ABSTRACT

Age Classification from Facial Images for Detecting Retinoblastoma

Tak Chien Chiam, M.S.

Mentor: Greg Hamerly, Ph.D.

Facial age estimation from images is a difficult problem, both because it is naturally difficult to tell the exact age of a person visually, and because of the variations in images, such as illumination, pose, and expression. We want to classify people into two groups, children (age ≤ 5) and adults (age > 5), to facilitate the detection of retinoblastoma, a type of pediatric cancer. Current regression based methods are ineffective, as they usually have mean absolute error of 5 years, which is too high for our purposes. We study the facial anthropometric measurements of humans at different ages, and build a system based on these growth patterns. We detect 76 facial landmarks using Active Shape Models, analyze all possible ratios computable from these landmarks, and use the best ratios as input into a Support Vector Machine. Our final system does very well on our problem, correctly classifying 85% of images.

Age Classification from Facial Images for Detecting Retinoblastoma

by

Tak Chien Chiam, B.S.

A Thesis
Approved by the Department of Computer Science

_____
Gregory D. Speegle, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Master of Science in Computer Science

Approved by the Thesis Committee

_____
Gregory J. Hamerly, Ph.D., Chairperson

_____
Gregory D. Speegle, Ph.D.

_____
James B. Farison, Ph.D.

Accepted by the Graduate School
August 2012

_____
J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Greg Hamerly, for his guidance, patience, and knowledge. I am very grateful for having you as my advisor and mentor for the past 3 years, including during my undergraduate research days. I express my gratitude to Dr. Greg Speegle for his guidance during my time at Baylor, and for being a member of my defense committee. I would like to thank Dr. James Farison for being willing to be a member of my thesis defense committee. I would also like to thank all the professors with whom I have taken classes, and also the professors with whom I have had the opportunity to work for as either a teaching or research assistant. To my friends and family, thank you all for your support and help all these years. Finally, thanks to Baylor University for providing me education, work, and research experience.

CHAPTER ONE

Introduction

Leukocoria is an abnormal condition where the pupil appears to be white. This is most commonly seen in photographs taken using a flash, where the pupil appears to be white rather than the normal red-eye effect, but leukocoria can also be observed in photographs taken without a flash. Leukocoria is a presenting sign for a number of medical conditions [7]. Leukocoria is the most common sign of retinoblastoma, an eye cancer that usually develops in early childhood, and can be seen in 60% of patients with retinoblastoma [7]. In this case, the white reflection is caused by the light from the flash reflecting off the white surface of the tumor at the back of the eye.

Retinoblastoma is a cancer that develops in the retina, and can develop in one or both eyes. The cancer most commonly affects children aged 5 and younger, and is commonly diagnosed in children between 1 and 2 years old [22]. Retinoblastoma is easily curable if it is detected early enough such that the cancer has not spread beyond the eye [22]. The main goal of our overarching project, of which this thesis is just a part, is to build a detector that can be used to provide an early warning to parents of children who exhibit leukocoria, such that they can be tested and subsequently, if necessary, treated for retinoblastoma.

In this project, it is important to be able to distinguish between images of children and adults, as leukocoria is a meaningful indicator of retinoblastoma only in young children. A white eye reflection may be an indicator of cataracts in adults. The ability to distinguish between children and adults will allow us to filter images such that only images of young children have to be analyzed for leukocoric eyes, and conversely, given an image with leukocoric eyes, be able to determine if it is significant (the eye belongs to a child) or not.

Therefore the goal of my portion of the project is to distinguish between images of children and images of adults. Specifically, given an image of a face, determine if the person shown in the image is a young child (5 years old or younger) or not (anyone older than 5 years old). This is a specification of the age estimation problem, which is a more general problem of determining the age of the person shown in an image of their face.

While there has been much research done on age estimation, to date most methods approach it as a regression problem, and have a mean absolute error of 5 years (correct to within 5 years). This is high for our purposes, as ± 5 years is likely to cause misclassification for children aged 10 and under. Therefore there is a need to develop a system that is more finely tuned for this age range to be able to make better predictions. Furthermore, many current methods transform the data into more abstract representations, which makes it difficult to interpret exactly what features of the face defines the age of a person. This study focuses on analyzing facial features and correctly modeling growth patterns as they are intuitively understood.

We will study anthropometric measurements to obtain a solid understanding of human growth patterns, which will give us the reasoning to choose ratios between the distances of facial features as our representation. We will explain why ratios are sufficient for our problem, and choose the best ratios in terms of correlation to age and independence from other ratios. Finally, we will optimize a learning algorithm to give us the best results. We will be using the FG-Net database [4], a database containing 1002 facial images from various subjects at various ages (more details in Section 5.2), as our primary dataset. Our final system does very well on our defined problem, correctly classifying images 85% of the time.

CHAPTER TWO

System Overview

As mentioned in the introduction, the goal of the project is to build a system which takes photographs as input, and identify if there are any children who exhibit leukocoria within the photograph. As this is a complicated task, the system consists of many components which form a pipeline.

In general, the system is as follows:

1. Face Detector
2. Age Classifier
3. Eye Detector
4. Leukocoria Detector

The face detector will identify the locations of any human faces in the photographs, and produce a set of facial images. The age classifier will be used to filter the facial images into the child and adult categories. The next step will be to detect eyes within the facial images of children, then to use the leukocoria detector to identify instances of leukocoria within these eyes. My contribution to the project is the age classifier. An example of the system on a sample image is shown in Figure 2.1.

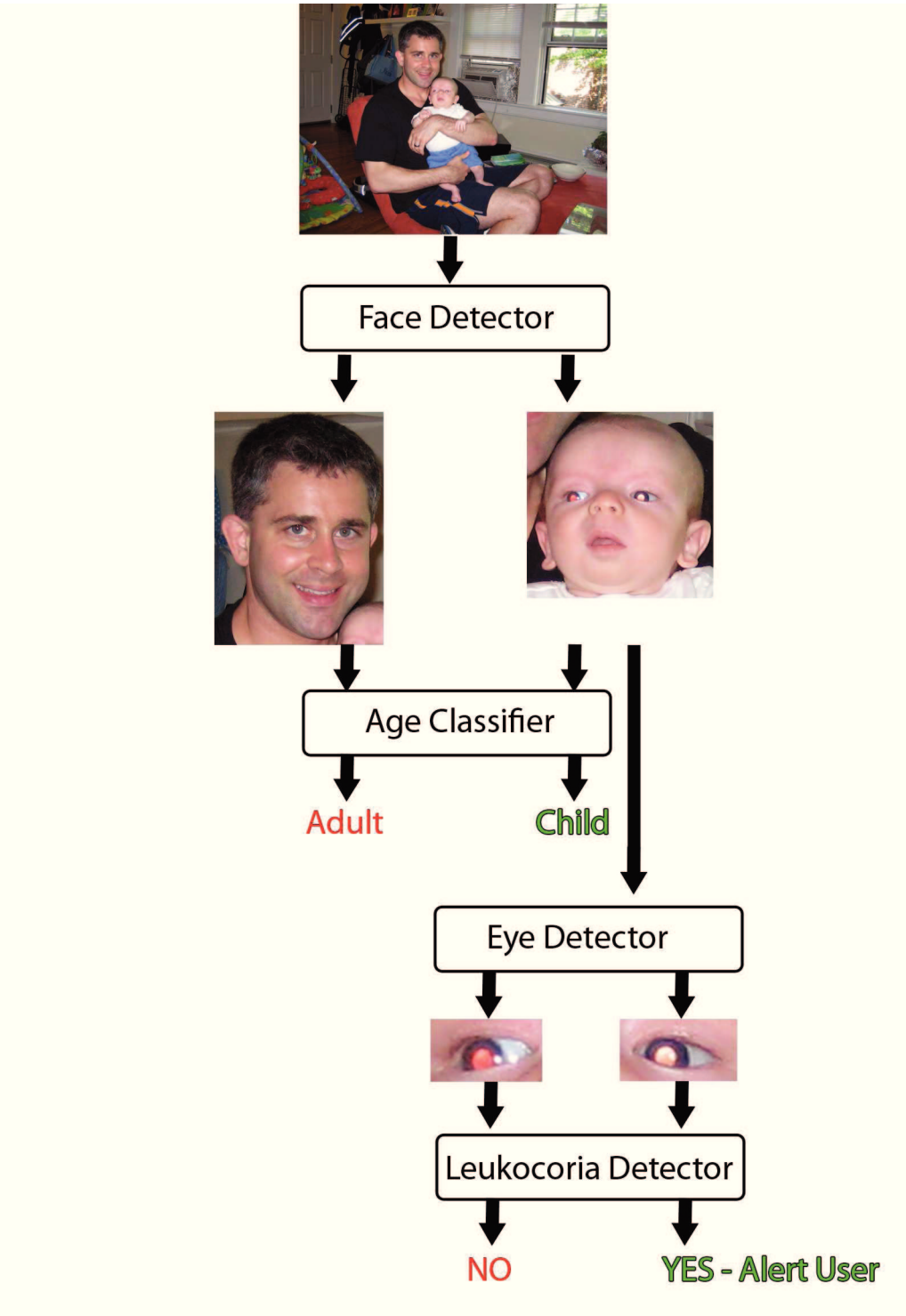


Figure 2.1: System process on example image. First, faces in the image are detected, then filtered to either child or adult labels. The faces labeled child are processed by the eye detector, and the eye images produced are given to a leukocoria detector. If leukocoria is detected, then the user (or family of the child) is alerted. Original image obtained from Dr. Greg Hamerly of Baylor University.

CHAPTER THREE

Problem Background

Age estimation, or more precisely, facial age estimation, is the task of determining a person's age based on a given image of the subject's face. This problem can be approached as either a regression problem or a classification problem. In a regression approach, we can build an age function that will compute a real value for the age of a person, based on some features extracted from the facial image. As mentioned in the introduction, regression is not suitable for our purposes as most methods have mean absolute error of 5 years, which is likely to cause misclassifications for young children. We know that the range of human age is typically from 0 to 130. Therefore, by treating each age as a class, age estimation can also be approached as a classification problem with 131 possible classes. Since the precision of age estimation algorithms is subject to a large margin of error, it is also common to simply classify the subjects into broader categories covering a larger range of ages, such as child, adult, and elderly.

The input image for age estimation problems should contain a single complete human face. In our case, since the age classifier is designed to work with the output of a face detector, this domain restriction is compatible with the main project. The image should be a frontal view of the face, as opposed to a profile view, but we do not make the strict assumption that the subject must be facing or looking at the camera directly, which is to say that there can be variance in pose. We also allow for variance in illumination and expression. This means our input domain includes casually taken images, rather than only passport style photographs, where the subject is assumed to have a neutral expression, face the camera directly, and have good illumination.

3.1 Anthropometric Study

The growth of the cranium vault (the space occupied by the brain) and eyes in their contained orbits (eye sockets) follow the rapid pattern of neural growth, while the rest of the facial complex is primarily related to the development of the dentition (teeth) and muscles of mastication (jaw muscles) [25]. In other words, the cranium of a child grows rapidly during the first 3 years and then slows down significantly, whereas the rest of the skull (the mandible or lower jaw and the rest of the facial bones) will continue growing. This results in a skull in the fetus, infant, and young child that is very different in proportions from that seen in adolescence and adult life, hence the large head and eyes and relatively small face of infants and young children. At birth, the face is 55-60% of the width, 40-45% of the height and 30-35% of the depth of the adult value [25]. From a frontal view, the cranium growth results in an increase in both width and height of the face, whereas the development of the mandible results in significantly more vertical growth (height-wise).

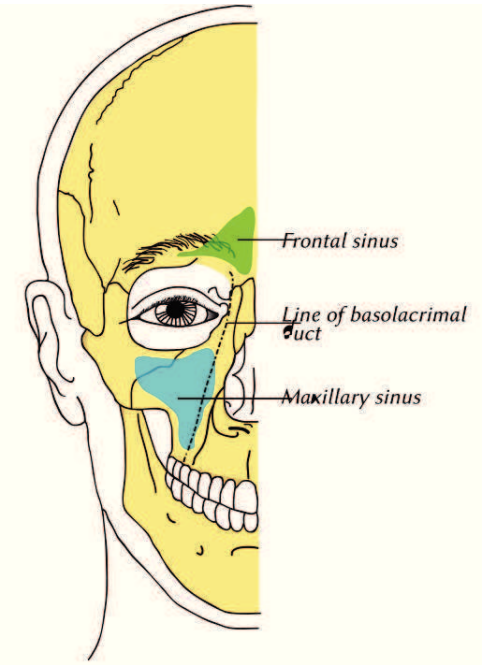


Figure 3.1: Outline of the facial bones from a frontal view. The maxillary sinus is labeled. Image obtained from *Gray's Anatomy* [5].

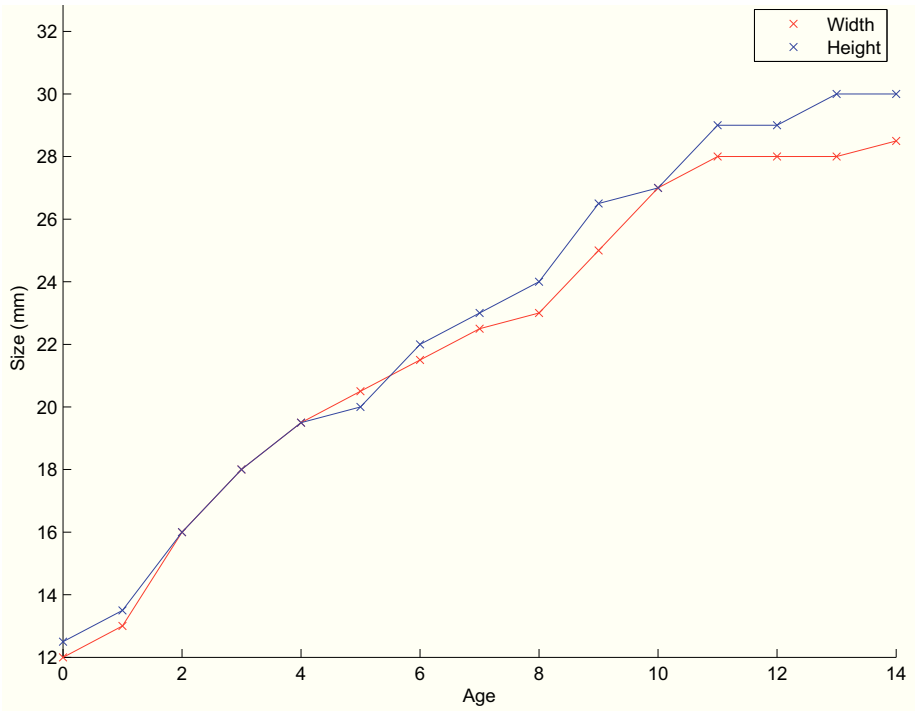


Figure 3.2: Growth of the maxillary sinus from birth to 14 years. Data is obtained from Table 5.19 in *Development Juvenile Osteology* [25]. Notice constant growth rate in both width and height until age 11, which then plateaus.

The maxillary sinus is an air-filled space around the nasal cavity (Figure 3.1), whose height can be roughly interpreted as the distance from eyes to mouth, shows a constant growth rate from birth until the age of 10 to 11 in both width and height. At birth the mean height is 12.5 mm, which increases to 20.0 mm by 5 years, and 29.0 mm at 11 years (pg 135) [25]. This shows significant growth, a 60% increase between the ages of 0 and 5, and another 50% increase between the ages of 5 and 11. Figure 3.2 shows the growth of the maxillary sinus from birth to 14 years. Notice how the growth rate is constant until around the age of 11, where it plateaus.

Likewise, the nasal bones (Figure 3.3), whose height determines the distance from the eyes to the nose, also exhibit significant growth in height from birth until the age of 13. At birth, the mean height is 8.3 mm, which increases to 16.2 mm at 5 years, and 22.8 mm at 13 years [25]. This represents a growth of 95% from ages 0 to 5,

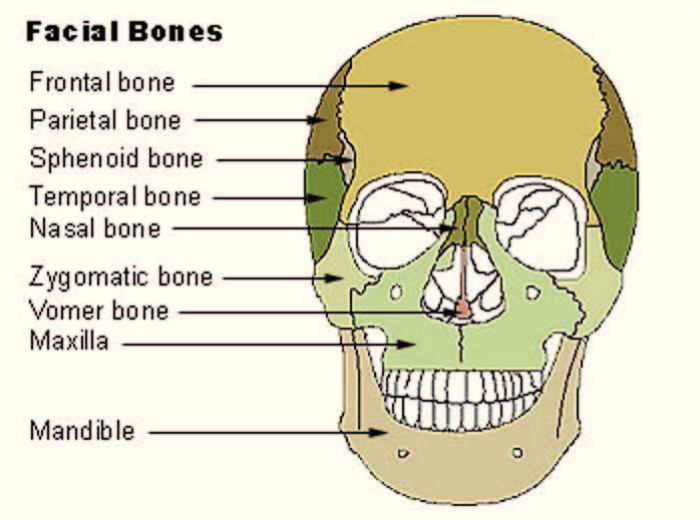


Figure 3.3: Facial bones from a frontal view. The nasal bones are the 2 small bones between the eye sockets as labeled. Image obtained from the website of National Cancer Institute (http://www.cancer.gov/) [1].

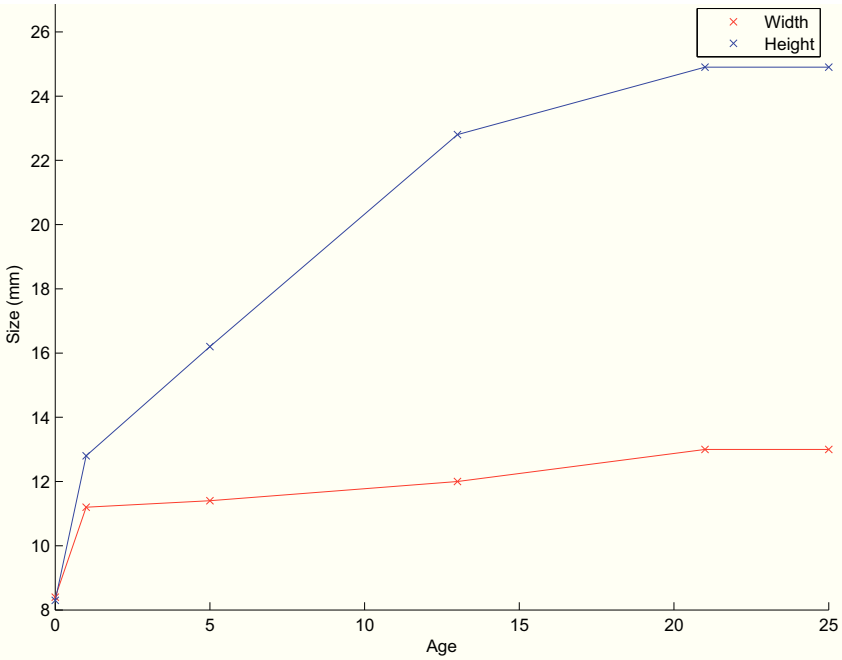


Figure 3.4: Growth of the nasal bones from birth to adult life. Data is obtained from Table 5.13 in Development Juvenile Osteology [25]. Notice significant growth in height until age 13, then significant reduction in growth, and plateau at adulthood. The width has significant growth in the first year, but very little growth past the first year.

and a growth of 40% from ages 5 to 13. Figure 3.4 shows the growth of the nasal bones from birth to adulthood. There is significant growth in height until age 13, then slower growth until adulthood. However, the width grows significantly in the first year, but shows very little growth past the first year.

The last 2 features showed significant and constant growth (at least height-wise) from birth until the age of 13, then slower growth until adulthood. In comparison, the growth of the frontal bone is different. The chord of the frontal bone (Figure 3.5), measured laterally from the top right of the frontal bone to the bottom left, is indicative of the width of the head. The chord of the frontal bone is 73.0 mm at 1 month, 111.5 mm at 4 years, and 117.9 mm at 12 years [25]. This is a growth of 53% between 1 month and 4 years, but only 6% between 4 years and 12 years [25]. Figure 3.6 shows the growth of the chord of the frontal bone from 1 month to adulthood. There is significant growth in the first 2 years, but the growth decreases significantly after 2 years. The difference in growth rates between different parts of the skull clearly indicates that the shape (or at least the aspect ratio) of the skull of a child changes greatly as the child grows.

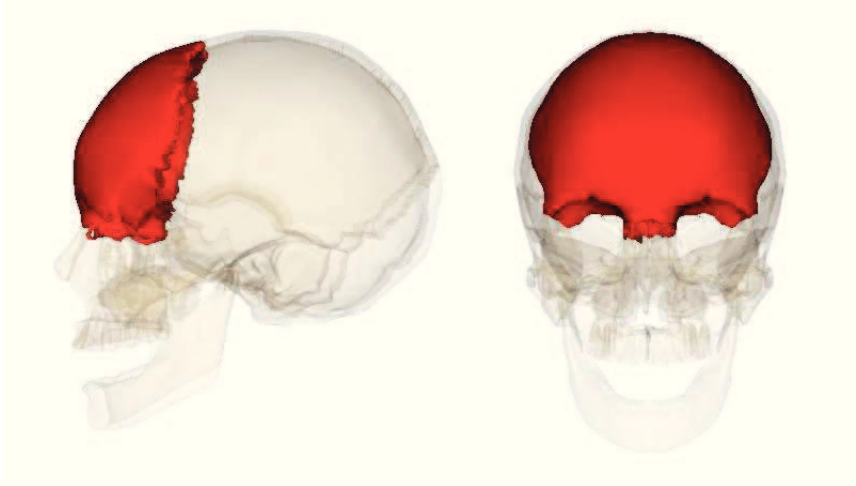


Figure 3.5: A 3-D rendering of the human skull from the profile and front. The frontal bone is colored red. Image obtained from the website of Life Science Databases (LSDB) (http://lifesciencedb.jp/ag/index.jsp) [2].

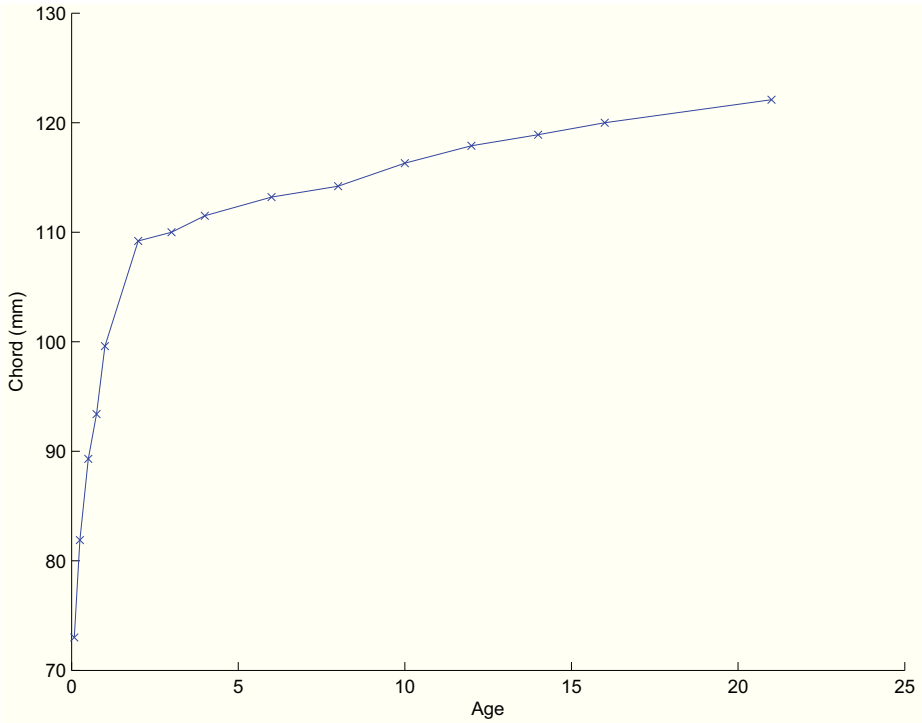


Figure 3.6: Growth of the chord of the frontal bone from 1 month to adult life. Data is obtained from Table 5.10 in Development Juvenile Osteology [25]. Notice significant growth in the first 2 years, but the growth slows down significantly after the age of 2.

Naturally as the skull grows, so does the face. The distance of facial features in a child relative to each other is different than those in an adult because of the facial bone developments. Figure 3.7 shows the measurements made in a study of facial measurements of children [33]. Figure 3.8 plots two of these measurements, the head width and nasion-menton distance (face length), as specified in Figure 3.7, by age. The measurements are also separated by gender after the age of 4. The head width exhibits very fast growth from birth until 2 years (21%), still significant growth between ages 2 and 5 (8%, in total 31% from age 0 to 5), then slows down significantly between ages 5 to 13 (7%). This is consistent with the growth of the frontal bone. The face height exhibits very fast growth until the age of 5, then slows down slightly between the ages of 5 to 17. This is also consistent with the growth of the maxillary sinus and nasal bones. The face height grows by 50% between the ages of 0 and 5,

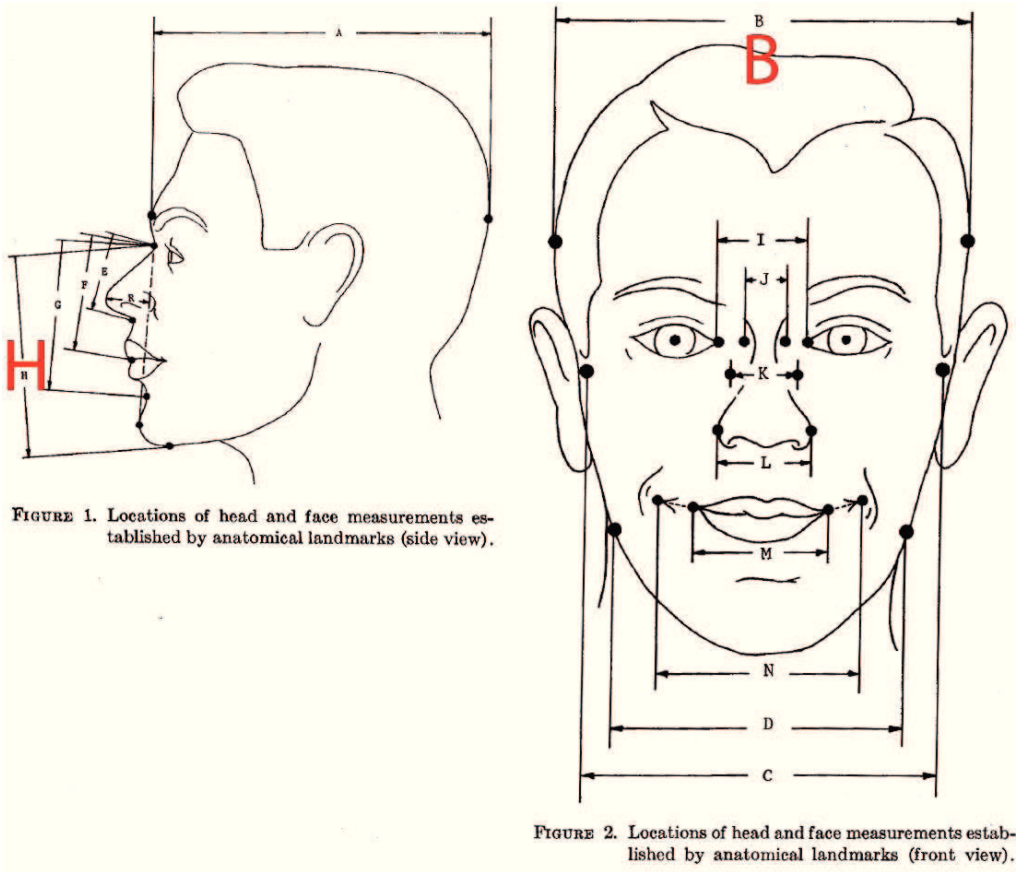


FIGURE 1. Locations of head and face measurements established by anatomical landmarks (side view).

FIGURE 2. Locations of head and face measurements established by anatomical landmarks (front view).

Figure 3.7: Measurements used in a study of facial measurements. Measurements H (nasion-menton distance or face length) and B (head width) are of particular interest to us. Image obtained from *Selected facial measurements of children for oxygen-mask design* [33].

and 26% between the ages of 5 and 13. This demonstrates that the length of the face exhibits more growth than the width of the face from the ages of 5 to 17. Therefore, the face of a person will elongate as they reach adulthood.

Based on the facial features which can be extracted from an image of a face, it is possible to distinguish between children and adults. This is the standard method of distinguishing between young children and adults used in most literature in the field [21, 8, 13]. Because the growth is not constant and features vary from person to person, there is no discrete age where it is possible to distinctly distinguish between young children and not young children. However, we can clearly distinguish between

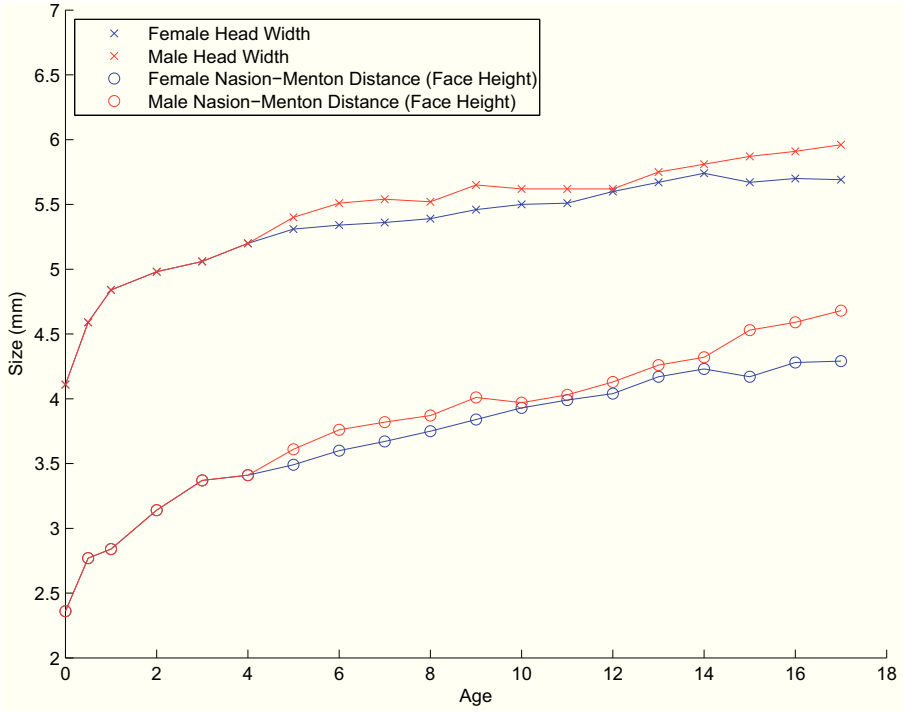


Figure 3.8: Growth of the head width and nasion-menton distance by gender and age. Data is obtained from Tables 2 and 8 in *Selected facial measurements of children for oxygen-mask design* [33]. Notice significant growth in the first 5 years, and slower, constant growth from ages 5 to 17.

a 5 year old and a 13 year old. However, it is unclear if it is possible (indeed we expect it is highly unlikely) to always distinguish between 5 year olds and 6 year olds, as variance in human features makes it very difficult to pin point the exact age.

Figures 3.9 and 3.10 show the growth of a person's face. The images are computed by averaging landmarks taken from faces from the FG-Net database (details in Section 5.2), grouped by age. The faces are then scaled to be of the same width. Notice the significant differences among the lengths of the faces. The growth that we mentioned (distance from the eyes to the nose and the eyes to the mouth) can be easily observed.

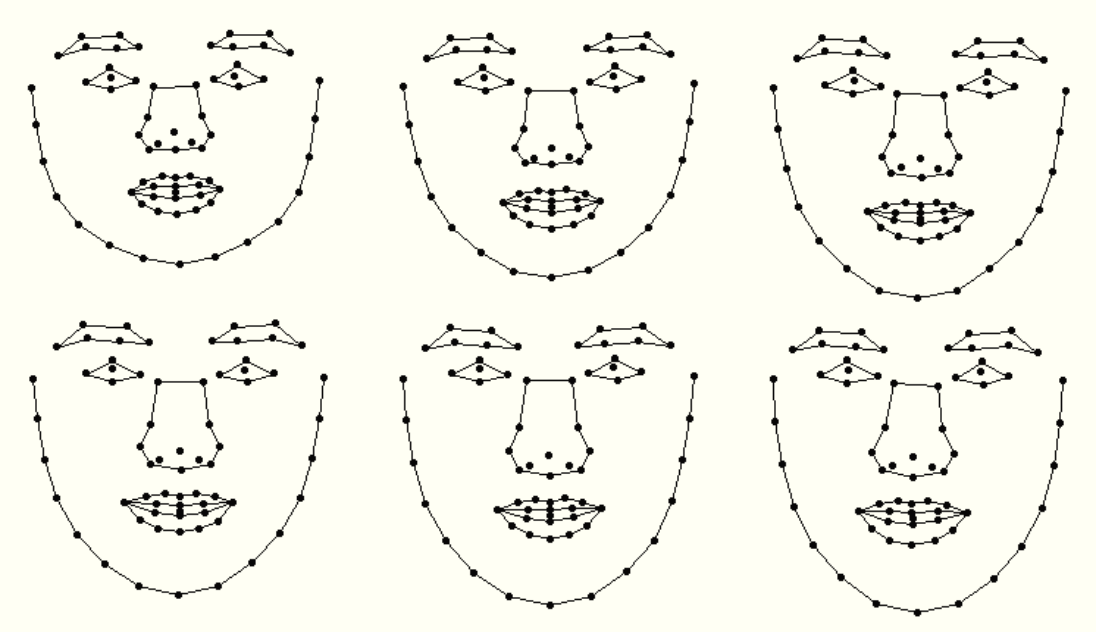


Figure 3.9: Average face by age computed from FG-Net database landmarks. Each face is computed from a sample size of between 15 to 32. From left to right, top to bottom, age 0, 2, 5, 10, 15, and 20. The widths of the faces are the same. Notice how the faces get significantly longer although the width is scaled to remain the same.

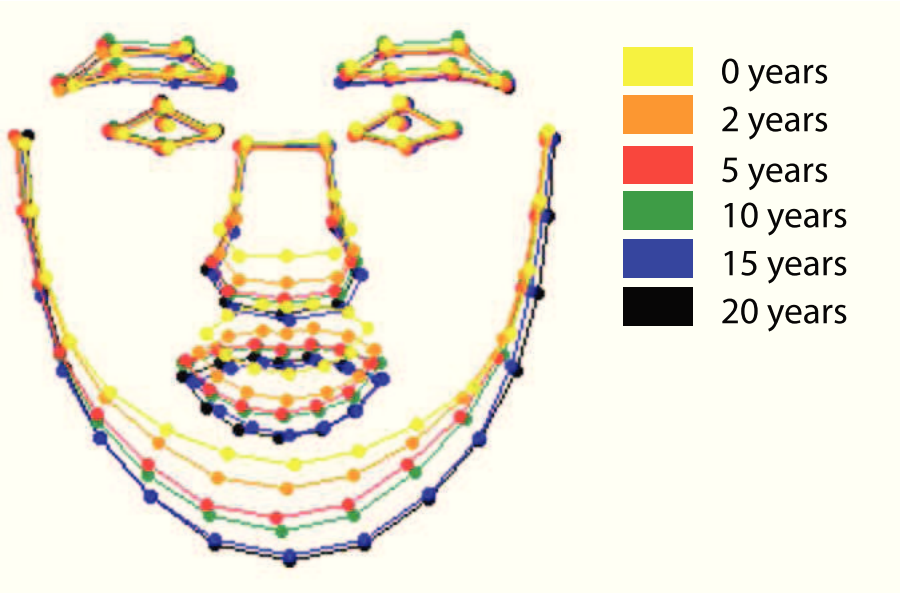


Figure 3.10: The average faces overlaid over each other, centered on the eyes. The smallest (shortest) face is the youngest, and the longer the face, the older the face is. Again, notice how the faces get significantly longer with age.

The vast majority of cases of retinoblastoma occur among young children, with almost two-thirds (63%) of all cases occurring before the age of two years and 95% occurring before the age of five years [22]. Therefore, for the sake of brevity, we will define *child* to be a person of 5 years old or younger, and *adult* for a person older than 5 years for the rest of this thesis, when used in our own context (more accurately, from Chapter Five onwards, as Chapter Four deals with related works that will reuse the words child and adult).

CHAPTER FOUR

Related Work

In general, age estimation is a two step process, the first being representation of the face in a workable format and the second being the classification/regression using the chosen representation. The representation process is necessary as pixel value data from an image is too high dimensional and complex to apply learning algorithms. We will discuss related work in 2 sections, the first being age estimation which is our problem, and the second being facial feature detection as it is a core component of age estimation methods.

4.1 Age Estimation

A detailed paper discussing the development in age estimation is *Age Synthesis and Estimation via Faces: A Survey* by Fu and Huang [13]. This thesis discusses the natural aging process on human faces, age synthesis (rerendering a face with aging effects) techniques, age image representation models, and age estimation techniques. In the age estimation section, the authors discusses a variety of techniques and approaches, some that treat age estimation as a regression problem and others as a classification problem. We will cover some of the mentioned techniques in this section.

Three common representations are anthropometric models, Active Appearance Models, and appearance models. There are other representations such as the Age Manifold. Anthropometry refers to the measurement of the human individual. This can be thought of as physical measurements or observations that humans intuitively understand, such as the distance between the eyes or the amount of wrinkles on a person's face. Active Appearance Model is a statistical face model that encodes the

shape and texture of a face [10]. Active Appearance Model is an extension of Active Shape Model [23], which captures the shape information of an object by fitting a deformable template to the object. AAM extends ASM by extracting the grayscale information within the detected shape. Appearance models are representations that do not attempt to explicitly locate facial features such as the eyes and nose, but apply a transformation process such as Local Binary Patterns in order to extract more meaningful information [32, 17, 26]. Appearance models are usually coupled with either PCA to reduce the number of dimensions of the data, or some boosting method to select portions of the data to work with.

4.1.1 Anthropometric Models

Kwon and Lobo produced the seminal paper in the field of age estimation [21], which used 2 different methods for age estimation. The first method is to compute 6 ratios by dividing the distance of 2 features by the distance of another 2 features (e.g. distance between the eyes over distance between the eyes and the nose). Using these ratios, they are able to classify images to the class labels baby or not-baby. The location of the features are found by template matching. The second method uses snakelets (curves in an image domain that can move due to internal forces within the curve and external forces from the image data [31]) to measure the amount of wrinkles on a person's face. This is used to further classify the not-baby images into adult and senior. Only the first method is of interest to us, as we do not care to distinguish among adults. The template matching method is quite expensive, as it includes a chin finding stage that uses snakelets. Furthermore, it does not account for much variance such as open mouths which will affect some of the ratios. The authors achieved 100% classification accuracy on a very small test set of only 15 faces.

Using the same basic classification structure as Kwon and Lobo, Horng, Lee and Chen classified facial images into 4 age groups: babies, young adults, middle-aged

adults, and old adults, using two methods [8]. The first method is to extract facial features (the locations of the eyes, nose and mouth) using a Sobel filtered version of the image and finding the regions that have the highest intensity (the area around the eyes has a lot of edges due to the complexity of the eyes). Using these features, they define two geometric features, R_{em}, the distance between the eyes and the mouth over the distance between the two eye centers, and R_{enm}, the distance between the eyes and the nose over the distance between the nose and the mouth. Using these ratios, they classify the images to baby or not baby, similar to Kwon and Lobo. The second method is again to measure the amount of wrinkles on a person's face, but this time using the Sobel filtered image instead of snakelets. Their method is much simpler and faster than Kwon and Lobo's; however, they use too few features and ratios, and their feature detection method is too simple and prone to error, and does not handle pose variance well. The authors obtained 99% accuracy using this method, but their test set was only 114 images, and the images used were taken in a controlled environment (passport style).

4.1.2 Active Appearance Models

Lanitis, Taylor, and Cootes utilized a method to model faces called Active Appearance Model (AAM) [10], which is based on Active Shape Models (ASM) [23]. Active Shape Models basically identify key points in a shape, and use these points as a deformable template to look for similar shapes. AAM uses ASM to determine the shape of the face, and uses a grey-level model to extract intensity information from the face found. Using AAM, the authors defined an aging function based on multiple images of a subject at different ages. This aging function can then be used to determine the age. The authors built multiple aging functions as different individuals age in different manners, and use the most appropriate version based on facial appearance

and lifestyle information. The results using this method have mean absolute error of about 5 years on their private database of 500 images through cross-validation.

Geng, Zhou and Smith-Miles explored the idea of using a sequence of an individual's aging face instead of dealing with each aging face image separately [15]. This method is called AGing pattErn Subspace (AGES). It is similar to Lanitis et al.'s method, "but the emphasis of the AGES method is to use face images of the same individual at different ages altogether to represent the aging pattern" [13]. An obvious weakness of this algorithm is that it requires images of the same individual. This method achieved mean absolute error of about 6 years on the FG-Net database through cross-validation.

4.1.3 Appearance Models

Wang, Yau, and Wang used Gabor wavelets and Local Binary Patterns to extract features, and then used Error-Correcting Output Codes with AdaBoost or SVM for classification [28]. The authors grouped the images into four age groups: children (0-11), teenage (12-21), adult (22-60), and senior adult (60+). The authors were able to achieve about 90% accuracy on a combination of the FG-Net and MORPH database [6], a facial image database similar to FG-Net, but does not include images of children.

Similarly, there are many more papers on using Local Binary Patterns and Gabor features [32, 17, 26]. Most of these papers report good results in both regression and classification representations of age estimation.

4.1.4 Other Models

Fu and Huang extracted a face patch (a closely cropped image of just the face) and used manifold learning to obtain a low dimensional representation in manifold subspace [14]. Using this, the authors obtain a common aging pattern which is then

used for determining the subject's age through regression. This method has been shown to be able to achieve mean absolute error of about 5 years on the FG-Net database [13].

4.1.5 Summary

Active Appearance Models, appearance models, and other types of models all produced good results; however, they seem to be too complicated for our simpler version of the age estimation problem. In our problem, there seems to be a clear cut, simple solution as shown by Kwon and Lobo, which is to obtain anthropometric measurements and use them, rather than extracting a low dimensional representation of a face and assuming that it sufficiently encodes the aging variance. We only desire to know if the subject in the image is a child or adult, and we do not need to know the exact age of the subject. Therefore we only need high precision in the ages around 5, where it is not immediately clear if the subject is under or above age 5. For ages outside of the range, it is tolerable to have low precision provided that it is sufficient to correctly classify the subject. By taking a more sound approach, we aim for comparable results using an easier to understand model.

4.2 Facial Feature Detection

We have seen that there are many ways of approaching age estimation, the most natural being the usage of anthropometric measurements through the two-step approach of facial feature detection and classification based on the features (usually ratios of distances between features). Facial feature detection has been thoroughly explored in the field of face detection. A survey paper in the field of face detection that provides a good summary of facial feature detection methods is *Face Detection: A Survey* by Hjelmas and Low [18], which divides the problem into three areas: low-level analysis, feature analysis, and Active Shape Models.

One method of facial feature detection is to approach it in two parts: low-level analysis for detecting potential features, and feature analysis for choosing among the identified features and mapping them to the facial features. The most popular methods for detecting potential features is perhaps using edge operators (such as the Sobel filter) and Haar classifiers [29, 11]. Edge operators are usually filters which are convolved with an image, and are able to detect edges in an image, which usually corresponds to facial features. Haar classifiers calculate Haar-like features, which are the values obtained by placing two (or more) same-sized rectangles side by side over a collection of pixels, and calculating the difference between the sum of the pixel intensities within each rectangle.

Since the detected features from low-level analysis are likely to be ambiguous (it is unknown whether the feature is a nose or a mouth), feature analysis is required. There are two approaches in feature analysis: feature searching and constellation analysis. Feature searching looks for prominent features such as the eyes, then searches for the other features [8, 20]. Constellation analysis looks at the relative position of potential features, and decides if their locations correspond to a face [18].

Another method of facial feature detection is template matching, best demonstrated by the Active Shape Model, as described in Section 4.1.2. More generally, template matching is where we define a face pattern, then search for this pattern in the image.

CHAPTER FIVE

Methodology

In this chapter, we outline the experiments we run, the datasets we use, and our evaluation methods.

5.1 Experimental Design

The experiments are divided into three steps. The first step is to find a facial feature detection method that works well for our purposes. It needs to detect a sufficient number of facial features with high precision. The second step is to then use the detected features and form some sort of higher level representation. We work mainly with ratios between two distances, where each distance is the distance between two facial features. This step requires finding possible ratios and pruning them to obtain the best ratios. The third step is to use this representation as input data into a learning method, optimize, and obtain testing performance.

5.1.1 Facial Feature Detection

Since the plane of separation between our two classes is at age 5, the most informative feature would be simply the relation between the location of the features of the face as mentioned in Chapter 3. Texture information is not important to us, as we do not want to be constrained to high resolution images (texture information is less informative for lower resolution images), in both classes there are subjects with similar texture (e.g. persons of age 4 and 7 both have smooth faces), and texture information is commonly only used for identifying the elderly by their wrinkles [8].

As discussed in Section 4.2, there are many different approaches for facial feature detection. We experiment with two types of methods in general, the first being

low-level analysis with feature analysis, and the second being Active Shape Model implementations. For the two step analysis approach, we implement Horng et al.'s method, and an ad hoc method developed. For the Active Shape Models, we try a few different implementations of ASM, namely asmlib-opencv, an open source ASM library using OpenCV [30], and Stephen Milborrow's Stasm, an extension of ASM [24]. Note that we do not need to use Active Appearance Models as we do not believe that texture information is relevant.

5.1.2 Input Feature Selection

One of the simplest input features that we can use is the distance ratio, which is the distance between two facial features divided by the distance between two other facial features. Distance ratios, or more simply just ratios, have been used to some degree of success in the past [21, 8]. As we investigated in Chapter 3, certain ratios are able to capture the difference in shape of the head between a a child and an adult. We mainly focus on ratios as our input features in this thesis. We expect to find ratios that directly correlate to age, where they either increase or decrease monotonically with age.

Using the maxillary sinus and frontal bone growth rates we identified in Chapter 3, a theoretical ratio is shown in Figure 5.1. After the age of 1, the ratio value decreases significant before staying stable at age 14. The initial increase in ratio value is caused by the rapid growth in the frontal bone for the first year. In general we can see that children will have high values for this ratio, and adults will have low values for this ratio, allowing us to differentiate between the two classes using this ratio.

Another theoretical ratio is the head width over face height based on the facial measurements in Chapter 3. This ratio shows a more consistent decrease in value from birth till adulthood. Again, this shows that we can easily distinguish between children and adults based on ratios such as these.

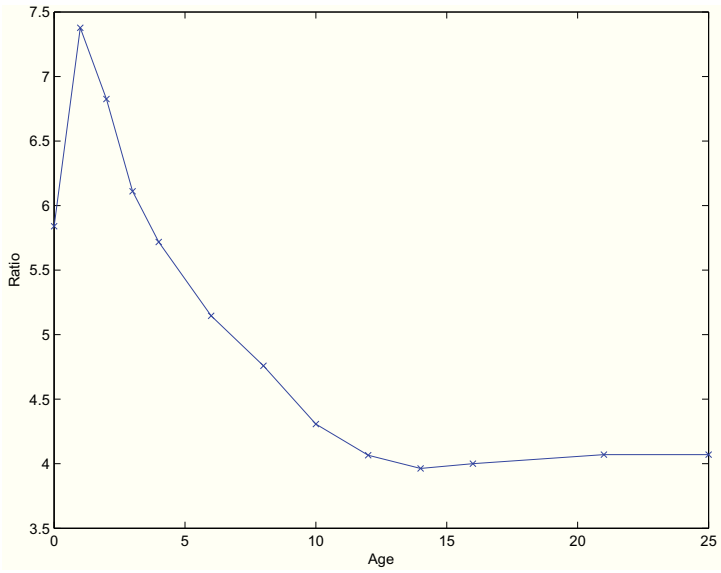


Figure 5.1: Chord length of frontal bone over height of maxillary sinus by age. After the age of 1, the ratio value decreases significantly before staying stable at age 14. The initial increase in ratio value is caused by the rapid growth of the frontal bone for the first year. In general we can see that children will have high values for this ratio, and adults will have low values for this ratio.

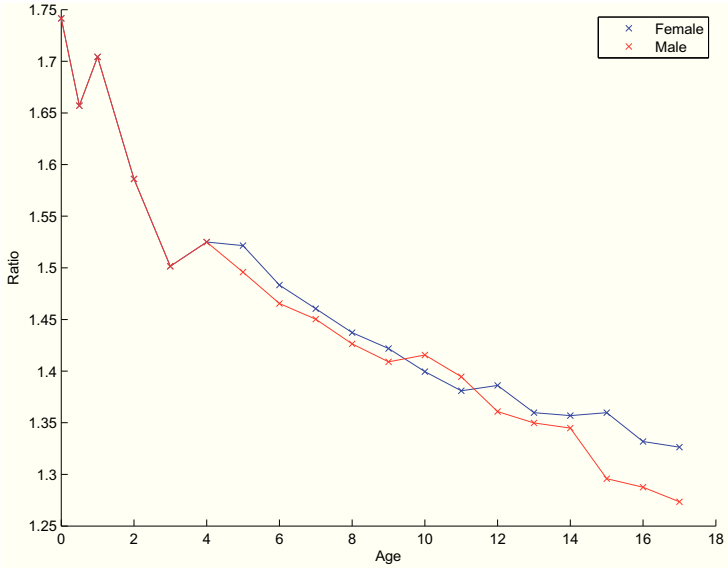


Figure 5.2: Head width over face height by age. This ratio shows consistent decrease in value from birth until age 17. In general we can see that children will have high values for this ratio, and adults will have low values for this ratio.

We divide our input feature selection process into 3 parts: manual selection of ratios, automatic selection of ratios, and complex features. For manual selection of ratios, we select and test ratios that are manually chosen and defined based on the growth patterns of the human skull. For automatic selection of ratios, we select ratios based on their tested performance according to our metric. Finally, for complex features, we test input features other than just distance ratios.

5.1.3 Classification Method

Classification is the task of assigning labels to unlabeled data (test data), or in other words, finding which class an unclassified datapoint belongs to, based on a training set. The training set consists of some number of instances (data), described by their attributes and their labels. For example, our sample input could be data from 1000 images, as described by some set of features obtained from the images such as 100 different distance ratios, along with the labels for the 1000 images (child or adult). A learning algorithm will process this input, and trains a model that will fit this data. This trained model is then presented with unlabeled data (images in which we do not know the true age of the subject), and it will make predictions on it (assign either the label child or adult).

Support Vector Machines (SVM) have been proven to work very well for many problems in computer vision including age estimation [32, 17, 28]. SVMs are one of the best supervised learning algorithms used for both classification and regression. Since we will only use SVMs for classification, we will describe SVMs in a classification context. The input to a SVM is a labeled dataset as described above. The SVM training algorithm will train a SVM model using this data, and this trained model can then be used to predict labels for unlabeled data.

The SVM training algorithm attempts to find a separating hyperplane between 2 classes in the data. For example, if we have data that is 1 dimensional, then we can

plot this data in a line, with circles as children and crosses as adults. A separating hyperplane would then be some point on the line, where all the points to the left of the line belong to one class (say all circles/children), and the points to the right of the line belong to the other class (all crosses/adults). This is a simplification of SVMs, and SVMs are capable of handling data that is not linearly separable through soft margins that allows for misclassification (e.g. allows for some crosses to be on the left of the line), and kernels which we will discuss later. Also, there could be many possible hyperplanes that separate the classes, so the SVM chooses the hyperplane with the largest margin, meaning it tries to leave as large a gap as possible between the hyperplane and the datapoints closest to it. These datapoints that are closest to the hyperplane lie on the margin, and are called support vectors.

Another core component of the SVM is the kernel. Simply put, a kernel maps data to a higher dimensional space. If the data is not linearly separable with its current dimension, by mapping the data to a higher dimensional space, the data can become linearly separable. We experiment with different types of kernels, and try to optimize them.

We also experiment with ensemble classifiers to be able to compare the performance of SVMs. Ensemble classifier use multiple classifiers to improve performance. The idea is that even though we may have classifiers that by themselves do not perform well (weak classifiers), if can combine many of these classifiers in an intelligent fashion to make a decision, then we should be able to make better decisions and perform better. An example of a weak classifier is a decision stump, which is a decision tree of height 1, or a linear classifier. There is no real restriction on what a weak classifier can be (provided it makes better predictions than random guessing does), but it is viewed in contrast to a strong classifier, which is a classifier that does well (makes very good predictions, e.g. SVM).

There are many different methods of combining classifiers, but we will only focus on boosting, and in particular, AdaBoost [12]. In AdaBoost we train multiple classifiers iteratively. Between each iteration, we reweight the data, such that the datapoints that were misclassified in the previous iteration gets weighted more. Each iteration trains 1 weak classifier, and this weak classifier is also assigned a weight according to its performance. At the end, we combine these weak classifiers into a single strong learner, by combining the outputs of each weak classifier using the weights of the weak classifier. Basically, we trust a weak classifier that does better more, so we assign a higher weight to its output, whereas for a weak classifier that does not do as well, we will have a smaller weight on its output.

5.2 Data

The primary database on which we run experiments on is the Face and Gesture Recognition Research Network (FG-Net) Aging Database [4]. The FG-Net database contains 1002 scanned photographs of faces from 82 subjects at different ages. The FG-Net database is a good database for our purposes because it contains images of the same subject at different ages, the age of the subject of the image, and landmark locations of the face of the subject (more detail on landmarks in Section 6.1.4). The landmark locations allow us to verify the performance of the feature detector easily. The age associated with each image provides the ground truth for testing the age classifier. Another advantage of this database is that a significant percentage (23%) of the images are of children (age ≤ 5), with 233 children and 769 adults. However, the quality of the images is not optimal, as the images include scanned old photographs that are visibly damaged and photographs with uncontrolled environments. This includes poor lighting conditions and a variety of poses. Figure 5.3 shows a subset of images from the database. For most subjects in the database, there are images of most stages of their life (as a baby, child, teenager, and adult).

Figure 5.3: A subset of images from the FG-Net database. Each row shows a single subject at different ages. The ages by column are very roughly (±3 years) about the ages 0, 5, 10, 15, and 20. The quality of the images vary greatly, along with pose and illumination.

A secondary dataset used for testing is the Vims (short for Video/Image Modeling and Synthesis Lab) Appearance Dataset for facial ANAlysis (VADANA) dataset [27]. The VADANA dataset contains 2298 higher resolution (compared to FG-Net) images of 39 subjects at different ages. The images are taken in uncontrolled environments so they cover a natural range of pose, expression, and illumination variation. The distribution across age is more imbalanced in this dataset, with only 8% of the images being of children (190 children and 2108 adults). Figure 5.4 shows a subset of images from the dataset. Most images in the database are from 2006 to 2010, so the images for most subjects in the dataset only cover a small range of ages.

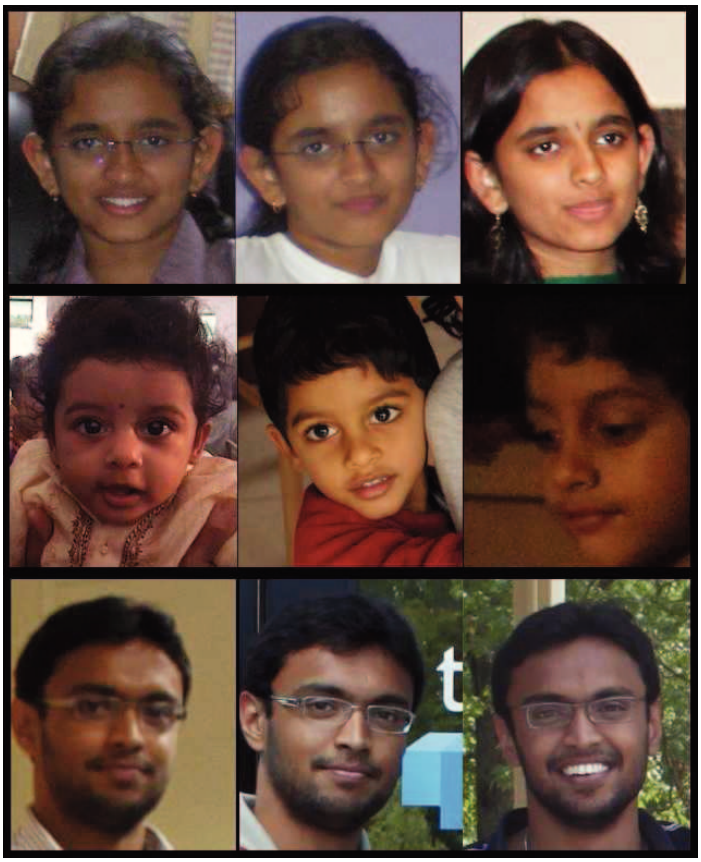


Figure 5.4: A subset of images from the VADANA dataset. Each row shows a single subject. Most images are captured recently (2006 onwards) so the quality is higher than the FG-Net database. Pose, illumination, and expression vary greatly.

Figure 5.5: A subset of images from the HPID dataset. The images in the first row have an upwards tilt of 15°, the middle row has no vertical tilt, and the bottom row has a downwards tilt of 15°. Across the row, the horizontal angle is: -45, -30, -15, 0, +15, +30, and +45. This subset does not contain all the images for this subject; images with vertical angle greater than 15° or horizontal angle greater than 45° are not included.

The Head Pose Image Database (HPID) [16] is an additional database that was used to investigate pose information. The HPID contains 2790 face images of 15 adults with varying poses in terms of vertical and horizontal angles. The vertical angle (tilt) which is the change in angle when a person looks up or down, covers the following angles: -90, -60, -30, -15, 0, +15, +30, +60, and +90. The horizontal angle (pan) which is the change in angle when a person looks left or right, covers the following angles: -90, -75, -60, -45, -30, -15, 0, +15, +30, +45, +60, +75, and +90. This database was used to investigate the effect of the subject's pose on age estimation. A sample of the database is shown in Figure 5.5. The sample consists of images from only one subject, and only a subset of the images available for that subject.

5.3 Evaluation

As mentioned in Chapter 3, our system treats age estimation as a binary classification problem. The labels are *child*, where the age of the subject ≤ 5, and *adult*, where the age of the subject > 5.

To simplify notation, we consider the child class to be the positive class, and the adult class to be the negative class. For a picture of a child, a correct prediction is a true positive, and an incorrect prediction (predicting the child is an adult) is a false negative. For a picture of an adult, a correct prediction is a true negative, and an incorrect prediction (predicting the adult is a child) is a false positive.

Two fundamental metrics that are useful for evaluating results are precision and recall. Precision is asking: "of the items that I predict to be in class X, what percentage actually belongs to class X?" It is a measure of how clean the predicted results are, in the sense that if the precision is high, then we can be more confident that the predicted class labels are correct. Recall would be: "of all the items that belong to class X, what percentage of them did I find?" Recall is a measure of the amount of data retrieved, where if the recall is high, then we can be sure that we found most of the items that belong to the desired class.

Precision:

$$\text{Precision}_{\text{child}} = \frac{TP}{TP + FP}$$

$$\text{Precision}_{\text{adult}} = \frac{TN}{TN + FN}$$

Recall:

$$\text{Recall}_{\text{child}} = \frac{TP}{TP + FN}$$

$$\text{Recall}_{\text{adult}} = \frac{TN}{TN + FP}$$

One of the metrics we use to measure performance of the system is accuracy. Accuracy is defined as the number of correct predictions over the total number of predictions.

Accuracy:

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

As the distribution of the classes is likely to be very uneven (probably fewer images of children than of adults), accuracy might not be the best measure of performance. This is because even if we have a high accuracy, we might be doing poorly on finding true positives (if 90% of the images are of adults, then always predicting adult will give an accuracy of 90%). Therefore another metric we use is the Balanced Success Rate (BSR). It is defined as the mean of the recall on both classes. A high BSR would indicate that we are correctly predicting labels for both classes. It is possible for one classifier to have high accuracy but low BSR because the classifier is predicting the majority class all the time, whereas another classifier could have lower accuracy but much higher BSR because it is doing a better job at finding the minority class.

Balanced Success Rate:

$$BSR = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$$

In our task, it is important for us to correctly identify as many images of children as possible, as we risk missing a case of retinoblastoma otherwise. We are willing to have this increase in recall at the cost of some precision. Therefore, we favor higher recall over precision, as we would prefer to misclassify adults as children, rather than having the classifier miss children because of misclassifying them as adults. As an alternative measure of performance, we could penalize classification errors using

different functions, with a softer function for misclassifying adults as children. Our alternate measure, which we call Age Sensitive Accuracy, is as follows:

$$A(x) = \text{mistake of classifying (5-x)-year-old as adult}$$

$$C(x) = \text{mistake of classifying (5+x)-year-old as child}$$

$$\text{penalty}(A(x)) = 1$$

$$\text{penalty}(C(x)) = 1 - e^{-((x/4)^2)}$$

$$\text{loss}(i) = \begin{cases} 0 & \text{if correctly classified;} \\ \text{penalty}(A(x_i)) & \text{if misclassified and class}(i) = \text{child;} \\ \text{penalty}(C(x_i)) & \text{if misclassified and class}(i) = \text{adult.} \end{cases}$$

$$\text{Age Sensitive Accuracy} = 1 - \frac{1}{n}\sum_{i=1}^{n} \text{loss}(i)$$

Figure 5.6 shows the loss function. When a child is predicted as adult, the loss is always 1, and when an adult is correctly predicted the loss is always 0. When a child is predicted as child, the loss is always 0, but when an adult is predicted as a child, the penalty starts out small but grows quickly. When a 12 year old is predicted as child, the penalty is over 0.95. This function allows us to favor predicting ambiguous faces (e.g. the 5 to 10 year old range) as children to minimize the loss. We will not be training to minimize this error, but this will be an alternative measure of performance of the classifiers.

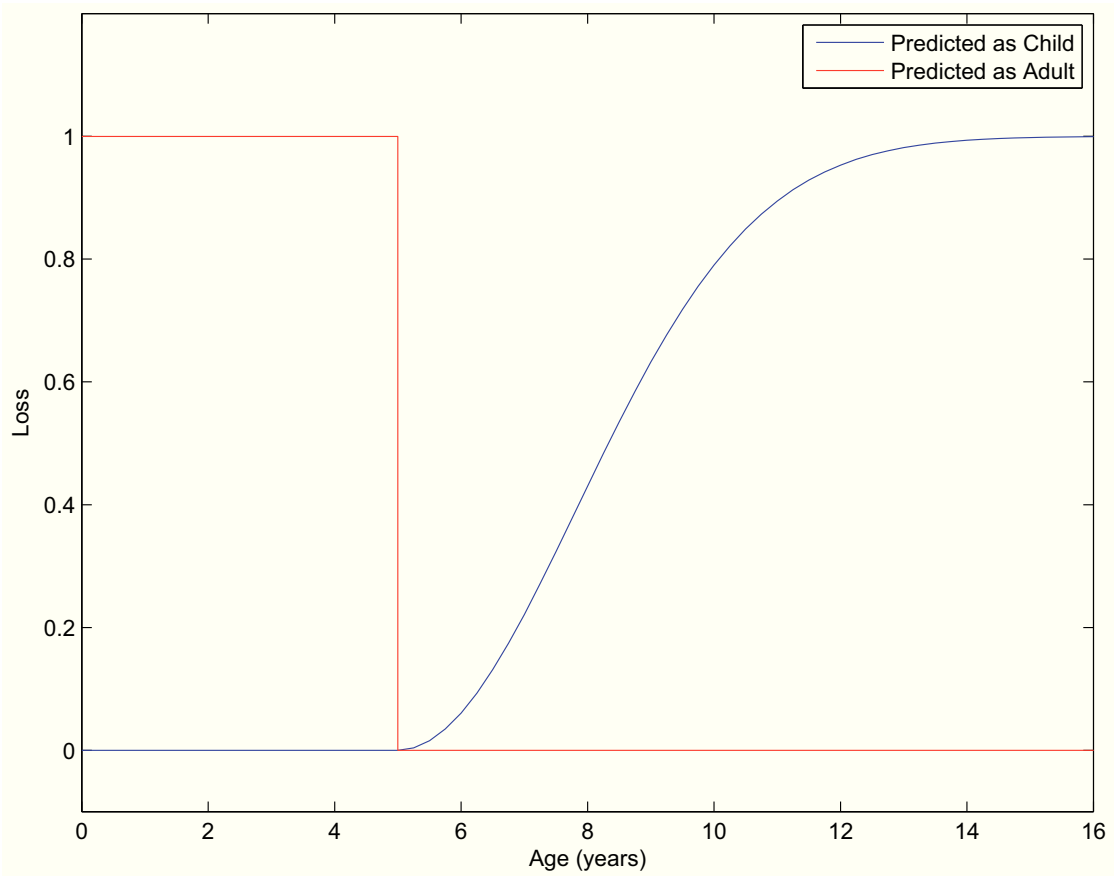


Figure 5.6: Loss function of predicting a person of a given age as either Child or Adult. If the prediction is correct, the loss is 0. Otherwise, if we mispredicted a child as adult, the loss is always 1. If we mispredicted an adult as child, then we have a function that grows with age, in order to penalize these mispredictions less. Note that the penalty is small for misclassification when the adult is just slightly over 5 years old, but quickly approaches 1 (the penalty is over 0.95 when age is 12).

CHAPTER SIX

Experiments and Analysis

In this chapter we look at the results from our experiments as outlined in Chapter 5, and analyze the results. The experiments are divided into 3 sections: facial feature detection, input feature selection, and classification method.

6.1 Facial Feature Detection

We tested 4 different algorithms for facial feature detection. First we tested Horng et al.'s method and our ad hoc algorithm which is similar. Next we test 2 ASM based implementations: asmlib and Stasm.

6.1.1 Horng et al.'s Method Performance

We found that Horng et al.'s methods will only work on images very close to passport style photographs, which requires good illumination, proper pose, and a neutral expression. This method will fail on the majority of the images from the FG-Net database.

6.1.2 adhoc Performance

Figure 6.1 shows the output of our ad hoc algorithm on a subset of images from FG-Net. Our algorithm only detects the location and approximate size of the eyes, nose, and mouth using a combination of both Sobel filters and Haar-like features. The predicted landmarks are mostly accurate. However, the algorithm is prone to confuse facial features, such as eyebrows for eyes. If the algorithm detects a feature correctly, the detection will be accurate to within 4% of the image's width of the true location of the feature. Our algorithm correctly detects both eyes in 90% of cases,

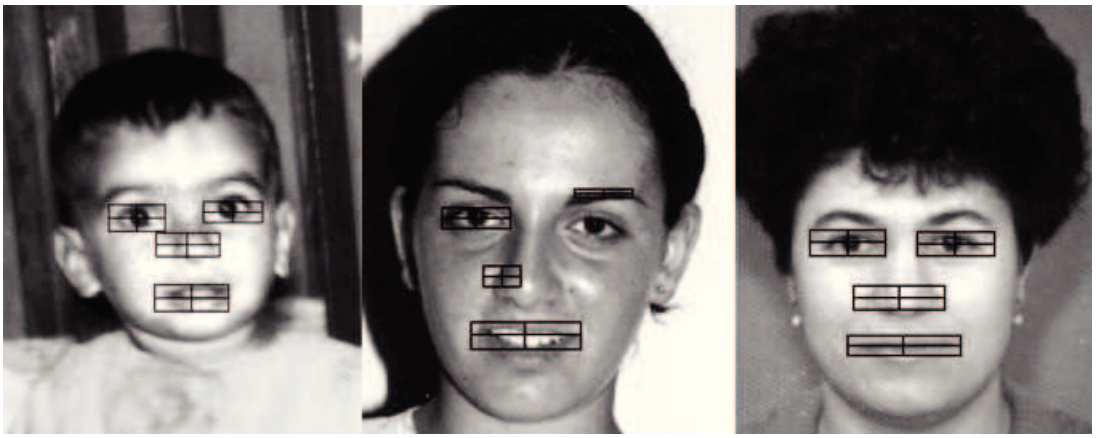


Figure 6.1: Sample output of out ad hoc algorithm. Predicted landmarks are mostly accurate, but is prone to confusion as seen in the middle image.

80% for nose, and 85% for mouth. This recall rate is low compared to Stasm. Stasm also detects more facial features (landmarks), therefore we will not continue exploring this algorithm.

6.1.3 Asmlib Performance

Asmlib is an open source ASM library using OpenCV. It will predict the location of facial landmarks as specified by the given model. In general, it will outline the face, eyebrows, eyes, nose, and mouth. Figure 6.2 shows the output of asmlib on a subset of images from FG-Net. The first two rows are results using different models included in asmlib. The bottom row shows the landmarks provided with the FG-Net database. The predicted landmarks are mostly inaccurate and differ greatly from the landmarks shown in the last row. More details on the landmarks are in Section 6.1.4. Most important features such as the eyes, nose, and mouth; were not correctly located, therefore asmlib is not suitable for our purposes, and we will not continue testing it.

Figure 6.2: Sample output of asmlib using differently trained models (provided with asmlib). The first 2 rows are the output of asmlib, and the third row shows the landmarks as provided in the FG-Net database for comparison. Predicted landmarks are mostly inaccurate.

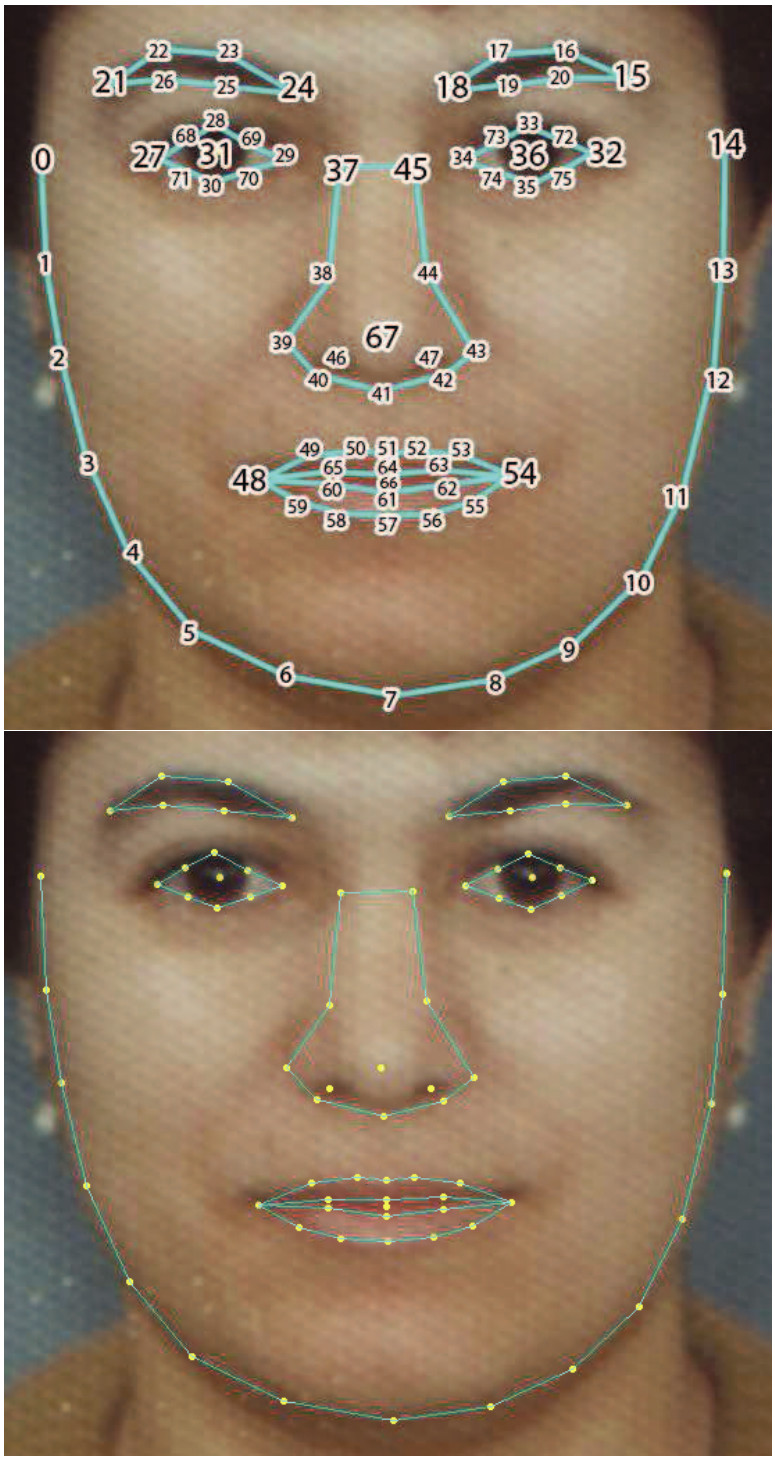


Figure 6.3: The 76 Milborrow / University of Cape Town (MUCT) landmarks. The first 68 are defined in the FG-Net markup, plus 4 extra points for each eye. The first image shows the corresponding number for each landmark, and the second image shows the landmarks more clearly.

6.1.4 Stasm Performance

Stasm is an extension to ASM by Stephen Milborrow, with various modifications to improve results. Stasm predicts a total of 76 facial landmarks, as shown in Figure 6.3. The 76 landmarks (or Milborrow / University of Cape Town (MUCT) landmarks) are obtained from joining the 68 landmarks as defined in the FG-Net markup, plus 4 extra points for each eye. Landmarks 0 to 68 in the MUCT landmarks are identical to the FG-Net landmarks. Landmarks 0 to 14 correspond to the edge of the face, landmarks 15 to 20 circle the subject's left eyebrow, landmarks 21 to 26 circle the right eyebrow, landmarks 27 to 30 and 68 to 71 circle the right eye, landmarks 32 to 35 and 72 to 75 circle the left eye. Landmark 31 denotes the center of the right eye, and landmark 36 denotes the center of the left eye. Landmarks 37 to 45 bound the nose, landmarks 46 and 47 denote the location of the subject's nostrils, and landmark 67 denotes the tip of the nose. Landmarks 48 to 59 circle the outer boundary of the mouth around the lips. Landmarks 60 to 62 mark the top of the lower lip, and landmarks 63 to 65 mark the bottom of the upper lip. Landmark 66 denotes the middle of the mouth.

On average, Stasm detects the 68 features specified in the landmarks file provided with the FG-Net database to within 8.95 pixels, or 2.21% of the width of the image. Stasm performs most poorly on detecting the features that correspond to the edge of the face. Without including the 15 landmarks that correspond to the edge of the face, Stasm is accurate to 7.82 pixels, or 1.93% of the width of the image. While not perfect, the performance is far better than the other methods that are usually only correct to within 4% or more of the width.

As shown in Figure 6.4, Stasm performs the worst on finding the edge of the face. It performs better on finding the eyebrows, and significantly better on detecting the eyes themselves. Stasm also finds the nose accurately although it performs badly on features 37, 38, 44, and 45 which correspond to the top of the nose. It also performs

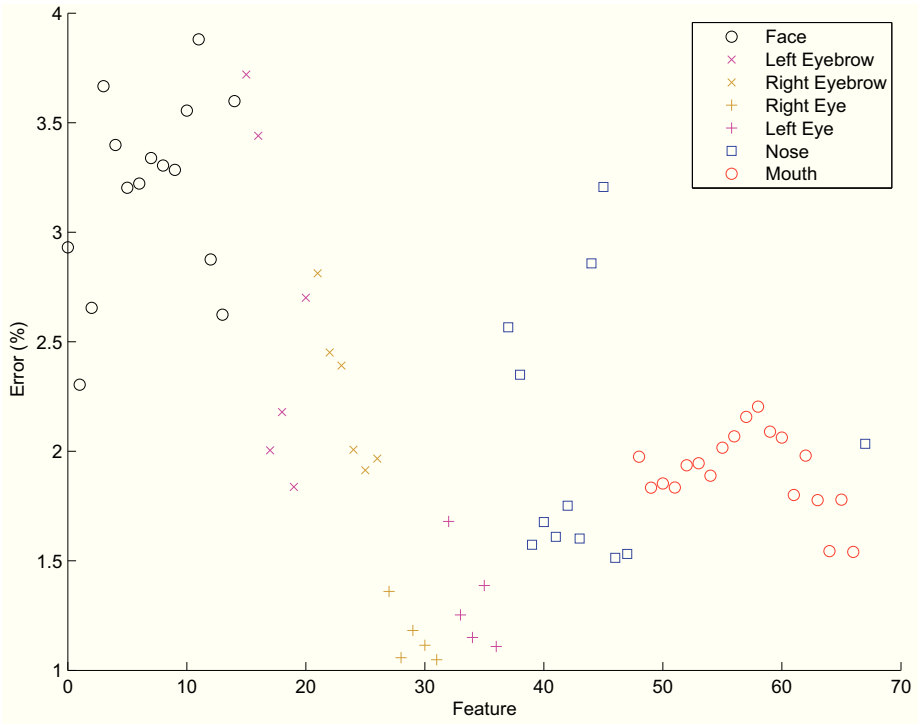


Figure 6.4: Error of Stasm on FG-Net database. The x-axis corresponds to the 68 FG-Net landmarks as numbered in Figure 6.3. Note the low error on the eyes, and the higher error on the face edges.

quite well at finding the mouth features, especially feature 66 which is the center of the mouth.

Figure 6.5 shows some results of running Stasm on FG-Net database. Notice that the landmarks are mostly accurate. However, there are some mistakes such as on the chin of the subject in the middle of the bottom row.

6.2 Input Feature Selection

In this section we analyze why ratios are a sufficient input feature, and the methods to select ratios. We analyze the performance of ratios chosen with a priori knowledge, and ratios chosen by a performance metric. We also look at more complex features, and decide on a final set of input features.

Figure 6.5: Sample output of Stasm. Predicted landmarks are mostly accurate (correctly aligned with the facial features). Note mistake on chin in bottom middle image.

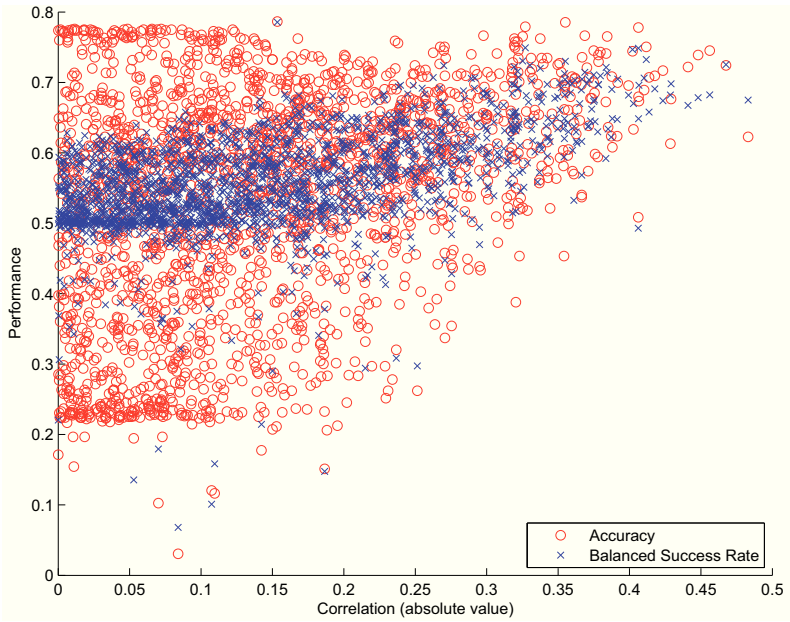


Figure 6.6: The dependence of ratio performance on Pearson's correlation (absolute value) to age. This figure shows a random sample of 2015 ratios from the 4 million possible. Each value is plotted twice, once for accuracy and once for BSR. Note that there is a visible trend where both accuracy and BSR increases along with correlation.

6.2.1 Ratio Correlation

Stasm produces the locations of 76 landmarks. We can form 2850 unique pairs from these landmarks, which gives us 2850 distance measures. From this, we can produce 4,059,825 unique ratios. Each of these ratios represent some feature of the face as a real value. Using some subset of these ratios, we will be able to represent the face completely.

To measure the usefulness of these ratios, we first computed the Pearson's correlation between the age and each ratio. The Pearson's correlation is a measure of correlation or linear dependence between two variables, X and Y (in this case, ratio value and age).

Pearson's correlation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Next, using leave-one-out cross-validation (LOOCV), we used a linear threshold classifier that maximizes BSR to measure the accuracy and BSR of each ratio. The classifier decides the direction of the threshold based on the sign of the correlation. The result for a randomly sampled set of 2015 ratios (ratios are chosen with uniform probability) is shown in Figure 6.6. In general we can observe that the BSR tends to be proportional to the correlation scores. We can see that most scores tend to be in the 0.5 to 0.6 range, even when the correlation is low (indicative that this ratio does not contain any information that can be used to determine age). This is because in such cases the threshold chosen is likely to be an extreme value such as to favor the majority class, resulting in near 100% accuracy for the majority class and near 0% accuracy for the minority class, which results in a BSR of 0.5.

However, we can clearly see a relationship between correlation and BSR, where BSR increases as correlation increases. The results show that there are many ratios which will result in high BSR, meaning that ratios are meaningful features that we can use as input features. Also, the correlation between a ratio and age is meaningful, and we can judge and choose ratios based on their correlation.

6.2.2 Odd Behavior of BSR with LOOCV in Ratio Correlation

Note: This section will explain why there are some data points in Figure 6.6 that have extremely high or low BSR although their correlation is close to 0. It is not necessary to understand this section for the rest of the thesis, and it can be skipped.

In this section we will be careful and define a few additional terms concerning BSR to be clear. In one single experiment during LOOCV, we will obtain a *trained BSR* on the training data, and a *test BSR* on the test data. Note that test BSR is either 0 or 1 as the testing data, or left out datapoint, is of size 1, meaning it will either be correct or wrong. Additionally, the BSR term used in the previous section will now be called *ratio BSR*, as it is the BSR obtained after performing LOOCV on

the ratio and computing the ratio's BSR from the test BSRs obtained for each left out experiment. The LOOCV process is as following:

1. For each datapoint in the training data,
2. temporarily remove it,
3. train a linear threshold classifier on the data with the point removed (maximizing trained BSR),
4. use the trained classifier to make a prediction on the removed datapoint (obtain test BSR),
5. reinsert the removed datapoint.
6. Compute the ratio BSR based on all the test BSR obtained.

There are ratios that have ratio BSR values that deviate significantly from 0.5, even though the correlation is near 0. This is surprising because the low correlation means that these ratios should not have much meaningful information, and therefore should have ratio BSR values close to 0.5. This is indicative that there is something non-intuitive happening in the LOOCV process.

We will analyze one particular ratio that has correlation close to 0 and ratio BSR of 0.068. We first analyze the trained BSRs that we obtain using different thresholds, by setting a threshold then classifying the whole dataset without withholding any data (nothing is left out). Figure 6.7 shows that for this particular ratio, there are few maximum points on the graph that are close to 0.5 trained BSR. This is strange, as this seems to indicate that if we pick any threshold in this range, the ratio BSR for this ratio should be close to 0.5, but instead it is only 0.068. We know that to maximize ratio BSR, we should maximize the test BSR in each LOOCV experiment. However, since we cannot directly maximize test BSR, we maximize trained BSR instead. Therefore we have should select the threshold that has the highest trained BSR in this graph. Ideally, there should have been only one maximum point which indicates the best threshold value for that particular ratio. For this particular ratio, having more than one maximum point means that the threshold chosen during

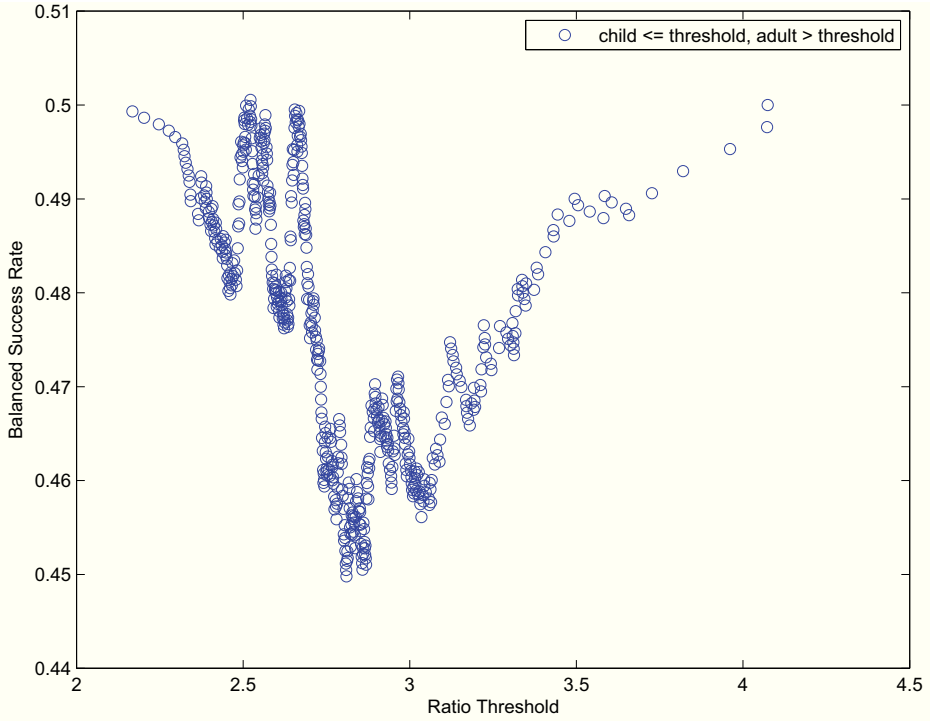


Figure 6.7: The dependence of test BSR on ratio threshold for ratio (6,14 / 18,43). The ratio is described as a pair of landmarks divided by another pair of landmarks, where we will compute the distance between the values in a landmark pair. Test BSR is maximized when choosing a threshold that corresponds to a maximum point. For this ratio, there are several local maximums at 2.2, 2.5 to 2.7 range, and 4.2.

LOOCV can change, as these maximum points will be chosen as thresholds as they have the highest trained BSR values. Since these values are very similar, any one of these values could be chosen as the threshold, and this, as it turns out, depends strongly on the label of the left out datapoint, which is non-intuitive.

During the LOOCV process, the threshold chosen alternates between these thresholds based on the left out datapoint. These thresholds are coincidentally near the minimum and maximum of the range. If the threshold is selected as the maximum of the range, it means that it will label everything as child. Conversely, if the threshold selected is the minimum of the range, it will label everything as adult, *except* for any datapoint that has ratio value of the minimum, as the classifier classifies anything less than or *equal* to the threshold as a child.

Therefore it happens that when the left out datapoint belongs to the adult class, the threshold is set to 4.074 in order to maximize trained BSR, as when the threshold is 4.074 the recall rate for the child class is 1 and for the adult class is 0 (everything is labeled as child), and therefore trained BSR is maximized at 0.5. Interestingly, it is not set to the minimum (2.168), because the 2.168 value belongs to a data point in the adult class. Since the classifier uses ratio $\leq$ threshold, the classifier will misclassify that single data point and result in a slightly less than 0.5 trained BSR. This is shown in Figure 6.8.

On the other hand, if the left out datapoint belongs to the child class, then the threshold gets set to either 2.522 or 2.668 but not the minimum (2.168) because of the distribution of the data that results in slightly higher trained BSR at those points. As these points are close to the minimum, the recall for the child class will be close to 0, and the recall for the adult class will be close to 1.

We now know that the threshold chosen alternates between 4.074 and 2.522/2.668, depending on the class of the left out datapoint, as seen in Figure 6.8. Now observe that whenever the left out datapoint belongs to the adult class, the threshold chosen is always 4.074. This means that in each of these cases, the classifier will classify anything with ratio value less than 4.074 as child. This in turn means that every single time during the LOOCV process when the left out datapoint belongs to the adult class, the linear threshold classifier built for that instance will always misclassify the left out datapoint! This is a very surprising result.

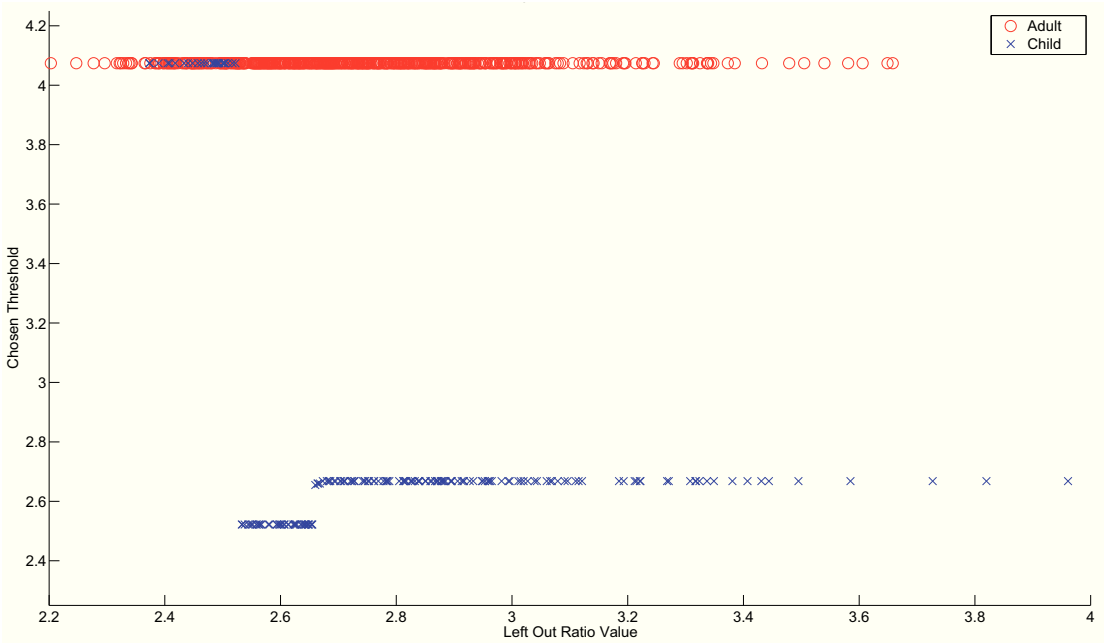


Figure 6.8: Threshold chosen based on the ratio value and classification of the left out datapoint for ratio (6,14 / 18,43). The ratio is described as a pair of landmarks divided by another pair of landmarks, where we will compute the distance between the values in a landmark pair. The left out datapoint has a value and can belong to either child or adult class. Note that for this particular ratio, the chosen threshold depends mostly on classification of the left out datapoint, as seen by the distinct clustering of the child and adult classes although they have overlapping ratio values. Ideally, the threshold chosen should not depend on the classification of the left out datapoint at all, meaning that if two points have the same ratio value, the chosen threshold should be identical.

If the left out datapoint belongs to the child class, then the threshold could be either 2.522 or 2.668 depending on the value of the left out datapoint. This is less extreme than the adult case since these values are not the minimum values of the range. As we can see in Figure 6.8, 2.668 is chosen most of the time. Since this figure also shows the distribution of the dataset, we can also see that some of the child datapoints are less than 2.668, and most are greater than it. So, for the case when the left out datapoint belongs to the child class and has ratio value less than 2.668, the trained linear threshold classifier will correctly classify that datapoint. However, if

the left out datapoint belongs to the child class but has ratio value greater than 2.668, then the trained linear threshold classifier will incorrectly classify that datapoint.

Since for this ratio the trained linear threshold classifier will always be wrong for the adult datapoints, and wrong more than half the time for the child datapoints, it will have a ratio BSR worse than 0.5 (0.068 in this case). The result is that the linear threshold classifier for this particular ratio is artificially bad, as the threshold alternates during the LOOCV process. As we can see in Figure 6.7, if we did not do LOOCV and just picked an arbitrary threshold, then the ratio BSR will always be close to 0.5 (at least 0.45, as compared to 0.068). This is an interesting effect of the LOOCV method that can result in extraordinarily bad or good results by alternating between extremely different hypothesis based on the left out datapoint. However, for our purposes we can ignore such outliers and focus on the observable upwards trend.

One last thing to note is that this occurs because the threshold, which has been set to point in a particular direction according to the correlation, is actually pointing the wrong way because the correlation is near 0 (since when the correlation is near 0, the direction of the threshold has very little meaning). Since the direction is wrong, the best ratio BSR achievable is usually 0.5 which occurs at the minimum and maximum of the range. If the direction was correct then Figure 6.7 will be flipped on the y-axis at 0.5, resulting in a mirrored image in which the maximums are near the middle, which will result in a much more reasonable result from LOOCV.

6.2.3 Manually Chosen Ratios

We chose 22 ratios that we felt cover most of the variation of the face that corresponds to age. We call this set of ratios as Manual22. These ratios were selected such that they would change according to growth in human heads. These ratios are shown in Figure 6.9 and their correlation to age, relative rank by correlation, accuracy, and

balanced success rate on the FG-Net database are shown in Table 6.1. The accuracy and BSR values are obtained by LOOCV using a linear threshold classifier.

The first 10 ratios concentrate on horizontal distance (distance spanning the width of the image) over vertical distance (distance spanning the height of the image). In particular, these 10 ratios all use the distance between the center of each eye. The first 5 also use the distance from the mean of the eyes to the location of a particular feature of the face. The second 5 use the distance from the particular feature to the chin. The first 5 have moderately low correlation to age, whereas the second 5 have moderately high correlation. This corresponds to our analysis that the secondary growth (growth related to dentition and mastication) is concentrated on the lower half of the face.

The next 5 (ratios 11 to 15) use 2 vertical distances, generally the distance from the mean of the eyes to a particular feature and the distance from that feature to the chin. In general, these ratios have correlation scores that are significantly lower than the ratios that have both horizontal and vertical distances.

The next 3 (ratios 16 to 18) focus on the distance from the mean of the eyes to the chin, a mainly vertical distance. The first 2 of the 3 ratios include a horizontal distance, and have very high correlation. The third includes another vertical distance and performs significantly worse. In particular, ratio 17 seems to be better than ratio 16, which is likely because it covers a larger horizontal distance. The next 3 (ratios 19 to 21) attempt to compensate for the noise in the eye to chin distance created when the subject has their mouth wide open (e.g. from talking or eating), by subtracting the distance between the upper lip to the lower lip. This normalization actually results in a significant reduction in correlation. However, this is caused by Stasm's inaccuracy in predicting the location of the upper and lower lip. In another experiment, using the landmarks given in the FG-Net database, this normalization

Table 6.1: Ratio, Pearson's Correlation to Age, Rank by Correlation, Accuracy and BSR. The correlations are negative because the ratios are inversely proportional to age (the older the person, the smaller the ratio value will be).

	Ratio	Corr	Rank	Acc	BSR
1	eyeToEye / eyeToNose1	-0.111	18	0.677	0.653
2	eyeToEye / eyeToNose2	-0.163	13	0.610	0.577
3	eyeToEye / eyeToMouth1	-0.151	16	0.560	0.636
4	eyeToEye / eyeToMouth2	-0.217	10	0.727	0.661
5	eyeToEye / eyeToMouth3	-0.185	12	0.691	0.649
6	eyeToEye / noseToChin1	-0.269	8	0.514	0.591
7	eyeToEye / noseToChin2	-0.302	7	0.541	0.626
8	eyeToEye / mouthToChin1	-0.261	9	0.528	0.617
9	eyeToEye / mouthToChin2	-0.313	5	0.542	0.568
10	eyeToEye / mouthToChin3	-0.304	6	0.634	0.649
11	eyeToNose1 / noseToChin1	-0.045	22	0.234	0.499
12	eyeToNose2 / noseToChin2	-0.099	19	0.755	0.514
13	eyeToMouth1 / mouthToChin1	-0.086	20	0.706	0.519
14	eyeToMouth2 / mouthToChin2	-0.159	15	0.611	0.548
15	eyeToMouth3 / mouthToChin3	-0.194	11	0.688	0.542
16	eyeToEye / eyeToChin	-0.385	2	0.754	0.746
17	faceWidth / eyeToChin	-0.390	1	0.752	0.751
18	noseLength / eyeToChin	-0.163	14	0.392	0.516
19	eyeToEye / (eyeToChin - mouthToMouth)	-0.330	4	0.699	0.726
20	faceWidth / (eyeToChin - mouthToMouth)	-0.350	3	0.732	0.737
21	noseLength / (eyeToChin - mouthToMouth)	-0.144	17	0.635	0.528
22	lowerFaceWidth / lowerFaceLength	-0.070	21	0.635	0.631

resulted in a significant improvement in correlation. However, since we use landmarks as located by Stasm, this normalization is detrimental for our purposes.

From these results, we can conclude that: horizontal distance over vertical distance ratios are the best because they capture the elongation that occurs in growth; the larger the distance, the better the ratio, as the measurements become less noisy; and normalization by means of accounting for distance between lips decreases correlation.

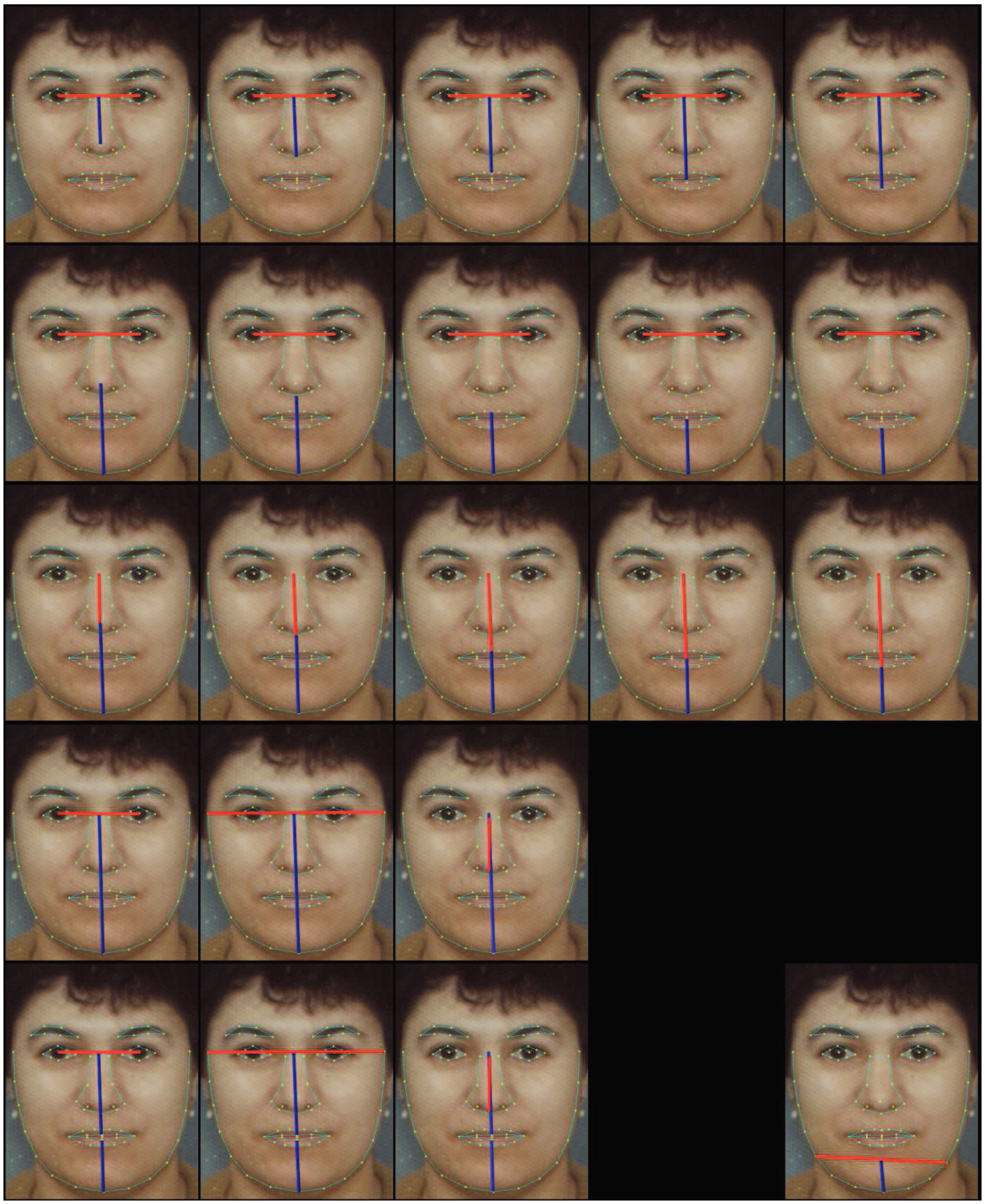


Figure 6.9: Manually chosen ratios. The images correspond to the ratios in Table 6.1 from left to right, top to bottom. Notice for ratios 19 to 21, there is a gap between the upper lip and lower lip.

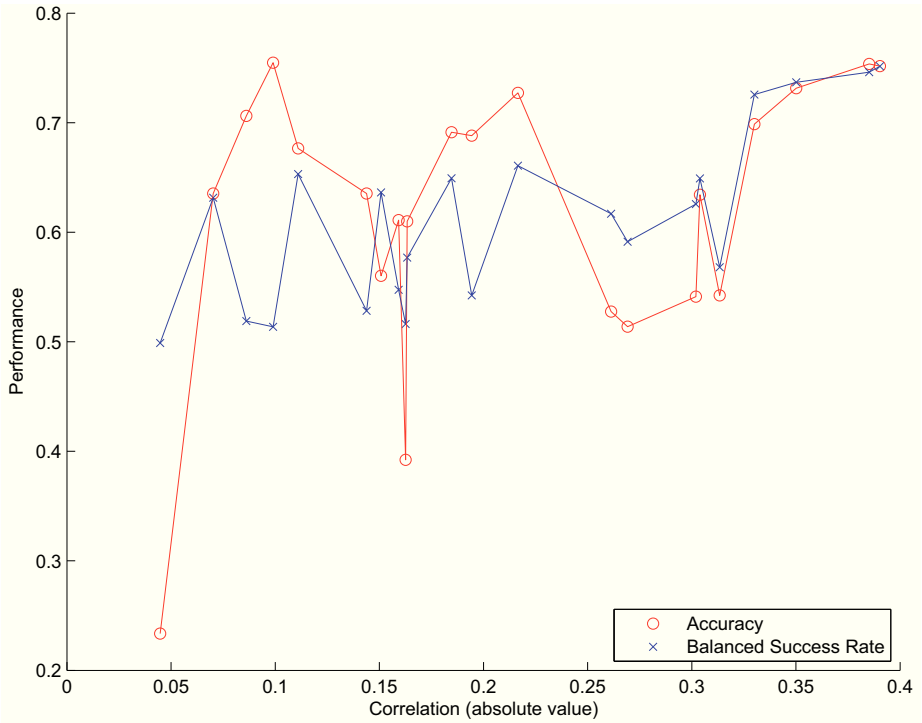


Figure 6.10: Performance of the manually chosen ratios in relation to correlation. In general, there is a noticeable trend where the performance increases along with correlation.

Figure 6.10 confirms our conclusions in Section 6.2.1, which is that correlation is a good indicator of the performance. This holds especially true for high correlation ratios, meaning that if the correlation is high, we can expect it to perform well (using a linear threshold classifier). We used all 22 ratios as input to an C-Support Vector Classification (SVC) SVM with a RBF-kernel using default parameters in libsvm and weighing scheme specified in Section 6.3.1, and the results are shown in Table 6.2. Note that only 946 of the 1002 images in FG-Net are used, as Stasm fails to detect faces (failure on the part of the face detector and not Stasm per se, as the Stasm software runs a face detector on the image first to obtain a more suitable starting point) on 56 images. The results from the SVM are slightly higher than the best performing ratios from the linear threshold classifier, meaning that the ratios contain different information that can be combined to achieve higher performance.

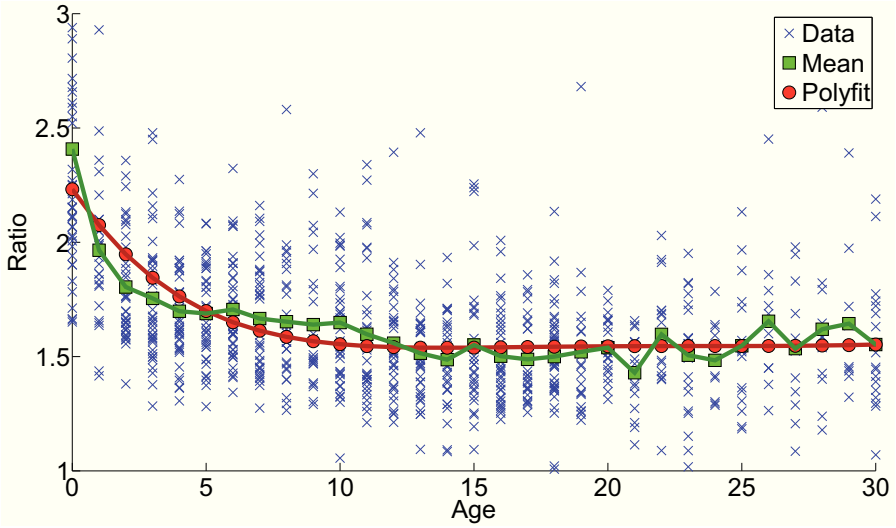
Table 6.2: Results from 10-fold cross validation using SVM on Manual22.

Performance Metric	Value
Child Recall	157/213 (0.737)
Adult Recall	595/733 (0.812)
Accuracy	0.795
BSR	0.774
Age Sensitive Accuracy	0.843

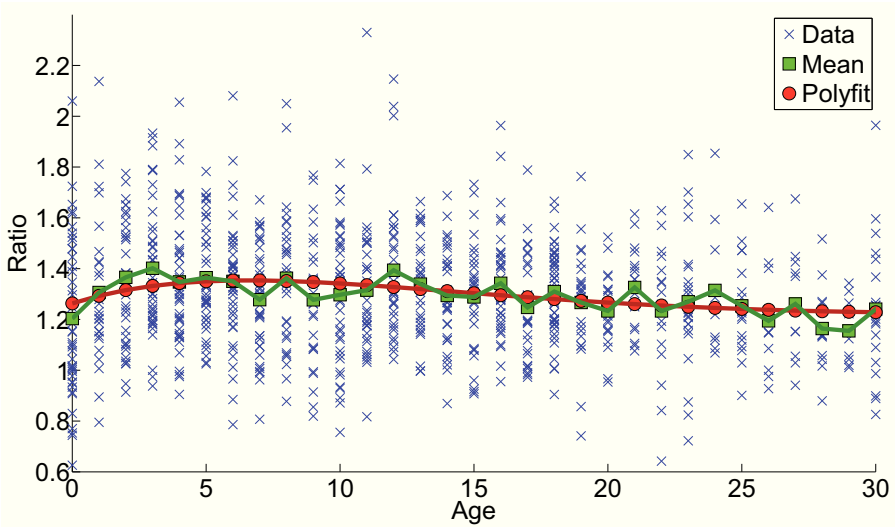
Figure 6.11 shows the distributions of 3 of the manually chosen ratios. The green boxes mark the mean of the ratio per age, and the red circles mark a 6th degree polynomial regression fit. For these ratios, the distribution becomes less meaningful after the age of 30, as the sample size per age decreases significantly. In general there is an observable trend where there is a great amount of monotonic change in the ratio within the first 10 years. In Figure 6.11(b), this monotonic change is lacking, and this corresponds to the poor performance of this ratio.

6.2.4 Pose Variation

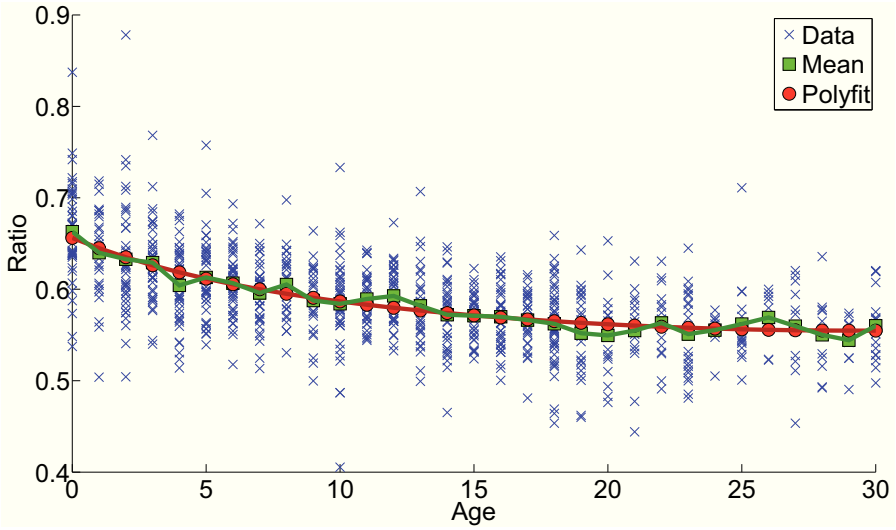
As the images we process are not guaranteed to contain faces photographed perfectly straight on, there will be significant pose variation in the image. We found that variance in the horizontal angle (looking left or right) does not significantly impact our results, assuming that the horizontal angle does not go beyond 30° (we will assume that the input images will not include images with more significant variance than this). However, we found that the vertical angle (tilt) greatly affects the ratios as it changes distance between features. We ran experiments on images obtained from the Head Pose Image Database (HPID). HPID contains 2790 face images of 15 adults with varying poses in terms of vertical and horizontal angles. A sample of this dataset and Stasm's result on the images is shown in Figure 6.12.



(a) eyeToEye / eyeToNose1



(b) eyeToMouth1 / mouthToChin1



(c) eyeToEye / eyeToChin

Figure 6.11: 3 sample ratios from Manual22. For (a) and (c) the change in ratio value according to age is monotonic. For (b) this monotonic change is not present and the line for the polyfit is a lot more flat. This shows that (a) and (c) are useful ratios, whereas (b) is less informative.

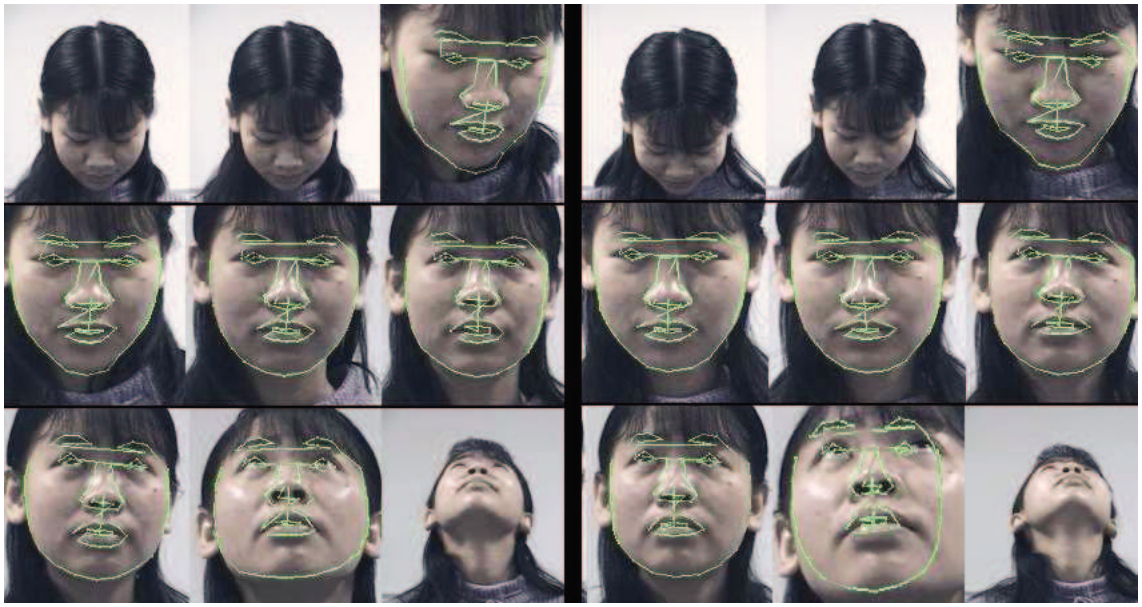


Figure 6.12: Sample prediction on 2 image series on same subject from Head Pose Image Database. The 3×3 on the left constitute 1 series, and the other 3×3 is another. Vertical angle (tilt) from left to right, top to bottom: -90, -60, -30; -15, 0, +15; +30, +60, +90. Note failure on detecting face, and therefore features in -90, -60, and +90, but success on +60.

Figure 6.13 shows a subset of the manually chosen ratios as the ratio values change as the vertical angle of the person changes. This information is calculated from the Head Pose Image Database. For our purposes we used only images with horizontal angle = 0 (facing forward). We also ignored the -90, -60, +60, and +90 vertical angle images because the face is too obscured most of the time. The values are computed as the mean of all images in the specified pose. As we can see, most of these ratios vary according to pose. For example, eyeToEye/eyeToNose1 increases greatly as the subject looks upwards. This makes intuitive sense as the nose gets closer to the eyes (when observed from the front of the subject) as the subject looks up. This is not a desirable trait as this means that pose greatly influences the ratios, making the face appear to be more "childlike" as the subject looks upwards.

However, there are ratios that are more resistant to pose variation, such as the eyeToEye/eyeToChin ratio. This ratio remains almost constant as subject's tilt varies.

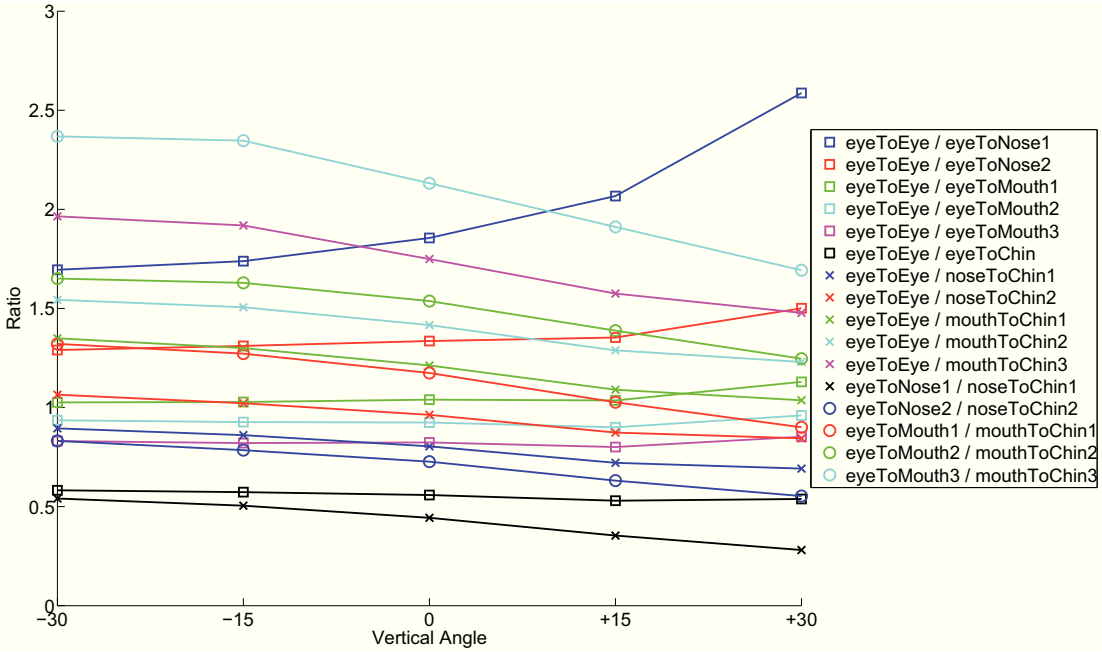


Figure 6.13: Average ratio value according to pose on images from the Head Pose Image Database. Each line describes a ratio and how its value changes according to pose (vertical angle).

This makes sense as the eye to chin distance is much larger and should vary less. We should also note that this is also because Stasm does not detect the same spot on the chin, in the sense that when the tilt varies, the predicted chin is not necessarily the same spot as shown in Figure 6.12. This is mainly due to illumination and Stasm trying to reduce deformation of the predicted face, since it is meant to be used on faces that are facing directly forward. The prediction of the chin greatly depends on where Stasm thinks the rest of the face edge locations are. This results in the subject having roughly the same detected face height regardless of pose. This is desirable, and the eyeToEye/eyeToChin ratio is discriminative with respect to age from Table 6.1, therefore this is a ratio that is robust to change, and able to distinguish between adult and children. This pose invariance also contributes to why the eyeToEye/eyeToChin ratio is superior to the other horizontal over vertical ratios.

Table 6.3: Results from 10-fold cross validation using SVM on Highest43.

Performance Metric	Value
Child Recall	171/213 (0.803)
Adult Recall	592/733 (0.808)
Accuracy	0.807
BSR	0.805
Age Sensitive Accuracy	0.860

6.2.5 Automatically Chosen Ratios

While the manually selected ratios perform well, they are selected by expert knowledge (or a priori human knowledge), which can be incorrect. We tried a different approach, a more methodical way of selecting ratios. Since there are over 4 million possible ratios, we need to do some sort of feature selection; otherwise, the data will be too high dimensional. As we know that the correlation between ratio value and age is indicative of its performance, we computed the correlation of all 4 million ratios.

One method is to take the best k ratios. In this case we just choose all ratios that have absolute value correlation over 0.5, giving us 43 ratios. We call this set of ratios as Highest43. We used this as input data into an SVM with the same settings in Section 6.2.3. The results are shown in Table 6.3. The results show an improvement over the manually chosen ratios.

However, Highest43 has many redundant ratios, as the best ratios tend to be very similar where only 1 end point (of the 4 facial features) in the ratio changes. Because of this redundancy, we do not use most of the information in the face as we actually cover very few facial features.

The second approach is to select ratios that are independent of one another. First we selected ratios that have correlation greater or equal to 0.4. This gives us 40,092 ratios. Then by computing correlation between the ratios, and removing any ratio that has correlation of over 0.9 with another ratio that has higher correlation, we end up with a more reasonable 110 ratios. We call this set of ratios as Independent110.

Table 6.4: Results from 10-fold cross validation using SVM on Independent110.

Performance Metric	Value
Child Recall	169/213 (0.793)
Adult Recall	634/733 (0.865)
Accuracy	0.849
BSR	0.829
Age Sensitive Accuracy	0.897

As in the previous section, we used these ratios as input into a similar SVM. The results are shown in Table 6.4. The results show significant improvement compared to the other methods.

The 16 ratios in Independent110 with the highest correlation are shown in Figure 6.14. It is difficult to draw any conclusion from these ratios, as they include a mix of both the vertical over horizontal ratios, but also include many dual vertical ratios (almost parallel vertically). The vertical over horizontal makes sense for the most part, but the dual vertical ratios are a little more interesting. Notice that the distances in these ratios are usually almost parallel and are of almost the same length, rather than in series such as the ones we experimented with in the manual ratios. To understand why these parallel ratios have high correlation, we look at an example. In ratio 3 in Figure 6.14 (top row 3rd from left), the blue line is slightly longer than the red. Imagine if the face belonged to a child, then the mouth would be closer to the eyes (vertically). As the blue line has a larger horizontal component than the red line, this decrease in the vertical component for both lines results in a larger decrease in the percentage of the length for the red line than for for the blue line. Therefore, the blue line would be even longer than the red, changing the ratio values.

We also took the top 40,092 ratios (ratios having correlation ≥ 0.4) and created Figure 6.15. This image shows which distances/lines appear most often in the top 40,092 ratios, meaning these are the most informative lines. The lines are colored from white to blue, where the more blue the line is, the more often it is used. The

landmarks are also colored similarly based on frequency from red to green. The more red the landmark is, the less it is used, and the more green the landmark is, the more it is used. It appears that there are 2 groups of lines that get used the most often. The first group contains lines from an eye to the bottom of the face. The second group contains lines that go across the face, either from eye to eye or left eye to right eyebrow and vice versa. This indicates that the best performing ratios are indeed those that consist of lines that extend across the face either width-wise or length-wise. Interestingly, landmarks 0 and 14 (the edges of the face) are rarely used, likely because their precision, as detected by Stasm, is lower than that of the eye landmarks. Independent110 selects from these ratios the ones with the highest correlation to age, and least correlation with each other.

6.2.6 Complex Features

In sections 6.2.3 and 6.2.5, the features that we used were simple distance ratios usually using only data from 4 landmarks. The most complex ratio, (faceWidth / (eyeToChin - mouthToMouth)), used a total of 7 landmarks as it used both eye center landmarks for computing eyeToChin, and included the additional mouthToMouth distance. In this section, we experiment with more complex features that uses more landmarks. The two complex features we explore are size ratios and features based on angles between facial landmarks. An example of a size ratio would be the size of an eye over the size of the face. Intuitively, since the eye does not grow as much as the face, this size ratio should be useful. An angle-based feature is the computed angle between 2 lines, instead of just computing the distance ratio. The distance ratio uses magnitude information from the line formed between two landmarks, but an angle-based feature would use the direction information.

The first complex feature we experiment with is size ratios, as opposed to just distance ratios. The size of each facial feature is estimated by computing the area

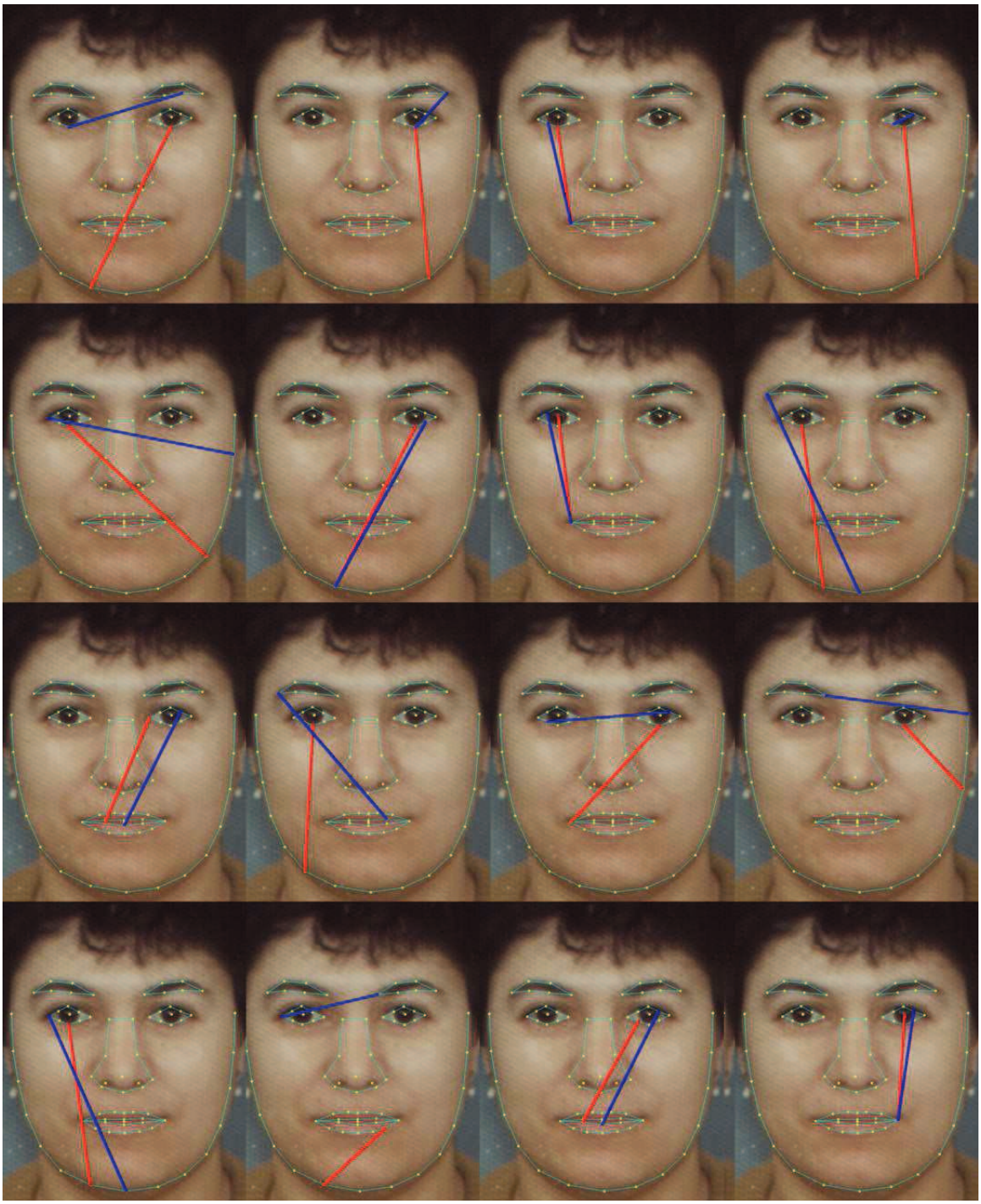


Figure 6.14: Top 16 of the Independent110 ratios. The ratio with the highest correlation is on the top left, and decreases across the row, then proceeds to the next row.

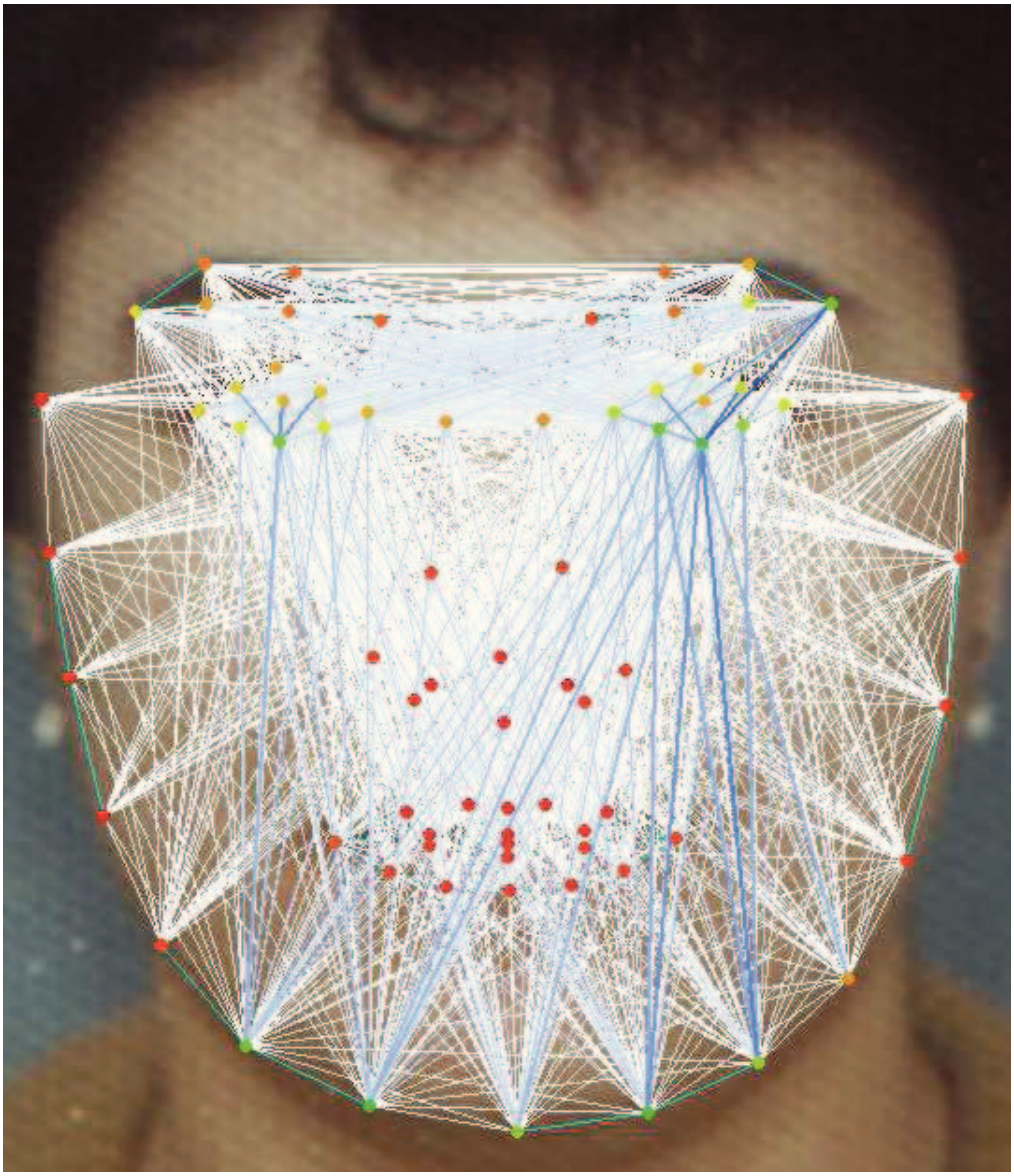


Figure 6.15: Image created from the top 40,092 ratios that have the highest correlation to age. Each ratio uses 2 distances or lines. The lines are drawn in this image with their color corresponding to the number of times they appeared in the top 40,092. If a line appears once, it will be colored white, and the more blue the line is, the more it appeared. The green lines are lines that never appeared in the top 40,092. Likewise for the landmarks, which are colored from red to green, the more red the landmark is the less frequent it is, and the more green the landmark, the more frequent it is.

Table 6.5: Complex feature and Pearson's correlation to age.

	Feature	Correlation
1	leftEyeSize / faceSize	-0.387
2	rightEyeSize / faceSize	-0.416
3	meanEyeSize / faceSize	-0.420
4	noseSize / faceSize	-0.080
5	mouthSize / faceSize	-0.030
6	chinAngle	-0.072
7	faceAngle	-0.379

of the bounding box of the feature. We computed the correlation for 5 size ratios, taking the size of both eyes separately and together, nose, and mouth; over the size of the face (shown in Figure 6.16). The results are shown in Table 6.5. As expected, the eye related size ratios have high correlation to age. The eyes do not grow as much as the face, hence the common conception that babies have large eyes. The nose and mouth ratios have very low correlation, as they grow along with the face, making the ratio of their size to the size of the face more constant.

We also experimented with features based on angles between facial landmarks. We computed 2 angles: the chin angle, which is the angle between the vectors formed from landmark 7 to landmark 4, and landmark 7 to landmark 10; and the face angle, which is the angle between the vectors formed from landmark 7 to landmark 0, and landmark 7 to landmark 14 (shown in Figure 6.16). The chin angle has correlation close to 0, which is not unexpected as while the jaw does grow and become more defined, the shape of the jaw varies greatly according to the individual. The face angle has high correlation as it is related to the width-height ratio of the face.

Using the high correlation features (features 1, 2, 3, and 7 in Table 6.5), called Complex4, as input into an SVM, we achieve good results considering we only use 4 features. The results are shown in Table 6.6.

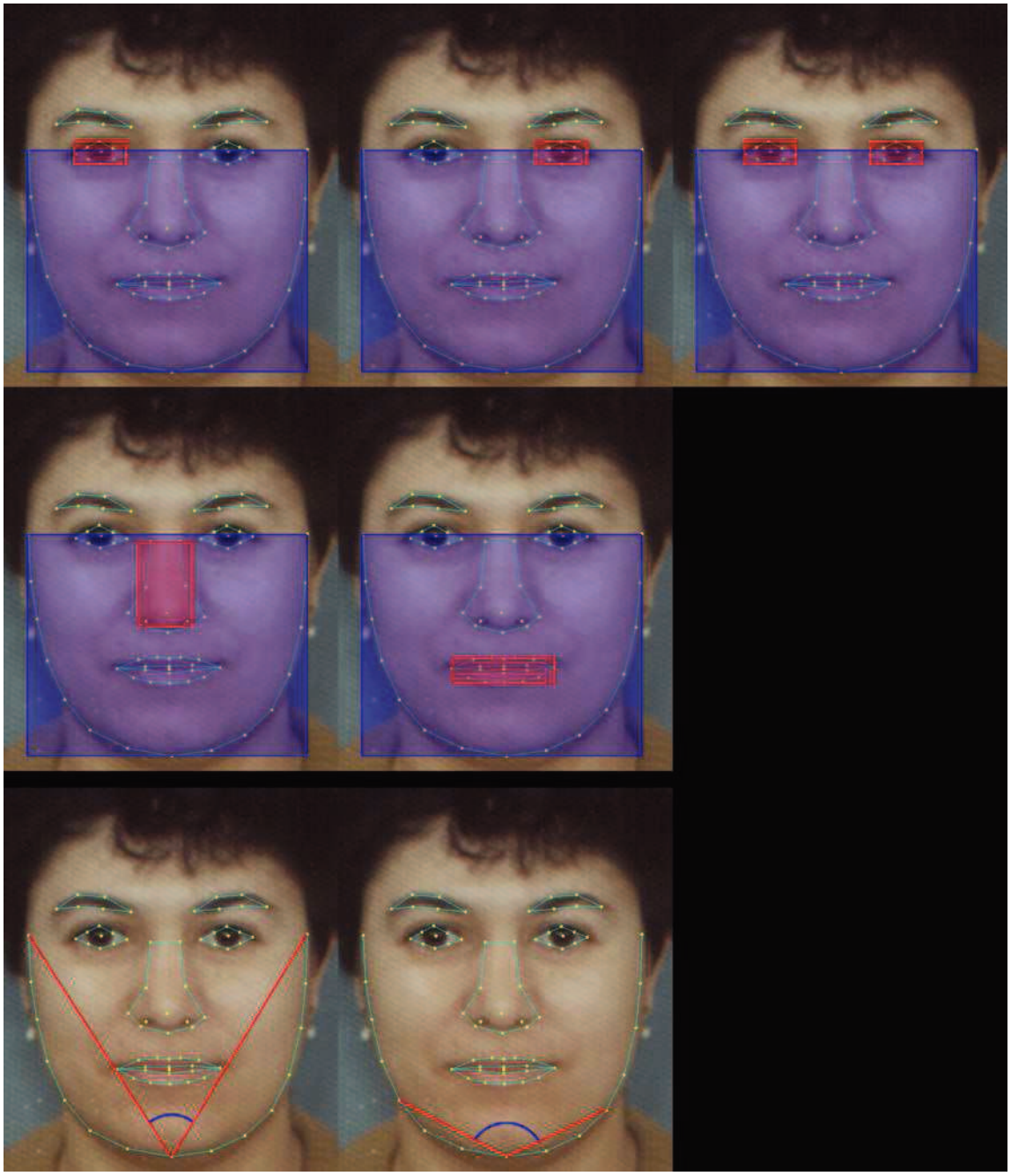


Figure 6.16: Complex ratios. The first two rows show the size ratios, and the third row shows the angle features (in blue).

Table 6.6: Results from 10-fold Cross Validation using SVM on Complex4 (ratios 1, 2, 3, and 7 from Table 6.5.

Performance Metric	Value
Child Recall	161/213 (0.756)
Adult Recall	570/733 (0.778)
Accuracy	0.773
BSR	0.767
Age Sensitive Accuracy	0.823

Table 6.7: Results from 10-fold Cross Validation using SVM on Combined136.

Performance Metric	Value
Child Recall	163/213 (0.765)
Adult Recall	626/733 (0.854)
Accuracy	0.834
BSR	0.810
Age Sensitive Accuracy	0.882

6.2.7 Combined Features

We combined Manual22, Independent110, and Complex4 as Combined136, and used it as input into an SVM. The results are shown in Table 6.7. The results are good, although they are slightly worse than the results for Independent110.

Additionally, the results from all the ratio sets (Manual22, Highest43, Independent110, Complex4, and Combined 136) are shown in Table 6.8.

Table 6.8: Results from 10-fold Cross Validation using SVM on all ratio sets.

Performance Metric	Manual22	Highest43	Inde110	Complex4	Comb136
Child Recall	0.737	0.803	0.793	0.756	0.765
Adult Recall	0.812	0.808	0.865	0.778	0.854
Accuracy	0.795	0.807	0.849	0.773	0.834
BSR	0.774	0.805	0.829	0.767	0.810
Age Sensitive Accuracy	0.843	0.860	0.897	0.823	0.882

6.3 Classification Method and Additional Datasets

We tried 2 different types of classifiers, Support Vector Machines and Ensemble Classifiers. We use libsvm [9] for SVMs, and the ClassificationEnsemble class in MATLAB's Statistics Toolbox for Ensemble Classifiers. We also show the results from testing on the VADANA dataset.

6.3.1 Support Vector Machines

We ran all tests and optimizations using the FG-Net database with a 10-fold cross-validation. We used a C-Support Vector Classification SVM, which is typical for classification problems. We decided to use RBF kernel SVMs, as we have no particular reason to choose polynomial or sigmoid kernels. It is also reasonable to use a linear kernel as we think the relationship between the class labels and attributes is linear, but we will stick with RBF kernels as a linear kernel is a special case of the RBF kernel [19]. Using linear kernels produces slightly worse results than RBF kernels in our experiments.

As we mentioned in Section 5.2, the frequency of the two classes (child and adult) are not balanced in the FG-Net database (only 23% are children). To avoid the SVM heavily biasing the adult (majority) class, we have to set the weights for each class. We decided that the most natural weighting scheme is to simply weight each class by the number of instances of the other class, such that each class will have equal influence in the primal optimization problem for C-SVC.

We also followed the guidelines of the authors of libsvm [19]. We used the Independent110 set of ratios as the data for optimization. First, we tried scaling the input data to the range [-1,1]. The purpose of scaling is to avoid attributes that have larger ranges from dominating attributes that have smaller ranges. However, scaling did not improve our results, likely because our data is fairly consistent in range in the first place. We also optimized the parameters c and γ. The parameter c is the

cost parameter to the SVM, which is a constant that balances the goals of having a large margin, and making correct predictions. A large value for c indicates that we wish to favor correct predictions, and a small value for c indicates we would like to maximize the separating margin. The parameter γ is a parameter to the RBF-kernel, which controls the flexibility of the decision boundary. A small value for γ means we will have a less flexible and smoother decision boundary. A large value for γ means we will have a more flexible decision boundary, but this can lead to overfitting.

However, we do not get significantly different results from using the default parameters of $c = 1$ and $\gamma = \frac{1}{110} \approx 0.0091$. Figure 6.17 shows the result of the grid search, varying c and γ by powers of 2 to obtain the BSR and accuracy. The best BSR we obtained is 0.834 (compared to 0.829 obtained with default values), obtained at $c = 16$ and $\gamma \approx 0.002$, which are relatively close to the default values. The best accuracy is 0.850 (compared to 0.849), obtained at $c = 16$ and $\gamma = 0.001$, which is again relatively close to the default values.

Further testing on Manual22 and Highest43 showed a more noticeable improvement. On Manual22, the best BSR is 0.792 (default was 0.774), and the best accuracy was 0.803 (default was 0.795). On Highest43, the best BSR is 0.825 (default was 0.805), and the best accuracy was 0.835 (default was 0.807). In conclusion, parameter optimization may improve results, but the default values produced decent results for our data. Note that the class weights specified at the beginning of this section will affect c, where the true value of c is actually $c * \text{weight}_{\text{label}}$.

We also tried the parameter optimization tools included in libsvm, but results were non-intuitive. The training process achieves high cross-validation accuracy 0.871 for Independent110, but when testing the trained model on the training data again, we obtained accuracy of 0.230 and BSR of 0.503, with 0.007 recall on adult and 1.000 recall on child. This is likely because their tool does not account for label weights, which is significant for us since our classes are heavily unbalanced.

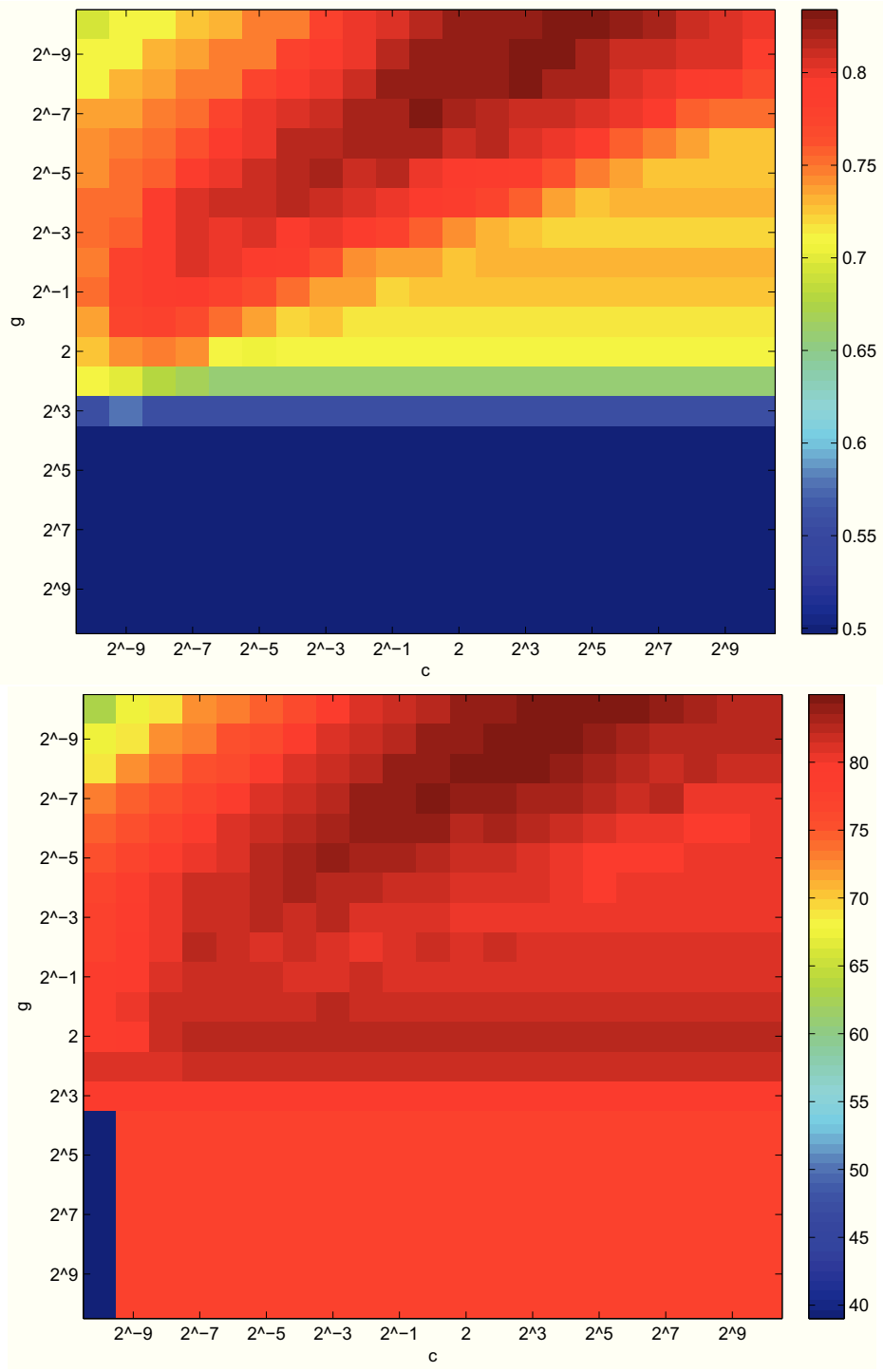


Figure 6.17: BSR and accuracy in grid search for parameters c and γ for Independent110. Note that the region of good parameters for BSR and accuracy coincide, and also coincide with the default parameter values.

Table 6.9: Results from 10-fold Cross Validation using Boosted Trees.

Performance Metric	Manual22	Inde110	Comb136	Comb121
Child Recall	0.657	0.718	0.671	0.676
Adult Recall	0.588	0.739	0.581	0.771
Accuracy	0.604	0.735	0.602	0.750
BSR	0.623	0.729	0.626	0.723
Age Sensitive Accuracy	0.649	0.800	0.632	0.796

6.3.2 Ensemble Classifiers

We used the ensemble classifiers available in Statistic Toolbox in Matlab. We used AdaBoost for the boosting algorithm and binary decision trees as the weak classifier, and built an ensemble classifier of 10 trees. To optimize for BSR rather than accuracy, we set the prior probabilities of both classes to be equal. Again, we used 10-fold cross-validation as the testing method. We tested several of the previous datasets and the results are shown in Table 6.9. Strangely, Combined136 was discovered to do very poorly in repeated tests. Therefore, we removed the low correlation ratios included from Manual22 (therefore only the 7 ratios with correlation greater than 0.3 are included) to create Combined121, which has comparable performance to Independent110. In general the results shown by the ensemble classifier were significantly worse than the results shown by the SVM. These are the best results we could obtain, after experimenting with more classifiers (up to 100), different boosting algorithms, and different weak classifiers. Without setting the prior probabilities, boosted trees are able to obtain results comparable to SVMs. It is likely that the worse results we are seeing are due to the prior probability settings. It seems that the penalty scheme does not work as well as SVM's (at least in the implementations we are using).

6.3.3 Testing on Other Datasets

We also did testing on the VADANA dataset. As we mentioned in Section 5.2, the distribution of the classes in this dataset is different from that of the FG-Net database,

Table 6.10: Results on the VADANA testset using SVM. The first test is using 10-fold cross-validation on the VADANA testset. The second test used the model trained using the FG-Net database, and tested on the VADANA dataset.

Performance Metric	CV	FG-Net
Child Recall	0.754	0.577
Adult Recall	0.812	0.852
Accuracy	0.808	0.830
BSR	0.783	0.714
Age Sensitive Accuracy	0.834	0.855

which will impact out results. First we did a 10-fold cross-validation on VADANA using the Independent110 ratios, which are the best performing set of ratios, and using weights according to the class distribution in VADANA. The results are comparable to the cross-validation results from FG-Net. For independent testing, we trained a model using the Independent110 ratios on all the images from FG-Net, then tested it on the VADANA dataset. The results are shown in Table 6.10. Although the child recall is low, the results are decent overall. One reason for the low recall of the child class using the FG-Net model, yet decent recall when doing cross-validation, is probably because the images of children in VADANA come from very few subjects. Therefore cross-validation can do well, but if one child looks like an adult to the FG-Net model, then it is likely for the FG-Net model to misclassify all the images of that child.

Additional, we also did testing on an internal dataset. We were unable to obtain permission to publish the images. This dataset contained 45 images of a single subject taken from ages 0 to 12. The images varied in pose, expression, and illumination. The purpose of this dataset was to observe our classifier changing the predictions made on images of the same subject, as the subject aged across images. We obtained very good results on this dataset using a SVM trained on FG-Net using Independent110, and the results are shown in Table 6.11. The trained SVM has never seen this subject before. The age sensitive accuracy is particularly high, because the misclassification

Table 6.11: Results on the internal dataset using SVM.

Performance Metric	Value
Child Recall	0.90
Adult Recall	0.76
Accuracy	0.82
BSR	0.83
Age Sensitive Accuracy	0.93

of adult as child only happened when the subject was close to age 5 (the subject was actually between the ages of 6 and 8 for all 6 misclassifications).

6.4 Comparison to Existing Methods

It is difficult to evaluate against other methods, for many reasons. There is the issue of regression vs. classification, class definitions for classification (what age ranges are grouped together), different datasets, and different testing methods. Since we use many more ratios and test more thoroughly, we can be confident that our system works better than other anthropometric methods mentioned in Section 4.1. Also, it is reasonable to assume that our method is better (in terms of our problem) than the regression methods that have mean absolute error of 5 years. Wang et al. achieved 90% accuracy on a combination of the FG-Net and MORPH databases. However, their definition for children is the ages 0 to 11, which seems to be a more natural boundary based on the growth patterns and sample ratios we have seen, so their slightly higher accuracy is not too surprising. We achieve 85% accuracy on our problem definition, which is still a comparable result.

We also used face.com for comparison. We used face.com [3] to make age predictions on the FG-Net database, then classified the images using their predicted age. Using this method, face.com achieves higher accuracy, but lower BSR. The results are shown in Table 6.12. The low recall on the child class means that face.com has

Table 6.12: Results from face.com.

Performance Metric	Value
Child Recall	96/213 (0.451)
Adult Recall	720/733 (0.982)
Accuracy	0.863
BSR	0.715
Age Sensitive Accuracy	0.889

problems finding the child faces, which is likely due to the error from their regression approach which we discussed.

CHAPTER SEVEN

Conclusion

We desire to build a system that is able to classify facial images into 2 classes, children (age ≤ 5), and adult (age > 5). The reason for defining child and adult is this manner is for the purpose of helping detect leukocoria, where the leukocoria is only significant if the subject in the image is age 5 or less. Age estimation is a well researched area and current algorithms perform well. However, these algorithms do not perfectly match our problem requirements. Regression approaches have mean absolute error that is too high since we deal with young children, and classification approaches have different class definitions. Therefore, there is a need to create a new approach that has high accuracy around the age of 5.

We know that distance ratios have been used with some success in the past, so we prefer to take this anthropometric approach to have a concrete understanding of the problem, rather than taking the more abstract appearance approaches. We introduced BSR as our performance metric of this problem to tackle the problem of imbalanced classes. We analyzed commonly chosen facial distance ratios, and devised a method to select better ratios. We also experimented with other more complex anthropometric features. Taking the best set of ratios, we optimized an SVM for our problem. The SVM performed better than boosted decision trees, and produced good results on an unseen testset. Our final system does very well on our defined problem, correctly classifying images 85% of the time. We analyzed our results in comparison to existing methods, and concluded that our method is comparable to existing methods.

One thing that we did not have the time to explore was to build a classifier that optimized age sensitive accuracy instead of BSR. This is a difficult problem as it requires modifying the penalty function of any learning algorithm. This is easy to do

on something simple such as a linear classifier; however, it will require more effort to tweak existing libraries of learning algorithms to include this modified penalty.

Another front to explore is to find more natural boundaries where humans can be classified by age, that is, finding the ages where it is easiest to separate humans. According to the ratio graphs that we have observed, ages such as 15 to 16 or around 21 exhibit large plateaus, indicating the growth of some portion of the face has stopped. It would be interesting to see exactly what features correspond to this change in growth at certain ages.

BIBLIOGRAPHY

[1] Axial skeleton. http://training.seer.cancer.gov/anatomy/skeletal/divisions/axial.html. Accessed: 29/05/2012.

[2] Bodyparts3d. http://lifesciencedb.jp/bp3d/. Accessed: 29/05/2012.

[3] face.com. http://face.com. Accessed: 29/05/2012.

[4] The fg-net aging database. http://www.fgnet.rsunit.com. Accessed: 29/05/2012.

[5] *Gray's Anatomy: The Anatomical Basis of Clinical Practice*. Churchill Livingstone, 0039 edition, November 2004.

[6] *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG 2006), 10-12 April 2006, Southampton, UK*. IEEE Computer Society, 2006.

[7] Aubin Balmer and Francis Munier. Differential diagnosis of leukocoria and strabismus, first presenting signs of retinoblastoma. *Clinical Ophthalmology*, 1(4):431–439, dec 2007.

[8] Wen bing Horng, Cheng ping Lee, and Chun wen Chen. Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering*, 4(3):183–192, 2001.

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

[10] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In Hans Burkhardt and Bernd Neumann, editors, *European Conference on Computer Vision (2)*, volume 1407 of *Lecture Notes in Computer Science*, pages 484–498. Springer, 1998.

[11] D. Cristinacce and T. Cootes. Facial feature detection using adaboost with shape constraints. In 14^{th} *British Machine Vision Conference, Norwich, England*, pages 231–240, 2003.

[12] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[13] Yun Fu, Guodong Guo, and Thomas S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

[14] Yun Fu and Thomas S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.

[15] Xin Geng, Zhi hua Zhou, and Kate Smith-miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:2234–2240, 2007.

[16] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. *International Workshop on Visual Observation of Deictic Gestures*.

[17] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S. Huang. Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition*, pages 112–119. IEEE, 2009.

[18] Erik Hjelmas and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–275, 2001.

[19] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A practical guide to support vector classification. *Bioinformatics*, 1(1):1–16, 2010.

[20] Kin-Man Lam Kwok-Wai Wong and Wan-Chi Siu. An efficient algorithm for human face detection and facial feature extraction under different conditions. *Pattern Recognition*, 34:1993–2004, 2001.

[21] Young H. Kwon and Niels Da Vitoria Lobo. Age classification from facial images. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 762–767, 1999.

[22] Ries LAG, Smith MA, Gurney JG, Linet M, Tamra T, Young JL, and Bunin GR (eds). Cancer incidence and survival among children and adolescents: United states seer program 1975-1995. (99–4649), 1999.

[23] Andreas Lanitis, Christopher J. Taylor, and Timothy F. Cootes. Automatic face identification system using flexible appearance models. *Image Vision Comput.*, 13(5):393–401, 1995.

[24] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *European Conference on Computer Vision*, 2008.

[25] Louise Scheuer and Sue Black. *Developmental Juvenile Osteology*. Academic Press, 1st edition, August 2000.

[26] Caifeng Shan. Learning local features for age estimation on real-life faces. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, MPVA '10, pages 23–28, New York, NY, USA, 2010. ACM.

[27] Gowri Somanath, M. V. Rohith, and Chandra Kambhamettu. Vadana: A dense dataset for facial image analysis. In *ICCV Workshops*, pages 2175–2182. IEEE, 2011.

[28] Jian-Gang Wang, Wei-Yun Yau, and Hee Lin Wang. Age categorization via ecoc with fused gabor and lbp features. In *Workshop on the Applications of Computer Vision*, pages 1–6. IEEE Computer Society, 2009.

[29] Phillip Ian Wilson and John Fernandez. Facial feature detection using haar classifiers. *J. Comput. Small Coll.*, 21:127–133, April 2006.

[30] Chen Xing. asmlib-opencv. http://code.google.com/p/asmlib-opencv/. Accessed: 29/05/2012.

[31] Chenyang Xu and Jerry L. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing*, 7(3):359–369, 1998.

[32] Zhiguang Yang and Haizhou Ai. Demographic classification with local binary patterns. In Seong-Whan Lee and Stan Z. Li, editors, *International Conference on Biometrics*, volume 4642 of *Lecture Notes in Computer Science*, pages 464–473. Springer, 2007.

[33] Joseph W. Young and United States. *Selected facial measurements of children for oxygen-mask design*. Federal Aviation Agency, Office of Aviation Medicine, Washington, D.C., 1966.