones took the work done by Papageorgiou et al. a step further. What they proposed was a two-fold solution to not only make the object detection more accurate, but also much faster.

2.5.4.1 Integral Image

The first improvement that Viola and Jones made was to create an integral image. By definition, the integral image is a summed area table that stores the sum of all pixel values in a certain range. In this case, the range is determined by an (x,y) pixel reference point. The sum of the values of those pixels in the rectangle defined by (0,0) and (x,y) is stored in the table at location x,y. For example, consider the pixel at location (x₀,yₘ). The sum of the pixel values is:

\[ \text{sum}(x₀,yₘ) = \sum (xᵢ, yⱼ) \text{ where } i \leq n, j \leq m \]

Creating the table can be done in linear time since it is just a matter of adding the previous sum to the next pixel value:

\[ \text{sum}(xₙ,yₘ) = \text{valueAt}(xₙ,yₘ) + \text{sum}(xₙ₋₁,yₘ) + \text{sum}(xₙ,yₘ₋₁) - \text{sum}(xₙ₋₁,yₘ₋₁) \]

The value at \( \text{sum}(xₙ₋₁,yₘ₋₁) \) is calculated twice; therefore one of these sums is subtracted. A graphical representation of this concept would look like this:
The calculation of the integral image and the subsequent look ups during image processing can all be done in constant time making this method extremely fast.

2.5.4.2 Haar Classifiers

Haar classifiers are based on the Haar wavelet described in previous sections. Recall that the Haar wavelet was useful in identifying sudden transitions in signals. Similarly, the Haar classifiers are used to find differences in intensities between two or more conjoining regions in an image.

Illustration 13: An example of a summed area table entry at point (x,y)
The value of the feature is the difference between the summed pixel values in the shaded and non-shaded regions. Referring back to the Haar wavelet in figure 9, the similarity between Haar's first wavelet and the Haar-like features is very obvious.

Using the Integral Image, or summed area table, finding the sum of the pixel values in any rectangular region is simply a matter of looking up four numbers and doing some simple calculations. For example, if we take one of the standard Haar features and place it at some random position in an image, it may look something like this:

Finding the value of any rectangular region in the image is simply a matter of looking up four values in the summed area table. Figure 16 is a graphical representation of this concept.
The value at position 4 would be calculated as $4 + 1 - 2 - 3$. [33]

The final step would simply be to take the difference between the two regions bounded by the feature. This number is then compared to the threshold in the trained classifier in order to determine if this is an area of interest.

2.5.4.3 Haar Cascades

In order to get more accurate results in the search for objects in an image, more complex classifiers are needed. However, as the complexity of the classifier increases, so does the amount of time needed to do the calculations. This negatively affects the near real-time processing speed that is advantageous when scanning a large number of images. In order to resolve this problem, Viola and Jones proposed a system of cascading classifiers [30].

The idea of cascading classifiers is that the scan of a region would first be performed with the basic simple classifiers. If an area of an image is identified by a classifier, it is then scanned with progressively more complex classifiers until it either
fails at some point or passes all classifier tests. In this manner, large portions of images can be rejected earlier in the detection process. The complexity of a classifier is determined by the manner in which it is trained and the number of Haar features used. For example, a simple classifier might use only the two-rectangle features, while increasingly more complex classifiers would use the three- and four-rectangle features. Also, each subsequent classifier in the cascade could be trained with false positives from the previous classifier as the negative training set, thereby ultimately reducing the number of overall false findings. By scanning regions with this method, it is possible to very quickly discard large areas of an image that don't contain the object that is being sought. A cascade of classifiers would look something like this:

![Illustration 17: A cascade of classifiers for a region of an image](image)

Based on all the previous research, it seemed that the logical choice for face and facial feature detection would be the Haar classifiers, and specifically, the Haar cascades proposed in [30]. The Haar cascades not only provide the object detection, but do it very quickly.
2.6 Technologies

The following sections describe the options for technologies that could be used to perform the image processing necessary to complete the task of finding faces and facial features. The second section touches on the options for machine learning that will be used to make the age classifications.

2.6.1 Image Processing

There were two primary choices for the image processing. I originally considered cxImage [22]. However, cxImage is used mainly for image transformations and not for extracting data. Our forensics group discovered that many groups who were involved in the area of pornography detection were using OpenCV. OpenCV was developed by Intel and provides computer vision libraries for real-time image processing. The main advantage of OpenCV is the fact that it uses Haar Cascades for object detection and has a built-in demo of face detection [2].

2.6.2 Decision Making

Part of the automation of the age classification is to have some type of decision making process. The three primary choices considered for this process are support vector machines (SVM) [4], linear discriminant analysis (LDA) [16], and thresholding [13].

2.6.2.1 Machine Learning

SVM and LDA are both types of machine learning that can be trained to classify data based on an attribute set. They are similar in that they both attempt to use the attribute set to find the optimal separation of the data into classes, but they are
different in the methods that they use to achieve this separation. Figure 18 shows how this separation is achieved.

Illustration 18: Machine learning used to separate data into classes

The method used by Support Vector Machines (SVM) to classify data is by use of support vectors. Support vectors are vectors calculated at the time that the SVM is trained and are derived from the points in each class that are closest in value to each other. These vectors describe a cushion, or separation, between the classes. The vectors are recalculated during training to optimize this cushion to be as large as possible while producing the fewest number of falsely classified data points. Linear Discriminant Analysis (LDA) on the other hand uses linear combinations of the data to find the separation of the classes. Thresholding is similar to LDA, but the thresholds typically have to come from human inspection of the data in order to find the separation point between the classes.
CHAPTER 3

METHODOLOGY

This chapter introduces the technology used in this project and how that
technology was employed to solve the problem. The first section describes in general
terms how the challenge presented by law enforcement was solved. The remaining
sections cover in more detail how exactly that challenge was met. The basic
methodology for classifying faces as either child or adult seems fairly straightforward
when described at a high level. In essence, it is simply a matter of finding a face in an
image; finding significant features of that face: eyes, nose, and mouth; calculating
some data points to be used in the attribute set; and finally using this attribute set with
a decision making process to make the age determination. The following sections
describe each step of this process in more detail.

3.1 OpenCV

As mentioned previously, there were a few choices for the image processing. I
chose OpenCV for image processing given the fact that it is widely used in computer
vision, uses Haar Cascades, and came with a sample program for face detection. I used
Microsoft's Visual Studio [18] as my programming environment with the OpenCV
libraries loaded into my project file. I started by loading, compiling, and running the
face detection sample program to see if I could indeed detect faces from sample
images. Using this sample as a base, I did some extensive editing to this code in order
to accomplish what I needed to do.
I began by adding in functionality that would call the Haar Cascades for the eyes, nose, and mouth. I made many improvements to how and where these classifiers searched for these features, and I will discuss that in much more detail in later sections. I also describe the tests I performed to find the best classifier for each feature.

Illustration 19: Detecting the face and the facial features

3.2 Calculating Proportions

After determining that I could find the face and the facial features that I needed, I calculated a few initial proportions to see if the numbers indicated that there was a separation between the classes. The cascades return a rectangular area that contains the feature. I used this information to make my calculations. Early results seemed to support the idea that there is a determinable difference between the child and adult classes.
3.3 Machine Learning

When it came time to choose the method for decision making, I ultimately went with LDA. There are several reasons for this. First, the proportions, or attributes, are continuous in nature. A person's face grows gradually as they age. Faces age and develop slowly over a long period of time indicating that there is some threshold at which a separation between childhood and adulthood could be detected in the attribute values. There was a relevant quote from the LDA page of Wikipedia to support this choice: "LDA works when the measurements made on independent variables for each observation (attribute) are continuous quantities." Second, LDA is faster. Not only is it faster to train, it runs faster. From working with SVMs in previous projects, I know that even the linear algorithm takes more time to make decisions than what I observed with LDA. And in those other projects, I was working with fewer attributes and a smaller data set than I was here with LDA. Lastly, I also drew on the experience of the Forensics group, which came to the same conclusions. In fact, code for LDA came from group development. LDA can accomplish the thresholding demonstrated by [13], but does it for all attributes in the set and not just each individual one.

For this project, the LDA code was developed using the two-class implementation, since we were working with a separation between child and adult. In this particular variation, LDA is trained with samples from both classes. It then calculates a vector onto which the multi-attribute samples are projected. This vector is defined by a linear combination of the attributes, and the classes are divided by a perpendicular to the vector. This perpendicular provides the best separation of the
classes and becomes the threshold to which new samples will be compared. A new sample is then compared to the threshold and classified according to which side it lands on [16].

3.4 Result Set

Since this project was meant to be used as a piece of a larger program, there are two result sets. The first result set is a list of file names and the associated attribute set. Each file name in this list is an image in which a face, eyes, nose, and mouth were found. The list is comma-separated with the file name first and then each attribute in the set following. This format is the standard output for many of the other programs designed by our forensics group.

The second result set is a group of four files. These files are simply lists of file names and are a result of passing the previous file as an argument to the LDA code. The four files are: a list of images classified as child, a list of images classified as adults, a list of images falsely classified as child, and a list of images falsely classified as adult. These files helped in the analysis phase of this project to determine which images were being correctly classified and which weren't.
CHAPTER 4

FINDINGS

Once I had determined that I could find faces and facial features using OpenCV and Haar Cascades, I started the process of finding the best set of classifiers. I then proceeded to fine-tune those classifiers to find the features in the correct place. After that, I calculated several proportions and measurements on the face and the facial features. I ran various combinations of these numbers through the machine learning algorithm in order to get the best separation of data possible. This chapter describes in detail the tests I ran and the findings that resulted from those tests.

4.1 Tests

The following tests were performed to improve many areas of the program. The first thing I needed to do was to test my program on some real faces, not the computer generated type I had been working with up to this point. I then researched the Haar classifiers and found a very informative paper [3] that did a comprehensive study of classifiers, which ones worked well and which ones didn't. I tested these classifier to verify the findings myself, and to come up with the best set of classifiers for my purposes. I also found a paper that had some suggestions for not only speeding up the detection process, but also for increasing the accuracy of finding the facial features [31]. I spent a significant amount of time testing different ways to implement this paper's suggestion of dividing the face into regions. I describe that process in more detail in a later section. During the course of these tests, I made some discoveries.
that helped increase the speed and accuracy; I will include those discoveries in the relevant sections. Finally, I needed to test the accuracy itself. There are three types of accuracy: finding faces, finding features, and classifying images. I will address each of these.

4.2 Image Databases

The initial testing I did was with a few images of computer generated faces. I wanted to verify that the feature detection would work with ideal faces first. The next step was to test it on real faces, and for this I used a set of random images that we had collected in our forensics group. After these tests proved to be promising, our forensics group met to formalize and define a new image test set. This new set divided images into age groups, ethnicity, and gender. Once testing was complete with this database, to further verify the results, I then tested against an independent image database. These databases are described in more detail in the following sections.

4.2.1 Initial Image Database

Our group decided on a general "Adult" category subdivided into Caucasian, African, and Asian. Those categories were further separated into male and female. The same was done with children with the additional division of ages. The child category was all children under the age of 18. Children 3 years old and older were divided by each year: 3, 4, 5, etc. Also, children under 3 were divided every six months: 0 - 6 months, 6 - 12 months, etc. Again, these age groups were subdivided into ethnicity and then gender.
I decided early on in the testing that I would use children 12 and under for my test group. I kept adults as 18 and older. There are two reasons for doing this. First, according to research [9, 23], after the age of 12, a child's face starts to become much more adult-like in appearance. This was proven to be true in later tests. Second, in many child pornography cases, investigators look for images of younger children. The reason for this is that even for a human, it gets more difficult to determine if older children are indeed children or just very young looking 18 and 19-year-olds.

For each of the following tests, I used a Python script to split my attribute files into training and test sets. The script chooses a random subset of images for training and testing. I verified that the files chosen for training were representative of the entire set - each age group, race, and gender were represented. I did some experimentation with the test set in which I chose only images that had ideal attribute sets, but this resulted in misclassification of more images due to the application not being able to recognize some of the borderline cases. It was better to have a pseudo-random test set in which there was more variance of the attributes.

4.2.2 FGNet Database

The FGNet database has been used in several previous facial feature analysis projects and literature, including the research by Ramanathan described in [23]. This database is also currently being used by our forensics group in a study of the Haar Classifiers [29]. This image set will be used as an independent database to verify the results of this project.

The FGNet database contains images of full-frontal faces that have been pre-classified by age. It also has manually plotted points on these faces that can be used to
make any number of calculations regarding the facial features. As well, the faces in this database show age progression by capturing the subjects at various ages. Therefore, the same human subject appears multiple times in this image set. Because of this feature, this particular database has been used in studies pertaining to age progression.

4.3 Original Set of Attributes

Based on Kwon and Lobo's work in [13], and from my own research, I came up with the following set of attributes that I felt were indicative of how a person's face changes with age.

- Eyes to Top and Bottom
- Eyes to Nose to Mouth
- Eyes to Mouth to Bottom
- Eyes to Nose to Bottom
- Area of Eye to Area of Face

Illustration 20: Facial proportions used for classifying images

4.4 Choosing the Haar Classifiers

Once I had an image database to work with and an initial set of proportions, I set out to determine which of the Haar Cascades would give me the most accurate
position of the facial features. In [3], Castrillon-Santana et al. performed a very good analysis of publicly available Haar Classifiers for facial features. They even trained some of their own classifiers and provided them for public use. In order to validate their work, and to find the set of classifiers that best suited my needs, I ran my own series of tests for each of the facial features using different combinations of classifiers.

As I mentioned previously, Haar classifiers are trained with a series of positive and negative images, and have general qualities that describe how they are trained. The positive images are pictures of the object of interest, and the negative images are pictures of anything else. Ideally, all of these images are the same size. So the first attribute of a Haar classifier is the size of the images that were used to train the classifier. It is not uncommon to see a Haar classifier name with this information included, i.e. "eyes45x11". This indicates that this is a Haar classifier used to identify eyes and was trained with images of size 45x11. The other bit of information that is useful when choosing a classifier is the number of stages. The number of stages is the number of cascades used for the classifiers. Again, this information may appear in the name of the classifier, such as "nose20stages". Again, this would indicate that this is a classifier for the nose and that it employs 20 cascades.

4.4.1 Faces

The following table shows the Haar classifiers I tested against my image database and the number of faces found by each.

<table>
<thead>
<tr>
<th>Haar Classifier</th>
<th>Number of faces found</th>
</tr>
</thead>
<tbody>
<tr>
<td>frontalface_default</td>
<td>4360</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>frontalface_alt</td>
<td>4025</td>
</tr>
<tr>
<td>frontalface_alt2</td>
<td>4162</td>
</tr>
<tr>
<td>frontalface_alt_tree</td>
<td>3085</td>
</tr>
</tbody>
</table>

Table 2: Haar Face Classifiers

My initial inclination was to use the frontalface_default classifier since it found more faces than the other classifiers. However, upon further inspection of the results, I found that this classifier was returning more false faces than the others. In the end, I opted for the frontalface_alt2 classifier which offered a balance between faces found and the fewest false positives. In later tests I also found that the alt2 classifier resulted in slightly better accuracy by about 1.5%.

Illustration 21: Samples of images where a face was either partially or incorrectly identified

4.4.2 Eyes

Locating the eyes in the correct position turned out to be one of the biggest challenges of this project. I will cover the exact steps I took to meet this challenge, but at this initial stage of testing, I was only concerned with finding the most sets of eyes possible and in the general area of where they should be. This table shows the results of the tests I conducted on the eye classifiers.
<table>
<thead>
<tr>
<th>Haar Classifier</th>
<th>Eye Pairs Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes22x5</td>
<td>0</td>
</tr>
<tr>
<td>Eyes45x11</td>
<td>0</td>
</tr>
<tr>
<td>Eye</td>
<td>3185</td>
</tr>
</tbody>
</table>

*Table 3: Haar Eye Classifiers*

At this point, I was using one search and one classifier to find both eyes. So it didn't seem likely that the left and right eye classifiers would be of much use. For this testing phase I did not use them. After running the tests on the other classifiers, I chose to use the "Eye" classifier. In later tests, I opted for the Left and Right Eye MCS classifiers. I will explain that decision in the section where I make that transition.

4.4.3 Nose

Finding the nose in the correct position also turned out to be very challenging and was compounded by the fact that there was really only one classifier worth using - the nose20stages. The nose.xml - or 15 stage classifier - did not return very accurate results as the two images of plots will show. In further tests, I will show how the nose continued to be a problem and my eventual solution.
There was only one mouth to classifier available, and it was the one trained by Castrillon-Santana et al. In most cases it located the mouth in the correct position, and the mouth turned out to be the least troublesome feature. There were other cases when once I made an improvement for finding one of the other features, the mouth location was improved as well.

4.4.5 Independent Haar Classifier Verification

As mentioned previously, there is another ongoing project in our forensics group that is testing the performance of the Haar classifiers [29]. In this paper, Tanner is using the FGNet database’s manually plotted points to generate a series of facial
proportions and then using the methodology and Haar classifiers from this study to calculate the automated proportions. Tanner performs an in-depth statistical analysis of using the manually plotted points versus the automated points found by the Haar classifiers. Results of these tests shows that the Haar classifiers perform with a relatively low margin error when compared to the more precisely plotted points. This independent verification supports the usage of the Haar classifiers for this project, and details of the tests will be presented with all other test results in later sections.

4.5 Face Division

Now that I had chosen the Haar classifiers, I began to make changes to my code that would hopefully result in finding the features with more precision. The first thing I did was to follow the example of Wilson in [31] and search for the facial features inside the face region. That is to say, I would not search for a face in the image, and then search the entire image again for the eyes, nose, and mouth. Wilson suggested using the region of interest returned by the face classifier as the new image in which to search for the facial features. This process would provide me with the face cropping that other projects were using, but would do it automatically with no human intervention.

Wilson also further suggested that the accuracy could be improved by dividing the face region such that the classifiers for each feature would be searching the general area where those features should be found. For example, searches can be conducted for the eyes in the upper region of the face, the nose in the middle, and the mouth near the bottom. Not only did this help locate the features with more accuracy, but also
reduced the processing time since smaller regions were being searched. It took several attempts to find the appropriate dividing lines for each section of the face. The search area needed to be large enough for the classifiers to be able to find the object, but not so large that objects were found in the wrong location.

The top of the face bounding box is 66% of the total face height and forms the search area for the eyes. The bounding box for the nose is 33% to 66%. And finally, the bounding box for the mouth is the lower 60% of the face region.

![Illustration 24: Dividing the face into search regions for the eyes, nose, and mouth.](image)

This table shows the results of the tests run with the chosen set of classifiers and after dividing the face into search regions.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>897</td>
<td>462</td>
<td>66.00%</td>
<td>65.81%</td>
</tr>
<tr>
<td>Adult</td>
<td>75</td>
<td>43</td>
<td>63.56%</td>
<td></td>
</tr>
</tbody>
</table>

*Table 4: Face Division Accuracy*
4.6 Resizing the Image

The next improvement to the program was a suggestion from our forensics group. Through work they had done on their own with Haar classifiers in [29], they found that they were able to achieve better results by resizing the images to a standard size. 350x350 was the ideal size suggested by the group. This size seemed to be the best compromise between being large enough for the classifiers to find the features and not so large as to cause a substantial loss of definition. This table shows the results of running the tests after resizing all images to the standard 350x350 size.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>1211</td>
<td>370</td>
<td>76.60%</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>72</td>
<td>74</td>
<td>49.32%</td>
<td>74.29%</td>
</tr>
</tbody>
</table>

Table 5: Image Resizing Accuracy

There are a couple of interesting things that should be noted. First is the overall accuracy that jumped to 74% from the previous 65%. The other interesting result is that the number of faces in which full feature sets were found has increased for both children and adults. 222 additional images were found for children and 28 for adults.

4.7 Modifying the Classifier Settings

When using Haar classifiers, there are several settings that can be adjusted. I found that the "flags" setting is the one of most interest for this particular application. The default setting is for the classifier to return a vector of objects that it found.
However, I discovered through experimentation that for the mouth, it worked best for the classifier to return the largest object found in the mouth region. I also limited the eye classifier to finding only two objects in the eye region. In many cases, this helped to locate the eyes more precisely, but there were times when it would find the same eye twice in slightly different locations. This is due to the fact that the default setting for "flags" causes the classifier to scan the image several times at various scales. The location of the eyes wasn't as precise as I would have liked, but it was much better, and the accuracy of the program increased.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>1259</td>
<td>370</td>
<td>77.29%</td>
<td>75.75%</td>
</tr>
<tr>
<td>Adult</td>
<td>78</td>
<td>70</td>
<td>52.70%</td>
<td></td>
</tr>
</tbody>
</table>

*Table 6: Accuracy after Modifying the Classifier Settings*

### 4.8 Picture Comparisons

The following set of images shows the progression of fixes I made to the code to locate the facial features in a more accurate position. From left to right, the pictures are: after dividing the face, after resizing the image, and after modifying the classifier settings.
4.9 Fixing the Nose

Finding the nose in the correct position turned out to be much harder than I expected. This may be due to the fact that the nostrils are the only areas of the nose with a high degree of light and dark contrast. Depending on shading and shadows, there are other areas of the face, even in the region near the nose, which can be mistaken for a nose. I tried setting the classifier to find the largest object as I did with the mouth, but this didn't produce consistent results.

To solve the problem of locating the nose in the correct position, I had to set the classifier to return the vector of noses that it found, and then I had to determine which of those was nearest the spot where the nose should be. I wrote an algorithm
that compared the center of each nose to the center of the face and chose the nose that was closest to the center. Since I had limited the search area for nose to the middle of the face, I was confident that this would give me an accurate location for the nose. As it turned out, in almost every case the nose was in the correct position.

Illustration 26: Sample of pictures before and after implementing the algorithm to find the best nose position

4.10 Working With The Eyes

After fixing the nose and getting more accurate results for that feature, I went back to the eyes to see if I could get better results there too. The eyes are used in calculating all of the attributes, so I felt it was of utmost importance to get those as accurate as possible.

4.10.1 Images Without Two Eyes

I made the decision to eliminate images in which both eyes could not be found. I knew that this would reduce the number of total images in the result set; however, having two eyes was necessary for calculating a more accurate proportion. From the research I performed, I knew that the distance between the eyes was the length of one
eye. So it follows that the distance between the centers of the eyes is two eye-lengths. Upon inspection of the images and the circles I was drawing around the eyes from the results of the classifier, I noticed that the circle was not consistent. Sometimes the circle would capture just the iris of the eye, and at other times it would even include part of the eyebrow. So the circumference of these circles was not a reliable source for calculating the length of the eye that I needed for my proportions.

I addressed this problem by using the center of the circle as the center of the eyes. In most cases, this center point was very accurate. By forcing the program to return only images with two eyes, I was able to use half the distance between the centers of the eyes as the length of the eye when calculating my proportions.

Limiting the results to images with two eyes did indeed reduce the data set. However, as it turned out, there were only 109 fewer child images - a reduction of less than 7%. Those images were evenly distributed between the correctly and incorrectly classified images. There were only 7 fewer adult images found.

There was a trade-off between the slightly fewer number of images found and the increase in accuracy. The increase in accuracy here was due to the fact that the proportions were now more accurate with the new eye measurement. In the end, I decided that the increase in accuracy was more important than losing a relatively few number of images. As it turns out, in later tests this proved to be the best decision.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>1203</td>
<td>317</td>
<td>79.14%</td>
<td>76.46%</td>
</tr>
<tr>
<td>Adult</td>
<td>67</td>
<td>74</td>
<td>47.52%</td>
<td></td>
</tr>
</tbody>
</table>

*Table 7: Accuracy with Two Eyes*
4.10.2 Face Division Revisited

The first thing I did to improve the accuracy of the eyes was to divide the eye search area vertically. I also tightened up the horizontal line as well. I moved the horizontal line from 66% of the image to 55%. I then modified the code to search for an eye in each of those new regions. At this time, I also changed the "flag" setting of the classifier to search for the largest object since now there was no reason to return more than one eye per region. The most significant result of doing all this was that the program was now finding one eye on each side of the face. It was also finding more images with two eyes than it had previously. And finally, in many cases, it was finding the eyes in a more accurate position.

Illustration 27: Further dividing the eye area into left and right search regions
4.10.3 Choosing the Classifiers Part 2

At the time that I was making the changes to the search method for the eyes, I realized that it might be worthwhile to investigate the possibility of using the left and right eye classifiers. I figured that I now had left and right regions of the face in which to search for eyes, so it would be interesting to see if the left and right eyes classifiers would make a difference.

After inspecting the results of the tests, what I found was that there were several cases where the eyes were found in a more correct position. The greatest benefit however, was that it was again finding more images with two eyes. This helped to make up for the reduction of images that occurred when I eliminated images without two eyes.

The accuracy of the program was much better now too, with a near 78% accuracy rate. It is important to note however, that in this and each of the previous tests half of the adult images are being misclassified. This was addressed later on by adding more and different attributes.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>1258</td>
<td>310</td>
<td>80.23%</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>71</td>
<td>71</td>
<td>50.00%</td>
<td>77.72%</td>
</tr>
</tbody>
</table>

*Table 8: Accuracy after Dividing the Eye Region*
A Word About Eye Distance

There is an interesting fact about the distance between the eyes that came out in this study. It’s fairly obvious that the face changes very gradually as we age. This is especially true for the eyes. According to the research papers and medical journals referenced previously, the size of the eyes change very little as we grow older. This is why an infant’s eyes appear much larger than an adult’s. In relation to the size of the

Illustration 28: Point plots before and after the eye region division and the new eye classifiers. The image on the right shows a much better grouping of the eyes in general.

Illustration 29: Before and after images of the eye fixes. The set in the upper right shows a very subtle but more accurate positioning of the eyes.

4.10.4 A Word About Eye Distance

There is an interesting fact about the distance between the eyes that came out in this study. It’s fairly obvious that the face changes very gradually as we age. This is especially true for the eyes. According to the research papers and medical journals referenced previously, the size of the eyes change very little as we grow older. This is why an infant’s eyes appear much larger than an adult’s. In relation to the size of the
face and head, an infant’s eyes are larger. From this fact, it follows that the area of the eye in relation to the area of the face is larger. One of my attributes was based on this proportion. It also stands to reason that since the distance between the eyes is in general one eye length that this measurement would only grow very slightly with age, but it would grow.

In later tests, I used the distance between the eyes as one of my measurement attributes and in the calculation of two other proportion attributes. The proportions I calculated using the distance between the eyes are Area of Face to Area of Eyes and Eye Diameter to Face Width. Ideally, these attributes should have resulted in better accuracy than they did. However, there is the small matter of how the face is detected and the region that is returned. The resizing of the image also has to be taken into account.

When the Haar classifier frames the face, it attempts to create a region that bounds the eyes, nose, and mouth. The classifier does not take into account the actual facial region, meaning that it will not frame the forehead, chin, or cheeks. Because of the way the face is framed by the classifier, there is a difference in the distance of the face in the frame. A child's face will appear closer than an adult’s since the adult facial features are more spread out. Because of this zoom, the actual distance between the eyes is counter-intuitive.

One would expect that the distance between the eyes of an infant would be slightly smaller than that of an adult. As presumed previously, this distance between the eyes would be an insignificant change since the eyes grow very little with age. Given this fact, it is also presumed that the distance between the eyes would not be a
good determining factor between children and adults. However, because of the zoom, there is a large enough difference in the distance between the eyes for this to be used as one of the attributes.

The counter-intuitive part comes into play with the actual measurement. Since the features of a child's face are a tighter grouping, the frame returned by the facial detection classifier is smaller. The inverse is true for adults. The upshot is that the measured distance between the eyes of a child is actually larger on average than that of an adult. The following diagram shows a sketch of how the difference faces are framed and why the distance between the eyes of a child appears to be larger when in actuality it is smaller.

Illustration 30: Example of how the framing of the face affects the eye distance

4.11 Plot of Feature Points

The following plots show the migration of features at the various age groups. Special attention should be paid to the fact that as age increases the nose (center blob) tends to move closer to the mouth (bottom blob), and the mouth tends to move closer to the bottom of the bounding box. Also note that the eyes (upper blobs) do not move
vertically as much as research indicated. This is due entirely to the manner in which the face regions are captured by the facial detection.

Illustration 31: Upper row left to right: infant, toddler, early childhood, late childhood. Bottom row left to right: preteen, early teen, late teen, adult.

4.12 New Adult Database

As I mentioned previously, the database of adult facial images was much smaller than that of the child faces. Up to this point, when I was testing the classifiers, the difference in database sizes didn’t have that great of an impact since I was merely trying to find the best set of classifiers by tracking the overall accuracy and how well each classifier worked on faces in general. However, when it came time to work with the proportions of the faces and making age classifications, it was more appropriate to have databases that were equal in size. I felt this would give me a better understanding of how the two groups are being divided and hopefully lead me to finding out why half of the adult images were being misclassified. To this end, I obtained a large database of adult faces that our group had used in previous work.
Now that I had two result sets of the same size, I used a Python script to randomly select a training and test set for the LDA machine learning portion. After several attempts at customizing the training set, what I learned was that randomly selecting the training sets produced better results. This is due to the fact that a random training set is more representative of the whole result set. When I tried to customize the training set, I wasn't taking into account some of the fringe images, the ones that were harder to classify, and this led to many more images being misclassified.

**4.13 Finding the Best Attribute Set**

As I mentioned, in the previous tests half of the adult images were being misclassified. With a new adult database of images, this reduced the accuracy rate to about 64%. I decided to create more proportions in an attempt to increase the accuracy of the program and obtain a better division of the two age classes. Some of the new proportions were based on observations of the plot points and how they migrated with age. It was also at this time that I realized that since I had resized the face area to a standard size, I could now use pixel measurements between the features as part of my attribute set. The list below shows the original and the new attribute sets.

**Original Proportions**

- EyesToTopBottom
- EyesToNoseToMouth
- EyesToMouthToBottom
- EyesToNoseToBottom
- AreaEyeToAreaFace
New Proportions

- Eye Diameter to Face Width
- EyesToMouth NoseToMouth
- EyesToMouth EyesToNose
- NoseToMouth MouthToBottom
- NoseToMouth NoseToBottom

Measurements

- Distance Between the Eyes
- Nose Width
- Eyes To Nose
- Eyes To Mouth
- Nose To Mouth
- Eyes To Bottom of Image
- Nose To Bottom of Image
- Mouth To Bottom of Image

4.14 Results of Attribute Set Tests

At this point, I felt that the attribute set was very comprehensive and representative of how the face changes with age. Some of the attributes were inspired by Kwon and Lobo; others were based on my own observations. The pixel measurements were based on the fact that the images were now the same size and dimension. The next step was to come up with the combination of attributes that resulted in the best accuracy and the best division of the classes. In order to determine which attributes were contributing to the correct division of the two classes and which
ones weren't, I tested each attribute individually. The table below shows the results of those tests. Close inspection of the results of the tests will also show that in many cases, as the accuracy for the child class increased, the accuracy for the adult class decreased. A similar phenomenon was documented by Kwon and Lobo in [13] and can be seen in Table 1. The fact that this relationship exists will be instrumental in determining which attributes contribute most to the overall classification of the two classes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToTopBottom</td>
<td>Child</td>
<td>425</td>
<td>916</td>
<td>31.69%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>1103</td>
<td>238</td>
<td>82.25%</td>
</tr>
<tr>
<td>EyeDiamToFaceWidth</td>
<td>Child</td>
<td>390</td>
<td>951</td>
<td>29.08%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>939</td>
<td>402</td>
<td>70.02%</td>
</tr>
<tr>
<td>EyesToNoseToMouth</td>
<td>Child</td>
<td>1102</td>
<td>239</td>
<td>82.18%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>415</td>
<td>926</td>
<td>30.95%</td>
</tr>
<tr>
<td>EyesToMouthToBottom</td>
<td>Child</td>
<td>996</td>
<td>345</td>
<td>74.27%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>604</td>
<td>737</td>
<td>45.04%</td>
</tr>
<tr>
<td>EyesToNoseToBottom</td>
<td>Child</td>
<td>1034</td>
<td>307</td>
<td>77.11%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>624</td>
<td>717</td>
<td>46.53%</td>
</tr>
<tr>
<td>AreaEyeToAreaFace</td>
<td>Child</td>
<td>390</td>
<td>951</td>
<td>29.08%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>939</td>
<td>402</td>
<td>70.02%</td>
</tr>
<tr>
<td>EyesToMouthNoseToMouth</td>
<td>Child</td>
<td>1127</td>
<td>214</td>
<td>84.04%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>362</td>
<td>979</td>
<td>26.99%</td>
</tr>
<tr>
<td>EyesToMouthEyesToNose</td>
<td>Child</td>
<td>567</td>
<td>774</td>
<td>42.28%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>1071</td>
<td>270</td>
<td>79.87%</td>
</tr>
<tr>
<td>NoseToMouthMouthToBottom</td>
<td>Child</td>
<td>851</td>
<td>490</td>
<td>63.46%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>576</td>
<td>765</td>
<td>42.95%</td>
</tr>
<tr>
<td>NoseToMouthNoseToBottom</td>
<td>Child</td>
<td>792</td>
<td>549</td>
<td>59.06%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>622</td>
<td>719</td>
<td>46.38%</td>
</tr>
<tr>
<td>EyeDistance</td>
<td>Child</td>
<td>390</td>
<td>951</td>
<td>29.08%</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>Child</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>-------</td>
<td>-------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td></td>
<td>939</td>
<td>910</td>
<td>402</td>
<td>431</td>
</tr>
<tr>
<td><strong>NoseWidth</strong></td>
<td>486</td>
<td>658</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EyesToNose</strong></td>
<td>70.02%</td>
<td>67.86%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EyesToMouth</strong></td>
<td>36.24%</td>
<td>49.07%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NoseToMouth</strong></td>
<td>52.05%</td>
<td>52.05%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EyesToBottom</strong></td>
<td>57.12%</td>
<td>56.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NoseToBottom</strong></td>
<td>49.07%</td>
<td>49.96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MouthToBottom</strong></td>
<td>37.06%</td>
<td>36.24%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Results of testing individual attributes

4.14.1 Attribute Set Trials

Previously, I had achieved an accuracy rate of approximately 78%, but the program was classifying half of the adult images incorrectly, and there were far fewer adult images than child images. The first solution was to add more adult images; however, there was still the problem of classifying almost half of the adult images incorrectly. To set about fixing this problem, I experimented with different combinations of attributes. This section summarizes those experiments and the results of the trial tests.

The first combination I tried was with the two attributes that had the best overall accuracy. For each iteration of this trial, I added the attribute with the next
highest overall accuracy rate. I continued this process until I could no longer achieve a better result.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom EyesToNoseToBottom</td>
<td>Child</td>
<td>538</td>
<td>803</td>
</tr>
<tr>
<td>Adult</td>
<td>1141</td>
<td>200</td>
<td></td>
</tr>
</tbody>
</table>

*Table 10: Attributes with the highest overall accuracy rate*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom EyesToNoseToBottom EyesToNose</td>
<td>Child</td>
<td>722</td>
<td>619</td>
</tr>
<tr>
<td>Adult</td>
<td>1024</td>
<td>317</td>
<td></td>
</tr>
</tbody>
</table>

*Table 11: Adding the third highest overall accuracy*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom EyesToNoseToBottom EyesToNose EyesToMouthEyesToNose</td>
<td>Child</td>
<td>745</td>
<td>596</td>
</tr>
<tr>
<td>Adult</td>
<td>994</td>
<td>347</td>
<td></td>
</tr>
</tbody>
</table>

*Table 12: Adding the fourth highest overall accuracy*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom EyesToNoseToBottom EyesToNose EyesToMouthEyesToNose EyesToMouthToBottom</td>
<td>Child</td>
<td>777</td>
<td>564</td>
</tr>
<tr>
<td>Adult</td>
<td>963</td>
<td>378</td>
<td></td>
</tr>
</tbody>
</table>

*Table 13: Adding the fifth highest overall accuracy*
At this point, it did not appear that the accuracy was going to get any better. In fact, the trend was that the accuracy was decreasing and the number of false positives was increasing. The one promising result was that the adult images were no longer being misclassified with the high percentage as in previous cases. However, the child images were now suffering from a higher rate of misclassification. I switched tactics and decided to start with the two attributes that had the highest accuracy for each class. That is to say that I started with the attribute that classified the most children correctly and the attribute that classified the most adults correctly. Following this pattern, in each iteration I added two more attributes.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToBottom NoseToBottom</td>
<td>Child</td>
<td>975</td>
<td>366</td>
</tr>
<tr>
<td>EyesToNoseToBottom</td>
<td>Child</td>
<td>975</td>
<td>366</td>
</tr>
<tr>
<td>EyesToNose</td>
<td>Adult</td>
<td>759</td>
<td>582</td>
</tr>
<tr>
<td>EyesToMouth EyesToNose</td>
<td>Adult</td>
<td>759</td>
<td>582</td>
</tr>
<tr>
<td>EyesToMouthToBottom MouthToBottom EyesToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 14: Adding the sixth highest overall accuracy*

*Table 15: Attributes with highest accuracy for each class*
Table 16 shows that by using these two attributes, the division of the classes has reversed from the results in Table 14.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToBottom NoseToBottom</td>
<td>Child</td>
<td>941</td>
<td>400</td>
</tr>
<tr>
<td>EyesToMouth EyesToNose EyesToMouthNoseToMouth</td>
<td>Adult</td>
<td>805</td>
<td>536</td>
</tr>
</tbody>
</table>

*Table 16: Adding the next two highest accuracy rates for each class*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToBottom NoseToBottom</td>
<td>Child</td>
<td>923</td>
<td>418</td>
</tr>
<tr>
<td>EyesToMouth EyesToNose EyesToMouthNoseToMouth EyeDist EyesToMouth</td>
<td>Adult</td>
<td>882</td>
<td>459</td>
</tr>
</tbody>
</table>

*Table 17: Adding two more attributes to the list*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToBottom NoseToBottom</td>
<td>Child</td>
<td>954</td>
<td>387</td>
</tr>
<tr>
<td>EyesToMouth EyesToNose EyesToMouthNoseToMouth EyeDist EyesToMouth NoseToMouth EyesToMouthToBottom</td>
<td>Adult</td>
<td>834</td>
<td>507</td>
</tr>
</tbody>
</table>

*Table 18: Continuing the trend of adding two attributes*
Table 19: Final attempt in this pattern

Table 18 shows a very good division of the two classes. After that point, adding more attributes showed that the accuracy wasn't getting any better, and the statistics were just moving slightly back and forth. Even though these results were an improvement, I decided to try one more run of trials, but this time trying a combination of both of the previous attempts. For this next series, I started with the attribute with the highest overall accuracy combined with the attributes with the highest accuracy for each class.

Table 20: Highest overall accuracy combined with highest for each class

Table 21: Adding the next highest overall accuracy
Table 22 shows a very good division of the two classes; however I was still concerned about the overall accuracy rate. Each class was still only being correctly classified at around 60%. The optimal solution would be to have a higher overall accuracy while maintaining a good division of the classes.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom</td>
<td>Child</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouth EyesToNose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToNoseToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouth NoseToMouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyeDist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 22: Adding the next highest for adult and child classes**

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom</td>
<td>Child</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouth EyesToNose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToNoseToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouth NoseToMouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyeDist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 23: Adding the next highest overall accuracy**

At this point in the testing I was starting to understand how each attribute was affecting the number of positives and negatives for each class. I took out the attribute for the eye to nose/nose to bottom proportion. The reason for doing this was to try to get better accuracy for the child class and reduce the number of false negatives. Table
shows that this was successful. When comparing tables 22 and 25, not only has the overall accuracy increased, but the number of correctly classified images in both categories has increased as well. I experimented with substituting in a few other attributes, but this result was the best I could achieve.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouthEyesToNose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToMouthNoseToMouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyeDist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToNose</td>
<td>Child</td>
<td>949</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>888</td>
<td>453</td>
</tr>
</tbody>
</table>

Table 24: Removing eye to nose/nose to bottom attribute

After these attempts, I began to notice that the attributes based on the actual pixel measurements seemed to be contributing significantly to the accuracy. This is most likely due to the fact that the proportions were calculated using these same measurements. In some cases, the proportions and the measurements even conflicted, resulting in all of the images of one class being incorrectly classified. I decided to run some tests using just the measurements. I performed these tests using the same methods as before. I started with the two measurements that had the best overall accuracy and added one more attribute with each iteration.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes to Nose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>870</td>
<td>471</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>879</td>
<td>462</td>
</tr>
</tbody>
</table>

Table 25: Measurements with highest overall accuracy
Once again, using just these two measurements, I was able to get a very good division of the classes. However, as was the case presented in Table 22, the overall accuracy is still lower than I would like.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes to Nose</td>
<td>Child</td>
<td>813</td>
<td>528</td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth to Bottom</td>
<td>Adult</td>
<td>938</td>
<td>403</td>
</tr>
</tbody>
</table>

*Table 26: Adding the measurement with the third highest accuracy rate*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes to Nose</td>
<td>Child</td>
<td>813</td>
<td>528</td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td>Adult</td>
<td>935</td>
<td>406</td>
</tr>
<tr>
<td>Mouth to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 27: Adding the measurement with the fourth highest accuracy rate*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyes to Nose</td>
<td>Child</td>
<td>808</td>
<td>533</td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Mouth</td>
<td>Adult</td>
<td>939</td>
<td>402</td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 28: Adding the measurement with the fifth highest accuracy rate*

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nose Width</td>
<td>Child</td>
<td>818</td>
<td>523</td>
</tr>
<tr>
<td>Eyes to Nose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Mouth</td>
<td>Adult</td>
<td>966</td>
<td>375</td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 29: Adding the measurement with the sixth highest accuracy rate*
Using all the measurements resulted in the best overall accuracy and a very good division of the classes. The outcome of this test shows a marked improvement. However for completeness sake, and to make sure there wasn't some combination that would provide even better results, I ran the next trial as before by starting with the two measurements that had the highest accuracy rate for each class.

Table 32: Measurements with highest accuracy rate for each class
Tables 33 - 35 show a better progression of the separation between classes and the overall accuracy rate. Adding the remaining two measurements would result in the same statistics as Table 32, so at this point I switched to the combination method of using the measurements with the highest overall accuracy and the measurements with the highest accuracy for each class. Table 36 shows the result of this test.
Having gotten a sense for how the attributes affect the number of positives and negatives, as well as the overall accuracy, I added the eye distance measurement in an attempt to correctly classify more adults. Table 37 indicates that this was successful.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyeDist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Nose</td>
<td>Child</td>
<td>933</td>
<td>408</td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td>Adult</td>
<td>864</td>
<td>477</td>
</tr>
<tr>
<td>Eyes to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 36: Adding the eye distance measurement

Continuing with the same philosophy that was successful with the eye distance measurement, I chose to add the nose width measurement in order to correctly classify more child images. Table 38 shows that this attempt was successful as well.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyeDist</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoseWidth</td>
<td>Child</td>
<td>954</td>
<td>387</td>
</tr>
<tr>
<td>Eyes to Nose</td>
<td>Adult</td>
<td>865</td>
<td>476</td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 37: Adding the nose width measurement resulted in correct classification of more child images

This accuracy and split were better, and both attempts at capturing more images of both classes had been successful, so I decided to try one more measurement. I added the mouth to bottom measurement in order to increase the number of adult images that were classified correctly. In this case, as shown in Table 39, the results
were not as meaningful as I had hoped. In fact, it reduced the number of child images by exactly the same amount as adult images increased, resulting in the same exact overall accuracy.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyeDist</td>
<td>Child</td>
<td>951</td>
<td>390</td>
</tr>
<tr>
<td>NoseWidth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Nose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Mouth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eyes to Bottom</td>
<td>Adult</td>
<td>868</td>
<td>473</td>
</tr>
<tr>
<td>Nose to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouth to Bottom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 38: Adding mouth to bottom measurement

After running all these tests, I felt that I had reached the upper limit of overall accuracy for both the combined attributes and the measurement-only attributes. Tables 25 and 32 represent the best overall accuracy for each attribute set. For the combined set of attributes this accuracy was 68.49% and for the measurement-only set it was 68.98%.

The final set of tests I performed in my attempt to find the optimal set of attributes was to divide the result sets into training and test sets. I used a simulated bootstrapping technique by performing these tests several times with a different division for each attribute set. Tables 40 and 41 show the results of these tests. The combined attribute set in Table 41 presented the best overall accuracy, the most consistent results, and the best division among the two classes.
<table>
<thead>
<tr>
<th>Measurement Only Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>482</td>
<td>189</td>
<td>69.15%</td>
</tr>
<tr>
<td>Adult</td>
<td>446</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>Trial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>484</td>
<td>186</td>
<td>68.36%</td>
</tr>
<tr>
<td>Adult</td>
<td>432</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>Trial 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>468</td>
<td>203</td>
<td>68.26%</td>
</tr>
<tr>
<td>Adult</td>
<td>448</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Trial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>498</td>
<td>173</td>
<td>67.36%</td>
</tr>
<tr>
<td>Adult</td>
<td>406</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>Trial 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>499</td>
<td>171</td>
<td>69.47%</td>
</tr>
<tr>
<td>Adult</td>
<td>432</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>Trial 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>471</td>
<td>200</td>
<td>69.89%</td>
</tr>
<tr>
<td>Adult</td>
<td>467</td>
<td>204</td>
<td></td>
</tr>
</tbody>
</table>

*Table 39: Measurement-only attribute set with training and test split*

<table>
<thead>
<tr>
<th>Combination Set of Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>504</td>
<td>167</td>
<td>69.60%</td>
</tr>
<tr>
<td>Adult</td>
<td>430</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>Trial 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>482</td>
<td>189</td>
<td>67.96%</td>
</tr>
<tr>
<td>Adult</td>
<td>430</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>Trial 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>486</td>
<td>184</td>
<td>68.43%</td>
</tr>
<tr>
<td>Adult</td>
<td>431</td>
<td>239</td>
<td></td>
</tr>
<tr>
<td>Trial 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>464</td>
<td>207</td>
<td>68.70%</td>
</tr>
<tr>
<td>Adult</td>
<td>458</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>Trial 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>473</td>
<td>197</td>
<td>68.88%</td>
</tr>
<tr>
<td>Adult</td>
<td>450</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>Trial 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>471</td>
<td>200</td>
<td>70.42%</td>
</tr>
<tr>
<td>Adult</td>
<td>474</td>
<td>197</td>
<td></td>
</tr>
</tbody>
</table>

*Table 40: Combined attribute set with training and test split*
4.15 Optimal Set of Attributes

I chose the combined set of attributes as the optimal set. In the best case, this set of attributes had the highest overall accuracy rate and class division. In general, the combination attribute set resulted in a better separation of the classes, and was slightly more consistent in terms of accuracy rate and class division.

4.15.1 Alternate Sets of Attributes

One thing that became quickly apparent while running all of the previous tests was that the attribute set could have drastic effects on the number of images that were correctly and incorrectly classified. This led me to the realization that I could intentionally manipulate the attribute set to provide either more child images in the result set, or to minimize the number of false positives in the results. To that end, I attempted to find an attribute set that would return the most child images while at the same time keeping the number of false positives at a realistic level. The intent of this experiment was that I could provide the user with three options: an even division of the classes, more child images in the result set, or minimal false positives in the result set.

As mentioned previously, I could return near 90% of child images in the result set, but at the expense of having almost every adult image in that same result set. Obviously this would be unacceptable for law enforcement since the data reduction would be minimal. There were only two realistic possibilities for attribute sets that would accomplish this feat, and they are represented in Tables 42 and 43. The attribute set in Table 42 would be the better option since it has better overall accuracy, maximizes the number of child images, and has far fewer false positives.
Table 41: Attribute set returning a high percentage of child images

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EyesToNoseToMouth</td>
<td>Child</td>
<td>1033</td>
<td>308</td>
</tr>
<tr>
<td>EyesToMouthToBottom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToNoseToBottom</td>
<td>Adult</td>
<td>630</td>
<td>711</td>
</tr>
</tbody>
</table>

Table 42: Another, less desirable, option for returning a high percentage of child images. This option has many more false positives.

Table 44 contains the attribute set that returns a smaller set of positively classified child images, but it also minimizes the number of false positives. In child pornography cases, there only needs to be a handful of images found on a suspect's digital media in order to prosecute. In fact, one image would suffice, but in most cases, ten or more images are desired in order to show intent and criminal activity as opposed to the accidental browsing to an illegal website and unknowingly caching thumbnail images. This set of attributes provides the investigator with the option of radical data reduction while still providing enough results to make a case.

Table 43: Attribute set that minimizes false positives

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoseToBottom</td>
<td>Child</td>
<td>545</td>
<td>796</td>
</tr>
<tr>
<td>EyesToMouthEyesToNose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyesToTopBottom</td>
<td>Adult</td>
<td>1167</td>
<td>174</td>
</tr>
</tbody>
</table>
4.16 Determining the Age Cutoff

Now that I had the optimal attribute set, I next wanted to justify using children of age 12 years and under for the cutoff between children and adults. Practically speaking, digital forensics investigators look for images of younger children, since those images are easier to identify and use in a court of law. Medically speaking, a child's face becomes more adult-like after the age of 12. Programmatically speaking, I wanted to prove that the line between child and adult begins to blur after the age of 12. The following sections describe those tests and present the results.

4.16.1 Image Database Divisions

The following list is the categories I used to divide the database into age groups. These age groups are based on child development stages and represent the ages where more significant changes take place in the face.

- Infant 0 - 12 mos.
- Toddler 1 - 3 yrs.
- Early Child 4 - 6 yrs.
- Late Child 7 - 9 yrs.
- Preteen 10 - 12 yrs.
- Early Teen 13 - 15 yrs.
- Late Teen 16 - 17 yrs.
- Adult 18+ yrs.

4.16.2 Results of DB Division Tests

Tables 45 - 51 show the results of the tests I performed by comparing each child category to the adult category. In each test, I again used the same number of child and adult images. There were several significant findings that came from these
tests. Table 45 shows that the best accuracy is achieved when comparing infants to adults. This supports the work that Kwon and Lobo performed, and also shows that the accuracy rate jumps up to almost 77%. This is back to where it was when the databases were unequal. The next significant finding was that the accuracy does indeed decrease as a child's age increases, which leads to the final point represented in Table 51. In the late teen period of life, a child's face becomes almost completely adult-like leading to almost all adult images being misclassified.

<table>
<thead>
<tr>
<th>Infant to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>93</td>
<td>43</td>
<td>76.47%</td>
</tr>
<tr>
<td>Adult</td>
<td>115</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

*Table 44: Results of comparison between infants and adults*

<table>
<thead>
<tr>
<th>Toddler to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>140</td>
<td>81</td>
<td>71.04%</td>
</tr>
<tr>
<td>Adult</td>
<td>174</td>
<td>47</td>
<td></td>
</tr>
</tbody>
</table>

*Table 45: Results of comparison between toddlers and adults*

<table>
<thead>
<tr>
<th>Early Child to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>76</td>
<td>39</td>
<td>63.04%</td>
</tr>
<tr>
<td>Adult</td>
<td>69</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

*Table 46: Results of comparison between early childhood and adults*

<table>
<thead>
<tr>
<th>Late Child to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>66</td>
<td>16</td>
<td>60.37%</td>
</tr>
<tr>
<td>Adult</td>
<td>33</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>

*Table 47: Results of comparison between late childhood and adults*
Table 48: Results of comparison between preteen and adults

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preteen to Adult</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>82</td>
<td>6</td>
<td>60.80%</td>
</tr>
<tr>
<td>Adult</td>
<td>25</td>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>

Table 49: Results of comparison between early teen and adults

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Teen to Adult</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>36</td>
<td>40</td>
<td>60.53%</td>
</tr>
<tr>
<td>Adult</td>
<td>56</td>
<td>220</td>
<td></td>
</tr>
</tbody>
</table>

Table 50: Results of comparison between late teen and adults

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Teen to Adult</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>52</td>
<td>4</td>
<td>50.89%</td>
</tr>
<tr>
<td>Adult</td>
<td>5</td>
<td>51</td>
<td></td>
</tr>
</tbody>
</table>

4.16.3 Combination Test Sets

This next set of tests I executed with the intention of further proving the point that there is a definite line between distinguishable data sets. I started by combining the infant and toddler images and comparing this combined set to the adult images. In each subsequent test, I added the next age group until finally comparing all children under age 18 to the adult database. Once again, in each test I used the exact same number of child and adult images. The following list shows the groupings I used for the tests.

- Infant/Toddler
- Infant/Toddler/EarlyChild
- Infant/Toddler/EarlyChild/LateChild
- Infant/Toddler/EarlyChild/LateChild/Preteen
- Infant/Toddler/EarlyChild/LateChild/Preteen/EarlyTeen
- All Children under 18

4.16.4 Results of Combination Tests

Tables 52 - 57 present the results of the child combination tests described in the previous section. The most significant result of these tests is found in Tables 55 and 56 where there is a drastic decrease in the accuracy rate when the early teen set of images is added to the combination. This shows that once again ages 12 years and under do indeed provide the best and most logical cutoff for the child category.

<table>
<thead>
<tr>
<th>Infant/Toddler to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>234</td>
<td>123</td>
<td>73.25%</td>
</tr>
<tr>
<td>Adult</td>
<td>289</td>
<td>68</td>
<td></td>
</tr>
</tbody>
</table>

*Table 51: Comparing infants, toddlers to adults*

<table>
<thead>
<tr>
<th>Infant/Toddler/EarlyChild to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>336</td>
<td>136</td>
<td>69.49%</td>
</tr>
<tr>
<td>Adult</td>
<td>320</td>
<td>152</td>
<td></td>
</tr>
</tbody>
</table>

*Table 52: Comparing infants, toddlers, early childhood to adults*

<table>
<thead>
<tr>
<th>Infant/Toddler/EarlyChild/LateChild to Adult</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>292</td>
<td>262</td>
<td>66.61%</td>
</tr>
<tr>
<td>Adult</td>
<td>446</td>
<td>108</td>
<td></td>
</tr>
</tbody>
</table>

*Table 53: Comparing infants, toddlers, early/late childhood to adults*
The last test I performed within the area of age classification was to train the program with a set of images from the 12 and under child images and use this model to test all the images of children against the adult images. The results in Table 58 are predictable in that the overall accuracy rate has decreased as has the accuracy rate for children.
4.17 Gender and Ethnicity

This last set of tables represents data I collected on the number of images of each gender and ethnicity that were being correctly and incorrectly classified. I wanted to verify whether or not gender and ethnicity played any part in the accuracy of the program. Tables 59 - 63 indicate that gender and ethnicity did not have any bearing on the results of classification.

<table>
<thead>
<tr>
<th>GENDER</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>486</td>
<td>174</td>
<td>73.64%</td>
</tr>
<tr>
<td>Female</td>
<td>459</td>
<td>222</td>
<td>67.40%</td>
</tr>
<tr>
<td>Combined</td>
<td>945</td>
<td>396</td>
<td>70.47%</td>
</tr>
</tbody>
</table>

*Table 58: Accuracy of gender*

<table>
<thead>
<tr>
<th>AFRICAN</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>118</td>
<td>32</td>
<td>78.67%</td>
</tr>
<tr>
<td>Female</td>
<td>116</td>
<td>50</td>
<td>69.88%</td>
</tr>
<tr>
<td>Combined</td>
<td>243</td>
<td>82</td>
<td>74.05%</td>
</tr>
</tbody>
</table>

*Table 59: Accuracy of images of Africans*

<table>
<thead>
<tr>
<th>ASIAN</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>145</td>
<td>58</td>
<td>71.43%</td>
</tr>
<tr>
<td>Female</td>
<td>152</td>
<td>65</td>
<td>70.05%</td>
</tr>
<tr>
<td>Combined</td>
<td>297</td>
<td>123</td>
<td>70.71%</td>
</tr>
</tbody>
</table>

*Table 60: Accuracy of images of Asians*
<table>
<thead>
<tr>
<th>CAUCASIAN</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>223</td>
<td>84</td>
<td>72.64%</td>
</tr>
<tr>
<td>Female</td>
<td>191</td>
<td>107</td>
<td>64.09%</td>
</tr>
<tr>
<td>Combined</td>
<td>414</td>
<td>191</td>
<td>68.43%</td>
</tr>
</tbody>
</table>

*Table 61: Accuracy of images of Caucasians*

### 4.18 Independent Database Tests

In order to more fully validate the findings of the preceding tests, I ran a series of tests against the FGNet image database. The fact that this is an age progression database produced some interesting results, which, once analyzed, reinforced many of the findings of the original tests.

The tests against the FGNet database were initially run with the machine learning model I trained with the original image set. The intent of this test was to show that the original model would be able to correctly classify any set of facial images. However, the first round of tests produced significantly lower accuracy rates than previously reported from the original database.

A peer review performed by members of our forensics group uncovered the fact that there seemed to be a high number of images in which the mouth was found in an incorrect position. This was not a trend that was noticeable in the tests with the original database. Regardless, I modified the search area for the mouth and rectified this problem. I reran tests against both the original database and the FGNet database, but there was not a significant change to the accuracy rate of either set.

I also ran tests on the individual attributes – proportions and measurements – and verified that the trends of those attributes were the same. Attributes that I chose for classifying more children or adults were indeed still classifying more children and
adults. The attributes that generated the highest overall accuracy rate were still the highest albeit relatively low in this case. So the attribute set was performing the way it should, and the Haar classifiers were finding the features in the correct position.

Upon further examination of the results of the tests against the FGNet image set, I discovered that the cause of the low accuracy was due to the fact that the image database contained multiple instances of the same person, and the classification of that person was consistent for all instances. Meaning, if a person was misclassified, then most, if not all, instances of that person were misclassified. The same holds true for correct classifications. There are several conclusions that can be drawn from these findings.

In the case of adults being classified the same at various ages, this supports the claim that the proportions of an adult’s face don’t change much with age. A person’s appearance will change, and they will look older, but the underlying facial structure does not change significantly.

The inverse is true for children. The proportions of a child’s face do change as the child ages. When I closely inspected the duplicate results for children, I discovered that a child was being classified the same for images that were close in age, but not necessarily for all age ranges. As an example, a child would be correctly classified at ages 1, 2, and 4, but incorrectly classified at 8, 9, and 10. The implication of this finding is that the groupings I proposed for testing the age cutoff were further validated and verified by these tests.

For the last test on the FGNet database, I removed most of the duplicate images of the same person. For the sake of maintaining a data set that was large
enough to produce meaningful results, I kept some duplication. For both children and adults, I chose images from a larger span of years. For example, I would choose images of a child at ages 1, 5, and 11. For an adult, I would try to choose a larger span such as 19 and 37. I also chose images that would represent each year equally if at all possible. After removing the duplicate images, I reran the tests using the model from the original database. These tests produced an accuracy rate of 72% when run with children 12 and under, and an accuracy rate of 65% using all children 17 and under. These results are very similar to the results of the original database and underscore the validity of the methodology used to perform age classification.

4.19 Performance Analysis

The final series of tests I performed were an attempt to determine if I could reduce the amount of time spend processing each image, and also to determine where the bulk of processing time was occurring. With the current configuration, each image was being processed at an average speed between 800 and 900 milliseconds. This processing time includes face and feature detection, extracting data from the features, calculating the proportions and measurements, and file I/O for the results. I set a series of timers throughout the code to determine how much time was spent at each stage of the image processing.

When I analyzed the results of the timers, I discovered that the majority of processing time was spent on object detection. On average, detecting the face, eyes, nose, and mouth accounted for 90 - 95% of the processing time. In most cases, over half of the total detection time was spent detecting the nose. Detecting the face took
approximately 25% of that total time as well. The remaining 25% was fairly evenly split by the other features - left eye, right eye, and mouth. If there was to be any increase in performance, in terms of speed, it seemed likely that it would come from taking a closer look at the face and nose detection.

Since nose detection was by far the most time consuming function, I made the assumption that it was due to the fact that this was only feature for which I had set the classifier to find all instances of the object. To verify this assumption, I changed the classifier setting to find the largest object. I knew from previous experimentation with the classifiers that this would result in a lower accuracy rate, but I wanted to determine if the reduction in accuracy would be outweighed by the increase in speed. I found that the detection time for the nose was reduced by an average of 400 to 500 milliseconds per image. This reduced processing time to nearly 1/2 of the original time per image. However, the accuracy rate did indeed drop quite drastically as well. The total accuracy for adults and children ranged between 64% and 66%. The accuracy for children 12 years of age and under dropped to 52%. There was also a significant reduction in the number of child images found with a full feature set. So in the case of the nose detection, the trade-off between speed and accuracy seemed to be in favor of accuracy.

For face detection, I couldn’t modify the settings of the face classifier since I want that classifier to find every face in an image. I tried a different classifier to see if that would affect the processing time, and it in fact had very little impact. In the end, even though the stated goal was speed over accuracy, there were in fact instances where an increase in speed had an undesirable effect on the program as a whole.
4.20 Summary of the Results

This section is a high level review of all the previous sections of tests and findings. The intent is to place all the results in a brief section for the purposes of quick reference. The following subsections provide a summary of the attributes used in machine learning to separate the classes, the accuracy of the program, and the speed.

4.20.1 Attribute Set

During the attribute tests, I realized that different sets of attributes could be used to provide very differing results. This discovery led to my creation of three sets of attributes, each serving a different purpose. The attribute sets and the corresponding reason of that particular set are summarized in the next two sections.

4.20.1.1 Optimal Attribute Set

The optimal attribute set is the set of attributes that divides the two classes evenly. It maximizes the true positives and negatives, while at the same time minimizing both the false negatives and false positives. There were two choices for this optimal attribute set - a measurement only set, and a combined measurement and proportion set. I chose the combined attribute set since it provided the best overall accuracy when split into training and test sets, and also because it returned the best division of the two classes. It also was slightly more consistent in both areas when tested multiple times with different result set divisions. The optimal attribute set is listed here:
- EyesToMouthEyesToNose
- EyesToMouthNoseToMouth
- EyeDist
- EyesToNose
- NoseToBottom
- EyesToBottom

### 4.20.1.2 Alternate Attribute Sets

There are two alternate attribute sets. Both of these sets were intended to give
the user the option to customize the results. With the first set of attributes, the user will
get more child images in the results, with the side effect of also getting more false
positives (misclassified adult images). The user will have more images of children, but
data reduction is not optimal with this option. The purpose of the second set of images
is to greatly reduce the result set by providing the user with the fewest number of false
positives. The disadvantage is that there are far fewer child images in the results, but
theoretically, enough images for an investigator to make a case if need be. The two
alternate attribute sets are listed here:

**Maximizing the number of child images**
- EyesToNoseToMouth
- EyesToMouthToBottom
- EyesToNoseToBottom

**Minimizing the number of false positives**
- NoseToBottom
- EyesToMouthEyesToNose
- EyesToTopBottom
4.20.2 Accuracy

The accuracy of the program is dependent upon five factors: the number of faces found, the number of full feature sets found, the correct age classification of the faces, gender, and race. This section provides a summary of the accuracy of each of these components.

The accuracy of finding the number of faces in an image is entirely dependent upon the accuracy of the Haar classifier itself. By testing the various Haar classifiers for the face, I was able to choose the one that had the best tradeoff between true and false positive results. 81% of the images in the database were correctly identified.

There were several things I could control with respect to finding full feature sets. First was choosing the best set of classifiers for each of the facial features - eyes, nose, and mouth. The second was to limit the search of the facial features to the face region itself. Third, by dividing the face into regions where the features were likely to be found, the classifiers were more accurate. By limiting the search area, the classifiers had a better chance of finding the features in the correct position. And lastly, by resizing the image, I was able to find more features since in many cases, the classifiers had a larger area to search. A full feature set was found in 71% of the images.

The classification of the images was another area where I had more control than with the face classifier. By defining more attributes, both proportions and measurements, and by testing various combinations of these attributes, I was able to find the set that resulted in the best overall accuracy. 70% of the images of children
age 12 years and under were correctly classified. 65% of the images of children under the age of 18 were correctly classified.

In terms of race and gender, I calculated statistics based on the number of correctly and incorrectly classified images of each category: male, female, African, Asian, and Caucasian. I found that race and gender did not play a significant role in the age classification.

4.20.3 Speed

The speed of the program was enhanced in several ways. As shown by Viola and Jones in [30], using the cascading classifier method greatly improves the speed at which objects are detected in an image. Using cascading classifiers was one reason I chose OpenCV for the computer vision libraries and also why I used the face detection sample for the basis of my program. Also, in [31] Wilson et al. show that limiting the search for facial features to the face region itself is faster than scanning the entire image for those features. This study also showed that dividing the face into regions also improves the speed. I employed these suggestions as well as my own improvements to the facial division to increase the speed of finding the facial features.

I found that by loading the classifiers into memory I was able to greatly impact the speed of the program. The first few times I ran the program, I was loading the classifiers for each image. This was extremely slow and did not provide the performance that I desired.

Lastly, the machine learning portion of the program - the decision making process and the classification - did not have any effect on the speed of the program. LDA turned out to be very fast, even when classifying thousands of images.
To calculate speed, I set a timer when the processing of an image started and stopped the timer when the processing ended. The average time taken to process a single image was between 800 and 900 milliseconds.
CHAPTER 5

CONCLUSION

5.1 Conclusions from Test Results

In the beginning of this endeavor, I set out to prove that I could determine if a face found in a digital image was that of a child or an adult by using the proportions of the main facial features. I've shown that with a moderate degree of accuracy this is possible. I've also shown that this entire process can be automated.

In the course of proving this thesis, I discovered the best set of attributes - proportions and measurements - that can be used to make this determination. I also showed that the age cutoff for distinguishing a child is 12 years of age and under. Images of children over the age of 12 are more likely to be classified as adults. Finally, I was able to show that race and gender did not have a role in the classification or misclassification of the images.

5.2 Limitations

The only real limitation of the program and the project as a whole is the inaccuracy inherent in the Haar classifiers themselves. It comes down to the program being only as good as the classifiers. If the classifier can't find a face or a facial feature, this impacts the overall accuracy of the program. The Haar classifiers are not 100% accurate, as has been shown throughout this project. However, the Haar classifiers are fast.
Saying that the program is 70% accurate in classifying children and adults is based on images where first a face was found, and second, where a full feature set has been extracted. Considering the fact that only about 80% of the faces are being identified, and a full feature set can be found in only approximately 70% of those images, that doesn't speak well for the program as a whole.

It could also be said that being able to correctly classify 70% of the images is a weakness. This would be a valid assumption, however, given the variances in faces of all ages, there is much more crossover in the attributes and proportions than might be expected. This is especially true with older children. With younger children, these variances aren't quite so extreme and the differences in proportions are more recognizable. This point was shown in the section of this paper describing the tests for the age cutoff.

Finally, there is the issue of this work being limited to images in which the face is clearly visible. The Haar classifier for faces will only detect a face if all of the major features of the face are present. In some instances, in a child pornography case, the faces of the children are obscured by some means. Typically, this could be done by placing a black bar over the face, blurring the face, or some other means of obfuscation. This is a problem in child exploitation cases in general and therefore would also be a limitation of any application using facial detection or recognition techniques.

In the end, even given these limitations, the results of this project are consistent with other published results. In many cases, this application performs better than what
has been previously released. This is explained in more detail in the Contributions section.

5.3 Future Work

For future projects attempting to classify children and adults according to law enforcement needs, the problem of face obfuscation would need to be addressed by exploring other techniques of determining a difference between children and adults. This might include some type of limb detection in which the proportions of limbs to torso are used in the attribute set and in making a determination of age.

Also, this project provided results of gender and ethnicity in terms of children only. A future study could be performed with a dataset of adults that is tagged with the same gender and ethnicity that was used here for children.

5.4 Contributions

Despite the aforementioned limitations, this project accomplishes quite a few new and innovative things as well as improving other areas of previously released research projects.

One of the most significant achievements is the discovery of the fact that the results can be tailored to the user's needs. Depending upon the attribute set used to classify the images, the program can return either a nicely distributed result set of children and adults with minimal false positives and negatives. Alternately, it can return a larger set of child images, but with more adult images. Finally, the program can return a smaller set of child images, but with very few adult images in the results.
The user can determine the degree of data reduction and the percentage of false negatives.

When comparing this program to the one Kwon and Lobo described in [13] and the other studies in the Related Work section, there are several key improvements I made that are worth noting. First, this program is fully automated. Kwon and Lobo relied on a user to pinpoint the middle of the face. Several other projects following Kwon and Lobo relied on manually cropping the images. This degree of human interaction is unacceptable for law enforcement since it would require that the user looks at every single image anyway. Second, by using Haar cascades along with the other enhancements I made, my program runs much faster. Kwon and Lobo relied on several passes of edge detection to determine the exact location of the facial features. Other studies used multiple passes of their algorithms to build the models used to classify faces.

The next point of interest is that the other studies relied on very precise locations of the attributes needed to build their models. The calculations, patterns, and models which attempted to precisely define the face did not provide a significant difference in the classification of their subjects. This is evidenced by comparing the accuracy of Kwon and Lobo’s attributes with the accuracy of the ones I used. Again, I was able to achieve these similar results with a much faster processing time.

In [29] Tanner shows that the results generated by the classifiers are comparable to those generated with precisely plotted points. An overall accuracy rate of approximately 75% was generated for children 12 and under using the plotted points in the FGNet database. Using the Haar classifiers, machine learning model, and
attribute set detailed in this project, and after having removed the duplicate subject entries in the FGNet image set, I was able to produce an accuracy of 72% with children 12 and under.

As well as relying on the precision of finding the facial features, most of the previous work was constrained by the resolution of the images. In order to find the wrinkles and other features that were used to make their models and calculate their classification, they needed images with a relatively high resolution. By using Haar classifiers, I was not as constrained by the resolution of the images.

Lastly, Kwon and Lobo limited their age cutoff to infants, adults, and seniors. Other groups used less than desirable age categories and limited their comparisons to images within those categories. While I showed that infants and adults did indeed provide the best classification accuracy, I also showed that I could classify children up to 12 years of age, and that this was in fact the optimal cut off age.

With respect to the claims laid out in [8, 17] by Memon, it is difficult to draw any conclusions due to the lack of specific information provided. There was no detailed description of the method employed to achieve the 70% accuracy that was stated in the presentation. Likewise, there were not any details of the database on which the 70% accuracy is based. Even so, I was able to show that using Haar classifiers and facial proportions, 70% accuracy is achievable. Further, my research and tests also show that an accuracy of 78% is realistic in cases where the division is more infants to adults.

Up to this point, all of the work done in facial age estimation and classification had explored the use of templates or patterns. Most of this work was founded on
previous solutions and simply attempted to make incremental improvements to either technique or accuracy. Therefore, a very significant contribution of this project is that by performing object detection through the use of Haar classifiers, I was able to move away from all of the previous work in the field and explore a new direction in facial age classification.

One of the main goals of this project was to provide a program that is fast. Accuracy was secondary to speed. Viola-Jones showed that using the cascading classifier greatly increases speed. Likewise, Wilson demonstrated that facial division increased speed and accuracy as well. Combining these with the other improvements I made myself provide a program that can analyze and classify thousands of images in under an hour.

Finally, I was able to meet or exceed law enforcement criteria which drove the goals of this project. The main stated requirement was that the program is fast with accuracy being the secondary concern. As stated above, the program I created is fast. Not only that, but it is sufficiently accurate to provide investigators with a meaningful result set. As an added bonus, the user can customize the result set to provide either greater data reduction and fewer false positives, or less data reduction with more true positives. This customized user experience was not a requirement, but is definitely an added strength of the application. And last but not least, the program is automated. By having an automated program, a forensics agent could launch the program at night at the end of a shift and come in the next morning to find the results waiting. No longer will agents have to manually sift through thousands of images to find the ones that are of the greatest interest.


[24] Safe Eyes – Parental Control Software from Internet Safety -


[25] Snitch from Hyperdyne Software -


[27] SurfRecon Pornography Detection Tools -


University of Rhode Island, May 2011.


