KWAME NKRUMAH UNIVERSITY OF SCIENCE AND
TECHNOLOGY, KUMASI

COLLEGE OF SCIENCE

DEPARTMENT OF MATHEMATICS

FINGERSPELLING GESTURE RECOGNITION USING PRINCIPAL COMPONENTS ANALYSIS

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS THROUGH THE NATIONAL INSTITUTE FOR MATHEMATICAL SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD
OF
MASTER OF PHILOSOPHY DEGREE
(SCIENTIFIC COMPUTING & INDUSTRIAL MODELING)

BY
RUDOLPH ELIKEM KLU
JUNE, 2017

Declaration

I hereby declare that, this thesis is the result of my own original research and that no part of it has been submitted to any institution or organization anywhere for the award of a degree. All inclusion for the work of others has been dully acknowledged.

Rudolph Elikem, Klu	.....	.....
Student (PG3325414)	Signature	Date

Certified by:

Dr. Peter Amoako-Yirenkyi	.....	.....
Member, Supervisory Committee	Signature	Date
Dr. Akoto Yaw Omari-Sasu	.....	.....
Member, Supervisory Committee	Signature	Date
Dr. Peter Romeo Nyarko	.....	.....
Member, Supervisory Committee	Signature	Date
Dr. R. K. Avuglah	.....	.....
Head of Department	Signature	Date

Abstract

Sign language is one of the most natural and raw forms of language and communication. which could be dated back to as early as the advent of the human civilization, when the first theories of sign languages appeared in history. This thesis presents an approach to recognize hand spelling gestures to aid the deaf/mute communicate with non-signers. Images were captured using a computer camera. Skin/hand areas were segmented out of the images using color, rotated onto their principal axis by the method of moments and transformed into a PCA feature space for gesture feature identification and characterization. Data was trained in this system and subsequently test data was classified using a single space euclidean classifier and a single space Mahalanobis classifier which utilized the Euclidean and Mahalanobis distances respectively. The two classifiers are compared for their accuracy and efficiency. The results of the work indicated that the single space Mahalanobis Classifier performed better than the single space euclidean classifier especially as the number of principal components increased. The number of principal components selected also greatly affected the accuracy of classification, with more principal components introducing noise in the images.

Acknowledgements

Glory be to the Almighty God who made this work possible.

It is with great pleasure that, my supervisor, Dr. Peter Amoako-Yirenkyi is also acknowledged, for his immense contribution, advice and assistance.

I would also like to acknowledge my parents and siblings for their prayers and moral support, to my friends, course mates and the countless others who in one way or the other aided in making this project a success. May the Almighty God richly bless you all.

Dedication

This thesis is dedicated to my father, Mr. Rudolph Klu, mother, Mrs. Patricia Klu, my siblings, Yaa Klu, Aku Brocke, Adjoa Klu, Emmanuel Klu, Ostwin and JNOKO

Contents

List of Figures

Abbreviations/Acronyms

HMM	Hidden Markov Models
SLR	Sign Language Recognition
ASL	American Sign Language
GSL	Ghanaian Sign Language
AdaSL	Adamorobe Sign Language
PCA	Principal Components Analysis
PC	Principal Components

Chapter 1
INTRODUCTION

1.1 Background

Human gestures constitute a space of motion expressed by the body, face, and/or hands. Among a variety of gestures, hand gesture is the most expressive and the most frequently used. A gesture is defined as a movement of part of the body, especially a hand or the head, to express an idea or meaning[The Oxford Dictionary]. It can also be said to be an action performed to convey a feeling or intention. As humans, communication is done with our voices. Other parts of the body such as our limbs and face are used to make various gestures. In a few decades, many attempts have been made to create systems with the concept of computer vision to understand and interpret gestures. Sign languages in the form of hand gesture is the main form of communication among the deaf and the hearing-impaired. This perhaps can be attributed to the fact that, There are special rules of context and grammars that support each expression.

Gesture based communication is a standout amongst the most normal and crude types of dialect and correspondence. Gesture based communication goes back to the early approach of the human development; from the fifth century BC,when the

principal composed records of communication through signing are recorded ever. It has utilized even before talked dialect/correspondence developed. From that point forward, the gesture based communication has developed and been received as a basic piece of our everyday correspondence forms. Presently, communications through signing are being utilized widely in worldwide sign language for the deaf and dumb, in the realm of games, for religious practices and furthermore at work places [Rockett (2003)]. Gestures are one of the main types of correspondence when a baby figures out how to express its requirement for nourishment, warmth and solace. It improves the accentuation of talked dialects and aides in communicating considerations and sentiments successfully. A basic signal with one hand has a similar importance everywhere throughout the world and means either "greetings" or 'farewell'. Many individuals go to remote nations without knowing the official dialect of the went by nation and still figure out how to perform correspondence utilizing motions and communication through signing. These cases demonstrate that gestures can be viewed as worldwide and utilized everywhere throughout the world. In various occupations around the globe, gestures are method for correspondence [Gupta and Ma (2001)]. In the air terminal for instance, a predefined set of signals make individuals on the ground ready to speak with the pilots and in this manner offer bearings to the pilots of how to get on and off the run-way and the arbitrator in any game uses gestures to convey his choices. Hearing disabled individuals have throughout the years built up a gestural dialect where every single characterized signal have an alloted meaning. The dialect permits them to speak with each other and the world they live in.

Sign comprises of three fundamental parts: Manual components including signals made with the hands (utilizing hand shape and movement to pass on significance), Non-manual elements, for example, facial expressions or body pose, which can both frame some portion of a sign or change the meaning of a manual sign, and Finger spelling, where words are spelled out by motions in the local verbal dialect. Actually

this is an oversimplification, Sign language is as intricate as any talked dialect, each communication through signing has a huge number of signs, each varying from the following by minor changes in the shape of the hand, movement, position, non-manual features or setting. Since signed languages advanced alongside talked dialects, they don't mimic each other.

Ghanaian Sign Language(GSL) is the national gesture based communication of deaf individuals in Ghana, slipped from American Sign Language. It was presented in 1957 by Andrew Foster, a deaf African-American teacher, as there had been no instruction or associations for the deaf already. Ghanaian Sign Language(GSL) is not related to indigenous Ghanaian gesture based communications, for example, Adamorobe Sign Language Adamorobe Sign Language (AdaSL) is a town gesture based communication utilized as a part of Adamorobe, an Akan town in eastern Ghana. It is utilized by around 1370 deaf individuals (2003).

The Adamorobe people group is prominent for its uncommonly high occurrence of innate deafness (hereditary passive autosome). As of (2012) around 1.1% of the aggregate populace is deaf, yet the rate was as high as 11% in 1961 preceding the local chief forbade deaf individuals from wedding equally deaf. Deaf individuals are completely incorporated into the town's life.

Under these conditions, AdaSL has created as an indigenous gesture based communication, completely free from the nation's standard Ghanaian Sign Language(GSL) ;which is identified with American Sign Language.

The recognition of signed gestures to words and sentences as they do in American Sign Language without a doubt speaks to the most troublesome recognition issue of those applications spoken of earlier. A working gesture based communication framework could give a chance to the deaf to speak with non-signing individuals without the requirement for a mediator. It could be utilized to create discourse or content making the deaf more independent.

While the automation of spoken speech recognition has now progressed to the point of being financially accessible, programmed Sign Language Recognition is still in its earliest stages. Right now most commercially available interpretation services are human based, and consequently costly, because of the experienced people required. There are different strategies for communication via gestures recognition. Some utilize electronic glove while other utilize vision based approach.
Be that as it may, as the electronic glove approach is the more costly one, vision based approach is generally utilized.

1.2 Motivation

The recognition of gestures to words and sentences as they do in Ghanaian/American Sign Language without a doubt speaks to the most troublesome recognition issue of those applications said above. Advancement of a framework that perceives straightforward hand spelling gestures/signs will ease correspondence between the deaf/mute and non-signers.

1.3 Problem Statement

Deaf and mute individuals have found it difficult communicating with non-signers over the years since the development of sign language in the country. In public institutions they are found unable to communicate their needs. It is expensive to train individuals in sign language to cater for their needs in public places like libraries and banks and equally expensive for the deaf/mute to have a dedicated interpreter when they need to communicate with non-signers.

1.4 Objective

The objective of this study are:

1. to develop a functioning system that can recognize finger spelling gestures
2. to test the performance of euclidean classifier and the mahalanobis classifier both in single space.

1.5 Outline of the Methodology

The following outlines the method employed in achieving the objective of the thesis:

1. Detection and extraction of skin regions in acquired images
2. Image vectors are constructed and reduced into a PCA feature space
3. Data is trained with training and test data is classified using the euclidean and mahalanobis classifiers

1.6 Justification of the Study

A working gesture based communication framework could give a chance to the deaf to speak with non-signing individuals without the requirement for a mediator. It could be utilized to produce audio discourse or text making the deaf more independent.

1.7 Organization of Chapters

Chapter 1 introduces the concepts of gestures, sign language and sign language recognition. It emphasizes the problem statements and states clearly the objectives of this study.

Chapter 2 discusses previous works related to this study

Chapter 3 explicitly discusses the methodology of this work. It explains image processing techniques used in this study; color extraction, hand region segmentation,feature extraction and classification using Principal Components Analysis.

Chapter 4 displays the results of the process of image analysis and feature classification discussed in Chapter 3.

Chapter 5 Concludes the study and provides recommendations for future references.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This discussion of related work concentrates on past work in gesture based communication recognition. Many early sign language recognition systems used markers or marked gloves to elude the problem of complex backgrounds. Glove-based signal systems however require the client to wear a bulky gadget, and by and large convey a heap of cables that hook up the gadget to a PC. This is also known as the sensor based detection These systems directly acquired specific data of the hands using data gloves and accelerometers. In [Waldron and Kim (1995)] used the Polhemus tracker to collect data from the data gloves and [Vogler and Metaxas (1999)] used the DataGlove sensonr to do such in. Such utilization of sensors implied the component extraction process was avoided since estimations, for example, hand as well as arm joint edges and spatial position acquired from the sensors were adequate to be utilized as features as was done by [Gao et al. (2000)] The use of gloves usually altered the signs performed due to the load of cables interfacing the gloves to a computer. [Dipietro et al. (2008)] given an extensive study of glove-based frameworks. Thirty sorts of gloves were talked about, illustrating their attributes and applications. These

incorporated the SayreGlove, MIT LED Glove, Digital Entry DataGlove, CyberGlove, and the Power-Glove, among others. The writers inferred that the DataGlove and the CyberGlove were the most normally utilized gloves for communication via gestures recognition.

Data gloves/digital gloves have been utilized broadly in earlier research. [Kadous (1996)] revealed a framework utilizing gloves to perceive an arrangement of disconnected Australian signing gestures with 80% precision.

[Lee and Xu (1996)] built up a glove-based gesture recognition framework that could perceive 14 of the characters from the handspelling gestures in the right sequence, adapt different signals and be ready to amend the template of every motion in the framework in online mode, with at a frequency of 10Hz. Consistently sophisticated gloves and gadgets have undergone development, for example, the "Sayre Glove", "Dexterous Hand Master" and "PowerGlove" [Watson (1993)]. By a long shot the VPL DataGlove is the most successful commercially available glove.

By and large most completely intuitive hand signal frameworks can be considered to include three essential levels: the detection level, tracking and recognition level. The tracking layer characterizes and removes visual components which could be ascribed to the existence of hands in the field of perspective of the catching camera. The association of data between successively captured frames is performed by the tracking layer of the framework. This records movement of the hands through progressive frames of images. Finally, the recognition layer is in charge of collection the information removed in the preceding layers, gathering and assigning names that will label them to specific classes of motions/gestures.

2.2 Detection

The essential step in gesture frameworks is the recognition of hands, detection and removal of the regions of interest. This selection of the region of interest(segmentation) is an extremely important step since it separates the information relevant to the recognition from the picture background,before it is passed to subsequent tracking level and then to a recognition phase. Countless strategies suggested in the writing use a few sorts of visual elements and, as a rule, a blend of them. These elements include the color of skin pigment, outline, movement and objective figures of the palm. [Cote et al. (2006)], discusses various hand segmentation and does a comparison of their performance.

2.2.1 Color

The segmentation of the color of the skin has been used in some methodologies for locating the hand. A noteworthy choice in the thinking of a model for the colour of the skin is the choice of the color space to be utilized. There are some proposed color spaces which include "RGB", "standardized RGB", "HSV", "YCrCb", "YUV", and so forth. "Luminance" (intensity or value) and chrominance (hue data) are the two components of a color signal. The favored color spaces are the ones that proficiently isolate the "chromaticity" and "luminance" information of the color signal. This is because of the way that by utilizing color spaces that do not have "chromaticity-luminance" independence parts of the color signal only, a level of invariability to brightening conditions could be accomplished. An analysis of various models of skin "chromaticity" was done in Terrillon et al. (2000).Their performances were further assessed

To greatly improve the robustness against variations in brightness a few techniques Martin and Crowley (1997), executed their work using the HSV space[Saxe and

Foulds (1996)], YCrCb [Chai and Ngan (1998)], YUV [Yang et al. (1998), Argyros and Lourakis (2004b)], or normalized RGB used by [Soriano,2003] colorspaces, so as to estimate the "chromaticity" of skin (or, generally, how much light in absorbs) instead of the skin's obvious chromatic value. These methods commonly do away with the "luminance" part, to reduce the effects of shadows, brightness and illumination changes, and in addition regulations of the direction of the skin surface in respect to the source(s) of light.

Various methods [Saxe and Foulds (1996), Kjeldsen and Kender (1996)] use prefigured color distributions extricated from various large data sets that have been analyzed statistically . For instance, in [Jones and Rehg (2002)], a statistical skin colour model was gotten from the examination of several photographs from the internet. Interestingly, techniques such as those portrayed in [Zhu et al. (2000)] construct a model of skin colour in view of sampled data of skin when the framework is being initialized.

The apparent shade of human skin has lots of variations over races and also among people of a similar race. Extra fluctuation might be presented because of varying light conditions as well as camera attributes. Accordingly, hue based ways to deal with hand recognition need to utilize a few means for adjusting for this changeability. In [Yang and Ahuja (1998), Sigal et al. (2004)], an invariant portrayal of the color of the skin against changes in light is sought after, yet at the same time with no definitive outcomes.

More elaborate systems depend on histograms coordinating, or utilize a basic look-into table method [Kjeldsen and Kender (1996), Yang and Ahuja (1998)] in light of the preparation information for the skin and conceivably its encompassing regions.

By and large, color extraction can be confounded by the objects in the background with similar properties to the human skin with respect to colour. An approach to adapt to this issue depends on removing the objects that are in the background.

Nonetheless, removal of background objects is ordinarily in view of the suspicion that the camera framework is fixed relative to the background objects. To take care of this issue, some exploration [Utsumi and Ohya (1998), Blake et al. (1999)], has investigated the dynamic redress of background models.

The color of the skin is just a single of many signals to be utilized for to hand identification. For instance, in situations where the countenances likewise show up in the camera field of view, further preparing is needed to recognize hands from the human face. In this way, the color of the skin has been used in mix with different signs to acquire better execution. In [Yuan et al. (1995)] skin location is consolidated with non-inflexible movement identification and in [Derpanis et al. (2004)] the color of the skin was utilized to limit the locale where movement components are to be followed. An imperative research bearing is, subsequently, the mix different prompts.

2.2.2 Shape

The trademark outline of human hands has been used in its recognition in pictures in different ways. Much data can be acquired by simply extracting the shapes of articles in the picture. In the event that effectively distinguished, the contours speaks to the state of the palm and is in this way not specifically reliant on perspective, the hue of the skin and brightness.

In the general case, extracted shape that depends on the detection of edges brings about a substantial number of edges that have a place with the hands additionally to immaterial foundation objects. Thus, modern post-handling methodologies are needed in building the unwavering quality of a method like this. In light of this , edges are regularly joined with (skin)color and the removal of backgrounds.

There are two types of gesture modeling. [Garg et al. (2009)] classifies gesture modeling into "spatial" and "temporal modeling". In spatial modeling the feature of position or gesture shape in the scope of HCI applications are considered [Garg

et al. (2009)], whereas temporal modeling relates the constantly changing features of the hand gesture (in reference to the gestures motion) in the HCI environments [Garg et al. (2009)]. Hand Modeling in spatial domain can be executed in either a 2D or 3D space [Pavlovic et al. (2000)].

In the works of [Krueger (1991), Krueger (1993), Utsumi and Ohya (1998)], the 2D/3D drawing systems of the users hands are extracted as a shape using the assumption of background homogeneity and executing edge detection on the image. Cases of the countours/shapes are utilized as elements can be seen in template and also in vision based techniques.

Certain strategies concentrate on the particular morphology of hands and endeavor to identify them in light of trademark hand shape elements, for example, the tips of fingers. The methods [Argyros and Lourakis (2006)] implemented use contour to signal fingertip position. A different procedure that was utilized in fingertip location is "template matching". These templates or models could be pictures of the tips of fingers or fingers [Rehg and Kanade (1995)]. Using extra image features like contours,such approches that match patterns could improve. [Rehg and Kanade (1994)]. Aside the high cost of its computation, template unable to adapt to neither scaling nor pivot of the objective of interest.That issue is tended to in [Crowley et al. (1995)] by a continuous update of the template. To the discovery of fingertips, numerous other aberrant methodologies to have been utilized, similar to image analysis utilizing special "Gabor kernels" [Meyering and Ritter (1992)]. The primary hindrance in the utilization of fingertips as components is the nuisance of they being blocked by whatever is left of the hand. An answer for this impediment issue includes the utilization of numerous cameras [Lee and Kunii (1995), Rehg and Kanade (1994)].

2.2.3 Detectors that learn from pixel values

Critical work has been done on discovering hands in grey scale images in view of their appearance and surface. In [Wu and Huang (2000)], the appropriateness of various classification strategies with the end goal of view-autonomous hand posture recognition was explored. Several methods [Triesch and Malsburg (1996), Triesch and Von der Malsburg (1998)] endeavour to distinguish hands in light of hand appearances, via testing classifiers with set of image samples. The essential supposition is that hand appearance contrasts more among hand motions than it varies among various individuals playing out a similar motion. Still, programmed include determination constitutes a noteworthy trouble. Several publications consider the issue of feature extraction and selection [Triesch and Malsburg (1996), Quek and Zhao (1996), Nolker and Ritterpages (1998)], with quite a handfull of results with regards to hand detection. All the more as of late, techniques in light of a machine learning method know as boosting have shown exceptionally powerful outcomes in face and hand location.

Boosting is a general strategy that can be utilized for enhancing the precision of a given learning algorithm [Schapire (2002)]. It depends on the rule that an exceedingly exact or classifier can be determined by the linear combination of numerous moderately non-exact or "weak" classifiers. By and large, a single weak classifier is required to execute just somewhat superior to arbitrary.

The AdaBoost algorithm gives a learning strategy to getting reasonable collection of weak classifiers. A collection of images is utilized for training by this technique that comprises of positive and negative illustrations (hands and non-hands, for this situation), which are related with its respective labels. Weak classifiers are included successively into a selection of weak classifiers keeping in mind the end goal to diminish the upper bound of the training error.

2.2.4 3D model-based detection

A classification of methodologies use 3D hand models for the identification of hands in images. One of the upsides of these techniques is that they can accomplish view-autonomous detection. The utilized 3D models ought to have enough degrees of freedom.

Hand postures are then assessed given that the correspondences between the 3D model and the observed image features are entrenched. Different 3D hand models have been proposed in the writing. In [Rehg and Kanade (1994), Stenger et al. (2002)], a full hand model is proposed which has 27 degrees of freedom (DOF) (6 DOFfor 3D location/orientation and 21 DOF for articulation). In [Goncalves et al. (1995)], a 3D model of the arm with 7 parameters is used.[Gavrila and Davis (1996)] proposes a 3D model with 22 degrees of freedom for the whole body with 4 degrees of freedom for each arm. In [MacCormick and Isard (2000)], the user's hand is modelled much more simply, as an articulated rigid object with three joints comprised by the first index finger and thumb.

2.2.5 Motion

The movement of the hand is a prompt used by a couple ways to deal with hand discovery. The logic being that movement based hand recognition requests for an exceptionally controlled setup, since it expect that the main movement in the picture is because of hand development. In reality, early works (for example,Cui and Weng (1996)]) accepted that the movement of the hand is the main movement happening within the frame of the image. In later methodologies, movement data is joined with extra visual signs. On account of fixed cameras, the issue of movement approximation lessens to background upkeep and resulting segmentations. For instance in [Martin et al. (1998)] that data was used in recognizing hands from other skin-hued protests

and adapt to illumination conditions forced by shaded lights. The distinction in "luminance" of pixels from two progressive pictures is near zero for background pixels. By picking and keeping up a proper edge, moving articles are distinguished inside a static frame.

2.3 Tracking

Tracking, or the frame-to-frame correspondence of the fragmented hand areas or components, is the second step in the process towards understanding the observed hand motion. The significance of robust tracking is twofold. To start with, it gives the inter-frame connecting of hand/finger appearances, which allows the rise to trajectories of features with time.

2.3.1 Template Based Tracking

This class of strategies shows incredible closeness to techniques for hand discovery. Individuals from this class conjure the hand identifier at the spatial region that the hand was distinguished in the past edge, to definitely limit the picture look space. The understood supposition for this strategy to succeed is that pictures are gained as often as sufficiently possible.

Correlation-based element following is straightforwardly gotten from the above approach. In [Crowley et al. (1995), OHagan and Zelinsky (1997)] connection based format coordinating is used to track hand includes crosswise over edges. Once the hand(s) have been distinguished in an edge, the picture districts in which they show up is used as the model to recognize the turn in the following casing. The presumption is that hands will show up in the same spatial neighborhood. The work in [Hager and Belhumeur (1996)] bargains likewise with variable light. An objective is seen under different lighting conditions. At that point, an arrangement of premise pictures that can be utilized to rough the presence of the question saw under different

brightening conditions is built. Following at the same time tackles for the relative movement of the question and the enlightenment.

Some methodologies distinguish hands as picture blobs in each casing and transiently relate blobs that happen in proximate areas crosswise over edges. Approaches that use this sort of blob following are for the most part the ones that recognize hands in view of skin shading, the blob being the correspondingly sectioned picture locale (e.g. [Birk et al. (1997), Argyros and Lourakis (2004b)]). Blob-based methodologies can hold following of hands notwithstanding when there are extraordinary varieties from casing to outline.

Augmenting the above approach, deformable shapes, or "snakes" have been used to track hand areas in progressive picture outlines [Cootes and Taylor (1992)]. Normally, the limit of this area is dictated by force or shading angle. All things considered, different sorts of picture components (e.g. surface) can be considered. The procedure is introduced by putting a form close to the area of intrigue. Snakes take into account ongoing following and can deal with numerous objectives as well as mind boggling hand stances. They show better execution when there is adequate difference between the foundation and the protest [Cootes et al. (1995)]. On the opposite, their execution is bargained in jumbled foundations. Following neighborhood hand highlights on the hand has been utilized in particular settings, presumably in light of the fact that following nearby components does not ensure the segmentation of the hands from whatever remains of the picture. The techniques in [Martin et al. (1998),Baumberg and Hogg (1994)], track turns in picture successions by consolidating two movement estimation forms, both in view of picture differencing. The principal procedure figures contrasts between progressive pictures. The second processes contrasts from a foundation picture that was already obtained. The reason for this blend is expanded invariability close to shadows.

2.3.2 Optimal Estimation Techniques

The tracking of features have been widely researched in PC vision. In this unique situation, the ideal estimation system gave by the Kalman filter [Kalman (1960)] has been generally utilized in turning perceptions (feature dection) into estimations (removed direction). The explanations behind its ubiquity are ongoing execution, treatment of instability, and the prediction of expectations for the progressive frames.The Kalman filter and hand blobs examination is utilized for hand tracking to get movement descriptors and hand locale. It uses skin color for gesture tracking in hand and is quite robust to background layers. In [Argyros and Lourakis (2004b)], the objective is held against situations where hands impede each other, or show up as one blob in the picture, in light of a theory definition and approval/dismissal scheme. The issue of different blob tracking was examined in [Argyros and Lourakis (2004a)], where blob tracking is performed in both pictures of a stereo match and blobs are related, across both cameras and frames The orientation of the client's hands is consistently evaluated with the Kalman filter to limit the point in space that the client demonstrates by expanding the arm and indicating with the forefinger in [Kohler (1997)]. In [Utsumi and Ohya (1999)], hands are followed from various cameras, with a Kalman filter in each picture, to evaluate the 3D hand gesture. Snakes incorporated with the Kalman separating structure (see underneath) have been utilized for following hands in images [Terzopoulos and Szeliski (1992)]. Treating the tracking of picture elements inside a Bayesian system has been for quite some time known to give enhanced estimation results. In [Bregler (1997)], a strategy is proposed for following human movement by gathering pixels into blobs in view of intelligent movement, shading and fleeting bolster utilizing a desire boost (EM) calculation. Each blob is in this way followed utilizing a Kalman channel. At long last, in [MacCormick and Blake (199), MacCormick and Isard (2000)], the shapes of blobs are followed

crosswise over casings by a blend of the Iterative Closed Point (ICP) calculation and a factorization technique to decide worldwide hand posture. In [Utsumi and Ohya (1999)], the 3D positions and stances of both hands are followed utilizing different cameras. Each hand position is followed with a "Kalman filter" and 3D hand stances are evaluated utilizing picture highlights. This work manages the shared hand-to-hand impediment inborn in following both hands, by choosing the perspectives in which there are no such impediments.

2.4 Recognition

The general objective of hand gesture recognition is the understanding of the semantics that the hand(s) area, stance, or motion passes on. Normally, the bigger the vocabulary, the harder the gesture recognition undertaking gets to be. An early framework that executed recognition was [Birk et al. (1997)]. It perceived 25 gestures from the International Hand Alphabet. The recognition of gestures is of subject of awesome enthusiasm all alone, as a result of gesture based correspondence. In addition, it additionally shapes the premise of various gesture recognition strategies that regard gesture as a progression of hand stances. Other than the acknowledgement of hand stances from pictures, gesture recognition incorporates an extra level of multifaceted nature, including segmentation, of the ceaseless signal into constituent components. In a various assortment of strategies, the transient cases at which hands speed is limited are viewed as gestures, while video outlines that depict a hand movement are some of the time ignored.

2.4.1 Template Matching

Template matching, central to pattern recognition methods, have been used with regards to both gait and gesture recognition. With regards to pictures, Template matching is achieved through the pixel-wise examination of a template and an image

under processing. The similitude of the image of interest to the template is corresponding to the aggregate score using a predetermined scale. For the recognition of hand gestures, the picture of an identified hand shapes the hopeful picture which is specifically contrasted against template images of hand gesture. The best template that matches the image (assuming any) is considered as the matching gesture. Obviously, in view of the pixel-by-pixel picture correlation, template matching is not robust to changes in orientation and size of image. It's one of the main techniques utilized to identify hands in pictures. To adapt to the inconstancy because of changes in orientation and size, a few creators have proposed startegies to normalize these two factors[Birk et al. (1997)],others too prepare the arrangement of templates with pictures from different perspectives.

2.4.2 Methods based on Principal Component Analysis

PCA strategies require an underlying preparing stage, in which an set of pictures of comparable substance is handled. Regularly, the intensity estimations of each picture are considered as estimations of a 1D vector, which has dimensions that are equivalent to the pixels quantity in the picture; it is accepted, or upheld, that all pictures equivalent in dimension.

For sets such as this, there were basis vectors developed to surmised any of the (training) pictures in the set. The preceding procedure is executed for every stance in the database of gestures, that the framework ought later have the capacity to perceive. In PCA-based signal detection, the coordinating mix of central segments shows the coordinating motion too.

This is on account of the coordinating blend is one of the agents of the set of motions that were bunched together in preparing, as expressions of a similar signal.

PCA was initially connected to gesture identification in [Sirovich and Kirby (1987)]and was stretched out in [Murase and Nayar (1995)]. A basic framework is displayed

where the entire picture of a man signalling is analysed, expecting that the principle part of movement is the gesture.

2.5 Complete Gesture Recognition Systems

As far as open motions, the communication via gestures for the hearing impeded has gotten huge consideration [Starner and Pentland (1995), Cui et al. (1995), Waldron (1995)].Other than giving a rather constrained and important dataset, it displays critical potential effect in the public eye because it encourages the correspondence of the hearing disabled with systems by a somewhat basic and natural mode for the signer. In [Imagawa et al. (1998)], a bidirectional interpretation framework between Japanese Sign Language and Japanese was actualized, so as to help the hearing weakened speak with typical talking individuals through gesture based communication. Among the most early frameworks is the one in [Starner and Pentland (1995)] which perceives around forty sign language from the american dictionary which was subsequently stretched out to watch the signer's gestures from a device fixed on a top hat worn by the signer. Other than the recognition of individual hand acts, the framework in [Martinez et al. (2002)] perceived movement primitives and full sentences, representing the way that a similar sign may have distinctive implications relying upon setting. The fundamental distinction of the framework in [Yang et al. (2002)] is that it extricates movement directions from a picture sequence and utilizations these directions as elements in gesture recognition in mix with perceived hand stances.

Chapter 3

METHODOLOGY

3.1 Introduction

The recognition of gestures based on image based techniques is not new in the area of image analysis. However this research seeks to analyse the detection of 6 finger spelling gestures using image based techniques and subsequently compare the accuracy of two classifiers(the single space "Euclidean" and "Mahalanobis" classifiers) in the detection of these 6 finger spelling gestures.

3.2 Segmentation

The first step of a hand gesture recognition process is hand detection from the background. The following steps of recognition process strongly rely on the influence of segmentation, hand segmentation is the key important step in gesture recognition process. Segmentation is the process of partitioning an image into multiple segments. This is done to simplify the image for easier analysis. Suppose we want to extract the important features within an image, the hand gesture in this case the image is processed to isolate or segment the hand gesture from the background of the image.

Edges are a usually a good signal for segmentation but color is used in this study for better robustness even with noisy images

3.2.1 Skin Detection(Colour)

Colour is typically the principal characteristic looked for finding the hands of the signer in a video frame since the skin is more distinct.

After finding skin pixel applicants, non-skin blobs can be removed by utilizing texture, shape or motion prompts.

Skin candidate pixels may then be utilized to perceive the hand with better quality while the background and other non-essential parts of the frame are removed. Segmentation and detection of hand reduces computation time while the accuracy of recognition of gestures in sign language and gesture recognition frameworks is increased.

Colour being a low-level feature makes it computationally less expensive to process. There exist some disadvantages nonetheless: the presence of color is delicate to light changes (in its brightness and "chromaticity"), camera alignment and shadows. A very significant challenge encountered with vision based methodologies is the segmentation of face and hand from non-uniform background with varying lighting conditions

A color signal has two constituents; the luminance which carries light intensity data or values and chrominance which is the color information data. It is frequently of much advantage chrominance values independent of the luminance values in a 2D space known as the "chromaticity" space. The skin distribution in "chromaticity" space has demonstrated invariability to changing brightening situations.

The skin segmentation technique's performance is greatly affected by the color space that is used for the skin color segmentation.

A few skin pictures were taken under different lighting/luminance conditions to build

up a skin locus. This is done to build up an offline skin model

Three color spaces are considered for skin color extraction in this study, specifically standardized or normalized RGB, HSV and YCbCr.

3.2.2 RGB and Normalized RGB Color Space

The sensors in human visual framework are coarsely split into three essential bands, "Red", "Green" and "Blue". So "RGB" color framework has been produced utilizing the three hues as the principal hues whiles alternate hues are depicted as the blend of the major hues. In this manner, hues are viewed as mixes of the alleged primary hues "red(R)", "green(G)" and "blue(B)".

Electronic displays and a few cameras showcase pixels as a triple $R, G, B \in \mathbb{R}^3$ of intensity values in "red", "green" and "blue", individually, in the "**RGB**" color space.

The "**RGB**" channels are however exceptionally corresponded: they all incorporate a property of brightness. The dependence of the luminance and "chrominance" data make the **RGB** color space somewhat less alluring. The conceivable number of hues that can be characterized by utilizing "RGB" shading space is

$$N = (2^{(3)})^8 = 16,777,216 \tag{3.2.1}$$

which is very adequate to show every single color that can be distinguished by human eye. $[R; G; B] = [0; 0; 0]$ conforms to black pixels and the $[R; G; B] = [255; 255; 255]$ is white. Between these two bounds lie all the other colours.

To get a linear independence between its "chrominance" and luminance information, the "**RGB**" color space is normalized. To normalize the "RGB" color space,

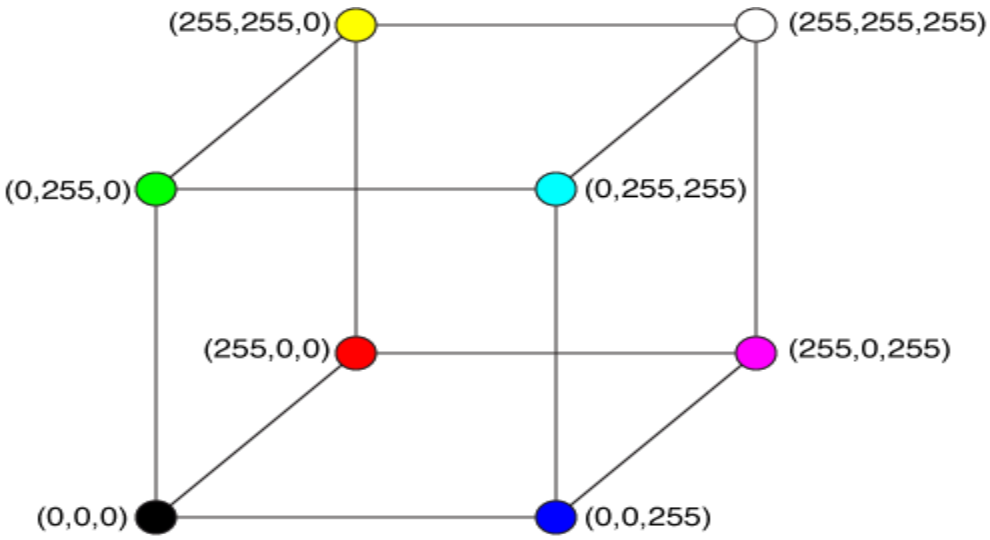

Figure 3.1: The RGB color space-RGB Cube

preceding equations are used

$$r = \frac{R}{R+G+B} \tag{3.2.2a}$$

$$g = \frac{G}{R+G+B} \tag{3.2.2b}$$

$$b = \frac{B}{R+G+B} \tag{3.2.2c}$$

Pure blue(b) is made redundant after the normalisation because $r + g + b = 1$. Figure 3.6 shows a histogram of the **RGB** color space showing the individual of the Red, Green and Blue colors for an .

3.2.3 YCbCr Color Space

The color space where we have "chrominance" and "luminance" information independence is the YCbCr color space. **Y** is the "luminance" and **Cb**, **Cr** are the "chrominance" information. The **RGB** color space can be converted into the **YCbCr** color

space by use of the following equations

$$Y = (C_1 * R) + (C_2 * G) + (C_3 * B) \quad (3.2.3a)$$

$$Cb = (B - Y) = (2 - 2 * C_3) \quad (3.2.3b)$$

$$Cr = (R - Y) = (2 - 2 * C_1) \quad (3.2.3c)$$

where $C_1 = 0.2989$, $C_2 = 0.5866$, $C_3 = 0.1145$ for standard images and $C_1 = 0.2126$, $C_2 = 0.7152$, $C_3 = 0.0722$ for HD standard images.

Figure 3.2 shows the YCbCr color cube

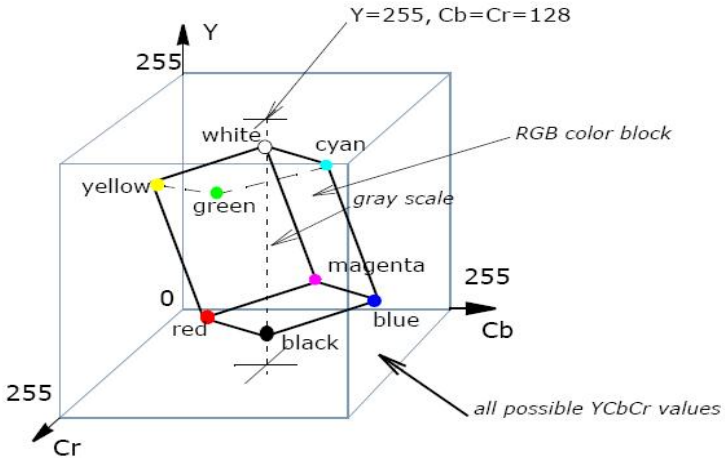

Figure 3.2: The YCbCr color space-YCbCr Cube

3.2.4 HSV Color Space

Hue, Saturation and Value depend on the craftsman ideas of Tone, Shade, and Tint, respectively.

The "**HSV**" (Hue, Saturation, Value) demonstrate characterizes a color space as far as three parts: Hue (**H**), the type of, (for example, blue, yellow). It has a range of 0 to 360 degrees, with 0 degree representing red, 120 degrees representing green, 240 degrees representing blue et cetera. Saturation (**S**) also has a range of 0% to 100%.

It is sometimes called the purity of the color. A lower saturation will represent more "greyness" an a higher, otherwise. Value (**V**), has a range of 0% to 100%. It also represents the brightness. A "hexcone" represents the HSV space where Hue is the point around the vertical axis, Saturation is the separate from the focal axis and Value is the separation along the vertical axis. Essential and secondary pure hues are completely saturated (S = 1).. Figure 3.3 shows the **HSV** color cone which shows the color distribution in the **HSV** color space. Figure 3.4 shows an image displayed in both the **RGB** and **HSV** color spaces. A histogram plot of the **HSV** color space is shown in figure 3.5

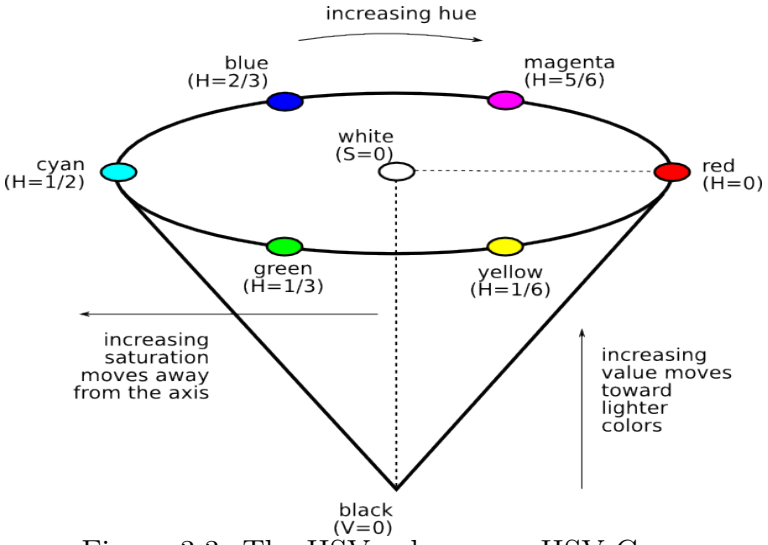

Figure 3.3: The HSV color space-HSV Cone

In this study the normalized **RGB** is used to extract the color. Chrominance-Luminance Independence is reached by removing the blue component after the normalization. This creates an **r-b** color space for skin color analysis. Colour distribution of human skin several nationalities have been explored in [Dour (2000)], being an area in the "r-g" colour space shaped like a shell (normalized RGB color space) which is known as the skin locus. [Soriano et al. (2003)], the values for the colour

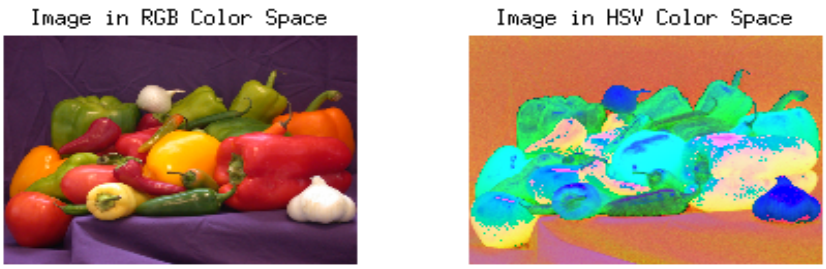

Figure 3.4: An image in the RGB and HSV Color Space

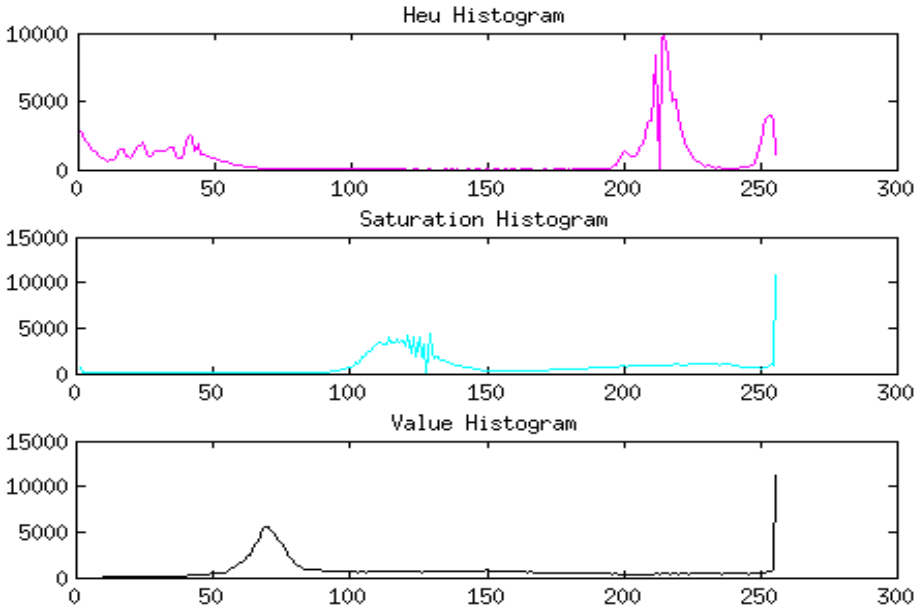

Figure 3.5: Histogram Plot of the HSV Color Space

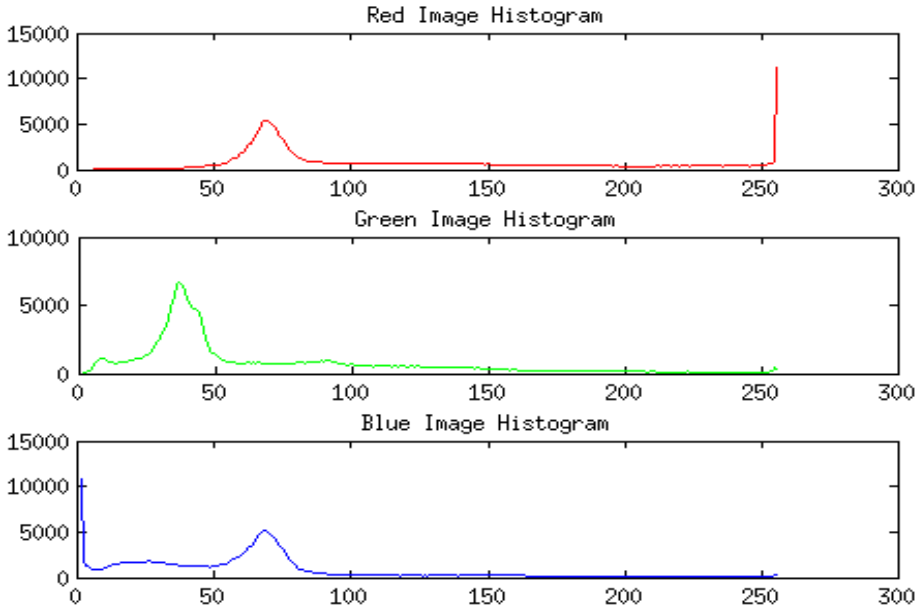

Figure 3.6: Histogram Plot of the RGB Color Space

space are given by

$$r = \frac{R}{R+G+B} \tag{3.2.4a}$$

$$g = \frac{G}{R+G+B} \tag{3.2.4b}$$

A physical based skin model was proposed for the skin color that can be utilized to identify the color of the skin, when the range of the source of light and the qualities of the camera are known. Segments of skin were removed of recordings and pictures under different lighting conditions and their standardized chromaticities were plotted in the **rg**-space as delineated in the figure 3.7. The **r-g** plot is the plot of the normalized red against the normalized green from the normalized RGB color space. The lower and upper limits of the "skin locus" are quadratics. The coefficients that for the quadratics that characterize the membership function for the skin locus were evaluated utilizing least-squares. These boundaries were derived by [Soriano et al.

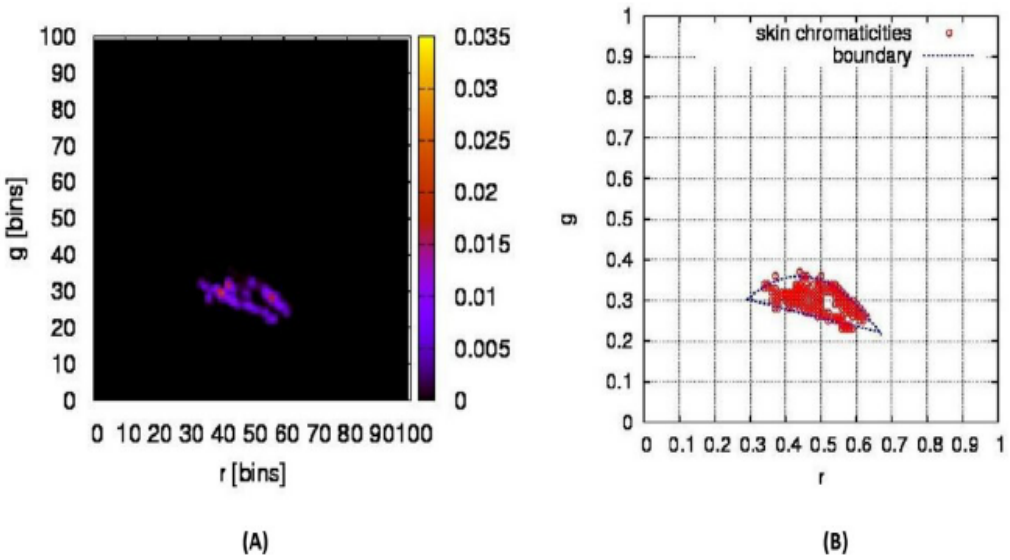

Figure 3.7: Skin Locus Proposed by [Soriano et al. (2003)]

(2003)] in her research towards the detection of skin pigments in sign language. The upper bound is given by

$$g = -1.3767r^2 + 1.0743r + 0.1452 \tag{3.2.5}$$

and the lower bound defined by

$$g = -0.776r^2 + 0.5601r + 0.1766 \tag{3.2.6}$$

A more broad skin locus is created by using the "r-g chromaticity" outline.

Figure 3.8 below delineates distribution of colour in the "**r-g** chromaticity" chart, where rosy hues, somewhat blue hues, and green-like hues are isolated. Notwithstanding the "line-p", a "line-n" is used here to remove skin colour by presenting the "coarse skin" and "fine skin" areas.

To begin with, the "coarse skin" district is characterized utilizing boundaries that are fixed, where skin and skin-like (hues close to the color of the skin) are separated.

At that point, skin color is separated from skin-like colors using the "fine skin" region with varying boundaries.

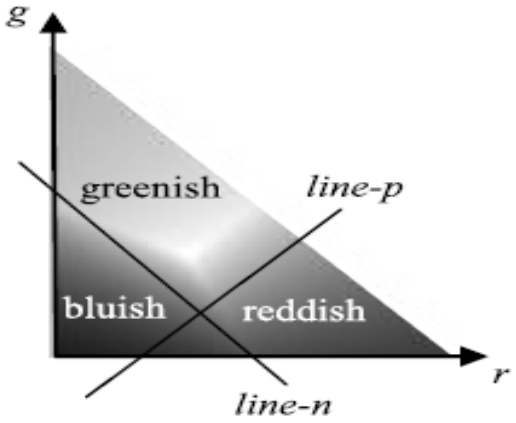

Figure 3.8: Skin Color Distribution Proposed by [Soriano et al. (2003)]

Coarse Skin Region

At the point when pictures are captured under typical brightness conditions, the "skin locus" involves a region close to the center of the "**r-g** chromaticity" chart. Limits of the "coarse skin" locale are characterized as outlined in Figure 3.9, where "line-G", "line-R", "line-B", and "line-up" are give by the following equations [Soriano et al. (2003)]

$$line - G : g = r \tag{3.2.7}$$

$$line - R : g = r - 0.4 \tag{3.2.8}$$

$$line - B : g = -r + 0.6 \tag{3.2.9}$$

$$line - up : g = 0.4 \tag{3.2.10}$$

The line-G(equation 3.2.7), line-R(equation 3.2.8), and line-B(equation 3.2.9) are utilized to expel the greenish, ruddy, and pale blue pixels individually, while the "line-up"(equation 3.2.10) is utilized to clear green-yellow pixels.

Fine Skin Region

When extraction is completed utilizing the "coarse skin" region, we obtain a picture

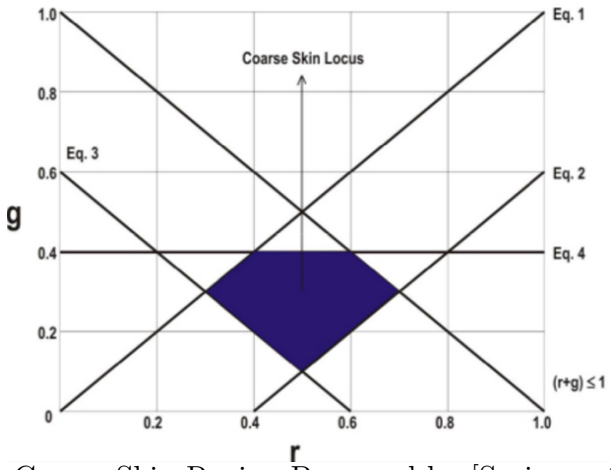

Figure 3.9: Coarse Skin Region Proposed by [Soriano et al. (2003)]

that includes both skin and non-skin pixels. To separate skin pixels from non-skin pixels, a "fine skin" area is created by moving the "line-p" and the "line-n" in Figure 3.9 to the specific positions as per the created histograms as described below. A sloping line with slope 1.0 called line-p could be utilized for isolating red shading adequately, as in by using the line, in the accompanying shows unmistakable peaks/dips for simple edge computation. A histogram describes the quantity of pixels in a picture at its various intensities. This procedure is initialized with a fine skin color segmentation which mimics the coarse skin segmentation process by removing (g-r) and (g + r) histograms of skin candidate pixels. Since skin candidate pixels have simply skin pixels and skin like pixels, front palms and back of hand(s) would compare to first or second greatest neighbourhood crests in the histograms. By considering those 4 neighbourhood crests , new limited skin color borders are created. By cross coordinating of these 4 local peaks, 4 new narrowed skin loci is separated and these loci are subjected to the skin candidate pixels and new images are held by utilizing new loci. One of these loci would relate to genuine skin locus for the present conditions. To choose the genuine skin locus, neighboring pixels of each picture are assembled to locales.

This can also be achieved with the constraint

$$S_{pixel} = \begin{cases} 1 & \text{if } (g < Q_+)\&(g > Q_-)\&(W > 0.004) \\ 0 & \text{otherwise} \end{cases} \tag{3.2.11}$$

where Q_+ is the upper bound quadratic function in the equation above and Q_- is the lower bound quadratic function given in equation 3.2.5 and equation 3.2.6 respectively. Where,

$$W = (r - 0.33)^2 + (g - 0.33)^2 \tag{3.2.12}$$

3.2.5 Image Rotation

After segmentation of the skin region, the selected candidate skin pixel are rotated onto their principal axis. Rotation is done for all image to standardize the analysis process. This allows for the correct principal features to be selected for classification.

The hand is rotated to it's principal axis. This is done by using the moments of the image. The moment of an image is defined as

$$M_{KL} = \sum_x \sum_y x^K y^L I(x, y) \tag{3.2.13}$$

Where $I(x, y)$ is the image intensity at point (x, y). The total energy of the image is given by

$$M_{00} = \sum_x \sum_y I(x, y) \tag{3.2.14}$$

A centroid(centre of mass) can be defined as well as the other moments of the image, if the intensity is regarded at each point (x, y) of the given image as the 'mass' of (x, y). In a two dimensional image the centroid is given by (M_{10}, M_{01})

$$M_{10} = \frac{\sum_x \sum_y xI(x, y)}{M_{00}} \tag{3.2.15}$$

$$M_{01} = \frac{\sum_x \sum_y yI(x, y)}{M_{00}} \tag{3.2.16}$$

The variance(σ^2) is given by the second moment about the centroid(M_{KL}^C)

$$\sigma_x^2 = M_{20}^C = M_{20} - M_{10}^2 \tag{3.2.17}$$

$$\sigma_y^2 = M_{02}^C = M_{02} - M_{01}^2 \tag{3.2.18}$$

σ_x^2 is the expansion or spread of the object in the x direction as σ_y^2 is the expansion or spread of the object in the y direction.

Orientation is defined as the angle of axis of the least moment of inertia; the second order moments are known as the moments of inertia. This determines how the object lies in the field of view [Horn (1987)]. The orientation of an image is given by

$$\tan(2\theta) = \frac{2M_{11}}{M_{20} - M_{02}} \tag{3.2.19}$$

unless $M_{11} = 0$ and $M_{20} = M_{02}$. Consequently

$$\sin(2\theta) = \frac{\pm 2M_{11}}{\sqrt{(2M_{11})^2 + (M_{20} - M_{02})^2}} \tag{3.2.20}$$

$$\cos(2\theta) = \frac{\pm (M_{20} - M_{02})}{\sqrt{(2M_{11})^2 + (M_{20} - M_{02})^2}} \tag{3.2.21}$$

$$\tag{3.2.22}$$

3.2.6 Principal Component Analysis(PCA)

Principal component analysis (PCA) is a standard apparatus in current data analysis. It is a straightforward, non-parametric technique for separating pertinent data from befuddling/complex data sets.

The dimensionality of the feature space is high,ordinarily, for instance, in image processing, the dimensionality can be as high as 320*240, or much higher. In any case, the groups inside the space will frequently lie on a low-dimensional subspace as a result of connections between's the features. A procedure is expected to discover the dimensionality of the subspace.

Lets take a set of training objects which are represented by their feature vectors, $x_i(1 < i < M)$, where M is given as the number of samples, the training can be written in the format of $X = \{x_1, x_2, ..., x_M\}$ whose mean vector is x_m and covariance matrix is R_x, defined by the following equations:

$$x_m = \frac{1}{M}\sum_{i=1}^{m} x_i \tag{3.2.23}$$

$$R_x = \frac{1}{M}\sum_{i=1}^{M}(x_i - x_m)(x_i - x_m)^T \tag{3.2.24}$$

The mean vector is a column vector and the covariance matrix is a real symmetric square matrix of size N by N, where N is the length of the feature vector. T defines the transpose of the matrix.

The training set X relates to a cluster of data points in a N dimensional feature space. There exists repetition in the feature space since the features, ie. the feature space dimensions, are not autonomous of each other. By PCA the repetition can be removed by changing the first feature space into an associated PC space as far as Principal Components. The change is in orthogonal, linear strategy, formalized in the accompanying way:

$$y = W^T x \tag{3.2.25}$$

where y is the new feature vector in the PC space, and W is defined as follows:

$$W = [e_1, e_2, ..., e_N] \tag{3.2.26}$$

e_i is the ith Principal Component(PC), or the orthonormal basis, of the feature space. All PCs must be standardized, that is, $e_l^T e_l = I$.

The dimensionality of the PC space is still N, the same as that of the first component space. However, the connection between the features vanishes. Consequently by

change a feature vector of the first feature space is mapped into the PC space as a isolated point with basis without any dependence on each other

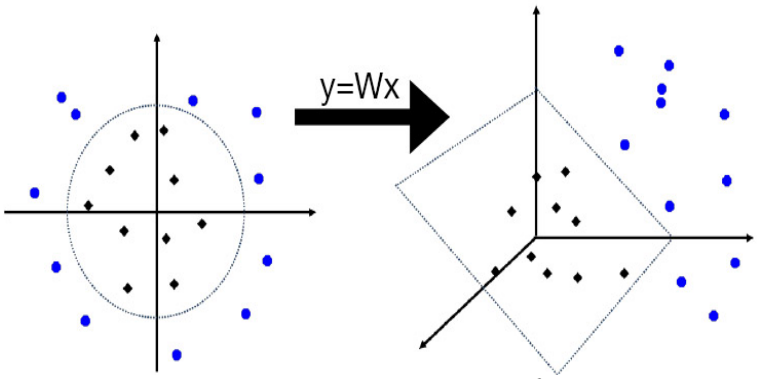

Figure 3.10: Feature space transformation.

No data is lost amid this change. The first component vector can be completely recreated utilizing the transpose of the PCs

$$x' = Wy = WW^T x \tag{3.2.27}$$

where x' is the reconstructed image. Since there is no loss of information, $x' = x$
The principal components now must be discovered . Different techniques have been produced. For instance basic neural network designs have been recommended for recursively assessing the subsets of the PCs, However, the most generally utilized technique is to exploit the covariance network of the feature values of the training set and tackle the *eigen value decomposition problem.*
Given the mean vector and the covariance matrix x_m and R_x,respectively, of a training data set, the eigen value decomposition prob;em is to discover the answer for the accompanying equation:

$$R_x e_i = \lambda_i e_i \tag{3.2.28}$$

where λ_i is the *eigenvalue* corresponds to the ith principal component. This solution can be found by solving its characteristic equation.

For an N by N covariance matrix, N eigen values and N essential segments exist. In the PC space, the information lies on a low dimensional subspace on account of the connection between the components, i.e. the changes in the data concentrates just on a portion of the measurements while the variety along the rest of the measurements is very nearly zero. Looking at the way that every eigen value speaks to the data variation along its PC, those eigen values comparing to the small data changes are near zero. See Figure 3.11

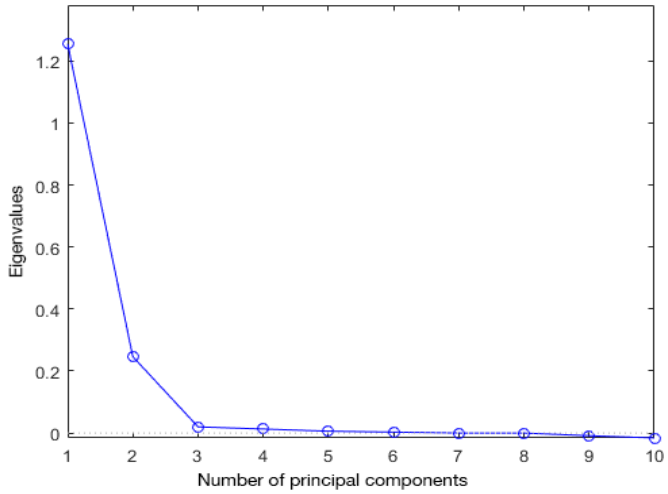

Figure 3.11: An example of a plot of eigenvalues, extracted from a data set of human hand images.

Arranging the eigen values in order of decreasing magnitude allows y to be represented in a compact way by maintaining only the PCs corresponding to the first few largest eigen values while ignoring the rest. That is, the component vector y is processed in the accompanying decreased way:

$$y = W_c^T x = [e_1, ..., e_K]^T x \tag{3.2.29}$$

where W_c is the matrix that contains just the principal K PCs that are selected and e_i relates to the biggest eigenvalue, e_2 the second biggest eigenvalue. This is done sequentially till the kth eigen value.

Normally $K << N$, and the dimensionality is reduced from N to K. The recreation of the first component vector is given by:

$$x' = W_c y = [e_1, ..., e_k][e_1, ..., e_k]^T x \tag{3.2.30}$$

However, at this stage, the reconstructed vector is different from the original one: $x' \neq x$. To measure the difference, the energy of the training set is defined in the equation below

$$E = \sum_{i=1}^{N} \lambda_i \tag{3.2.31}$$

The error between the reconstructed vector x' and the original feature vector x can be measured using the criterion of percentage of total energy remaining. The percentage of the remaining energy can be written as a function of the number of the retained PCs:

$$E_{remain}(K) = \frac{\sum_{i=1}^{K}}{E} \tag{3.2.32}$$

PCA minimizes the mean square reconstruction error of the training set:

$$\varepsilon = min\left[\sum_{i=1}^{M} \|x_i - x_i'\|\right] = \sum_{i=K+1}^{N} \lambda_i \tag{3.2.33}$$

$$= E - \sum_{i=1}^{K} \lambda_i \tag{3.2.34}$$

$$= E(1 - E_{remain}(K)) \tag{3.2.35}$$

The Equation above tells that the greater K, the smaller the reconstruction error, and the more PCs retained the more energy remaining.

In essence, PCA assumes the projection of the training samples in the PC space is bounded by a hyper ellipsoid. λ_i, is the variance of the data along the ith PC- more than 98% of distribution of the data is covered in the range $[-3\sqrt{\lambda_i}, -3\sqrt{\lambda_i}]$ along this PC.

3.2.7 Construction of Image Vectors

An image made out of $m \times n$ pixels, given m and n as the measurements of the image. A vector is developed by doing a row by row concatenation of the image pixels
The picture vector can be composed as $f = [a_{11}, ..., a_{1n}, ..., a_{m1}, ...a_{mn}]$ The number of columns and rows are no longer critical here, the image vector is changed into $f = [a_1, ...a_N]$. For a 32×32 picture matrix $N = mn = 1024$. Every pixel in the grayscale picture vector can be dealt with as a variable changing from 0 to 255. Subsequently the entire picture vector is an random vector which contains 1024 arbitrary variables for a 32×32 image matrix.

Each picture vector now relates as a solitary entity in an N dimensional component space. The dimensions in the element space are not independent from each other since connection exists between the pixels. Utilizing the system of PCA depicted over, the component space is changed into another PC space whose dimensionality is much lower than the first space.

Each PC in the PC space speaks to a method of changes of the hand structure. The principal PC relates to the variation of a huge scale structure where the biggest eigenvalue is its weight, the second PC compares to the variety of a less substantial scale structure where the second biggest eigenvalue is its weight, etcetera. Figure 3.12 gives an illustration.

The PC space utilized was prepared by the training set containing 6 autonomous gestures with more than 900 frames for every spelling gesture. The variations of the initial three PCs are displayed. Some normal reproduction of images utilizing the initial three PCs are given in 3.13, where the two static signals on the left side are the first pictures of "C" and "Point" separately, while the two pictures on the privilege are their reconstruction.

While applying PCA, one issue is to choose the dimensionality of the element space,

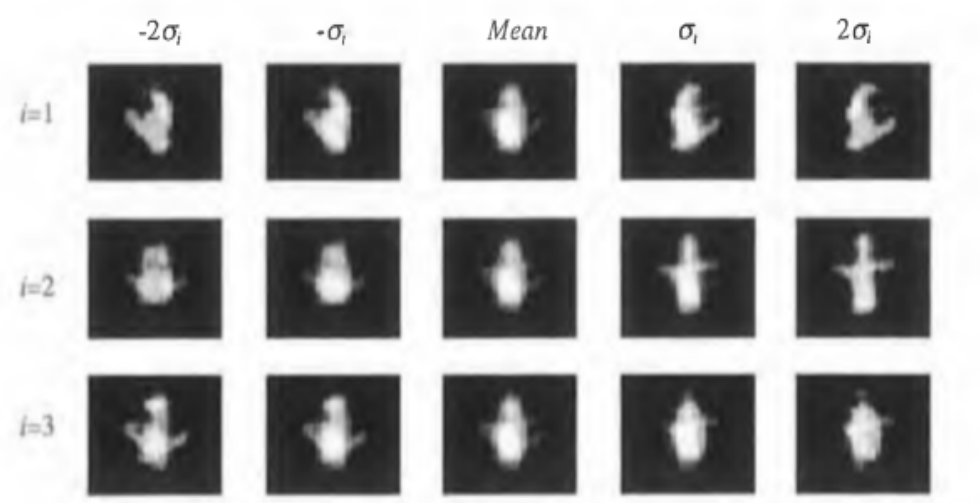

Figure 3.12: The first three PCs. σ_i, is the standard deviation along the ith PC, and $\sigma_i = \sqrt{\lambda_i}$.

i.e. what number of PCs to hold. As a rule the assignment of deciding what number of PCs to hold involves speaking to however much data as could be expected. The more PCs the more energy can be retained. In any case, holding more PCs additionally implies more calculation and less dimensionality reduction, and some of the time more PCs will acquire more clamour, for example, the variety of the light condition or other white clamour. Subsequently, this is a trade-off between the loss of the energy and the amount of calculation. The quantity of PCs to be held is chosen utilizing a straightforward threshhold T. The model is formalized as

$$E_{remain}(K) > T \qquad 0 < T < 1 \tag{3.2.36}$$

where $E_{remain}(K)$ is defined in equation(3.2.32). From trials it is found that 0.95 would be a reasonable threshold for some applications.

3.2.8 Training Phase

The initial phase in the recognition stage is to build a picture database. The securing of picture information is a key issue in the framework assessment, particularly for a

Figure 3.13: Image reconstruction using PCA.

framework utilizing an appearance-based hand model. In the event that the greater part of the pictures in the database focus on one or some particular view points, the inevitable recognition rate would be very great. Be that as it may, the framework would be touchy to the progressions of the hand configurations or positions. Thus the recognition rate would not gauge the performance of the framework well. For a fairly judged recognition, the accompanying techniques are utilized

1. A different database is developed for each of the finger-spelling gesture, with every database containing 100 pictures. Diverse view angles are incorporated into the database however much as could be expected.

2. A holdout technique is connected for choosing the training and test sets from the databases. Initial images, say 70, are randomly selected from every database for training, then another set of pictures, say another 70, are separated from whatever is left of the pictures in every database. The technique guarantees the independence amongst training and test tests.

3. The classifier is trained utilizing the training set and compute its precision

utilizing the test set.

4. Steps 2 to 3 are rehashed for 10 times to work out the average accuracy

3.2.9 Classification

Two classifiers will be discussed, a single PC space is calculated which contains all 7 hand gestures.The average vector relating to the gesture classes in the PC space is calculated.

1. Single-Space Euclidean Classifier

   The least demanding strategy to manage a characterization issue is to dole out the unknown object into the group with the least Euclidean separation between the unknown element and the average of the group. Assume there is an unknown element vector $p = [p_1, p_2, ..., p_k]$ and the ith class is characterized by its mean vector $w_i = [w_{i1}, w_{i2}, ..., w_{iK}]$, where K is the dimensionality of the feature space. The Euclidean distance between them is defined as:

$$d_E(p, w_i) = |p - w_i| = \left[\sum_{j=i}^{K}(p_j - w_{ij})^2\right]^{\frac{1}{2}} \tag{3.2.37}$$

   The classification can then be formulated as:

$$w_c = min_i(d_E(p, w_i)) \tag{3.2.38}$$

   where w_c is a whole number that represents to the picked class. This is a basic standard with next to no calculation. The classifier that uses the above grouping criterion is known as a single space Euclidean classifier since just a single PC space is processed from the general training samples. This classifier was used in our recognition.

2. Single-space Mahalanobis Classifier The single-space Euclidean classifier treats the distance along each measurement of the PC space in a similar way and does not think about the diverse fluctuation along various bearings. To defeat this issue , another classifier is executed: the single-space Mahalanobis classifier. It too works in a single PC space, however utilizes the Mahalanobis distance, characterized in equation(3.2.39), instead of the Euclidean distance in the PC space.

$$d_M(p, w_i) = |p - w_i| = \left[\sum_{j=1}^{K} \left(\frac{p_j - w_{ij}}{\sigma_j} \right)^2 \right]^{\frac{1}{2}} \tag{3.2.39}$$

where σ_j is the standard deviation along the jth PC. The single-space Mahalanobis classifier can be formalized by the following equation:

$$w_c = min_i(d_M(p, w_i)) \tag{3.2.40}$$

3333333 Since in the feature space, the jth eigenvalue, λ_j, represents the data variance along the jth PC, the above classification criterion can be rewritten as:

$$w_c = min_i \left[\sum_{j=1}^{K} \frac{(p_j - w_{ij})^2}{\lambda_j} \right]^{\frac{1}{2}} \tag{3.2.41}$$

Chapter 4

Analysis and Recognition Results

The system was trained to recognize 6 independent handspelling gestures. With over 900 images per gesture.

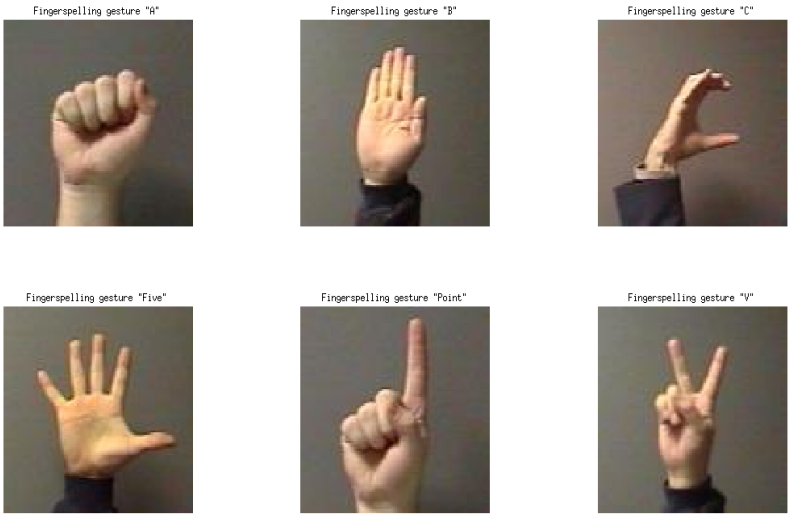

Figure 4.1: The 6 hand spelling gestures that were trained and recognized

4.1 Pre-processing(Skin Detection and Segmentation)

The system was trained to recognize the 6 gestures shown in Figure 4.1. Figures 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7 show the skin segmentation process of a random selection of the training data for each gesture. These images show the selection of skin-like pixel by our system given different brightness and contrast settings and different threshold values. Selection was of skin-like particles was either “over” or “under” exposed.

Figure 4.2: Hand region segmentation for gesture ”A”

After training the system a separate testing dataset is used to test the system. This dataset has over a hundred images for each of the 6 handspelling gestures. The figures below show some of the output from the testing data.

The system achieved an overall 60% accuracy in it’s classification. Comparatively the Mahalanobis Classifier performed 20% better at classifying the test gestures. Figures 4.8, 4.9, 4.10, 4.11 and 4.12 show some randomly selected classification of the system. It is worth noting once again that choosing different number of PCs can

Figure 4.3: Hand region segmentation for gesture "B"

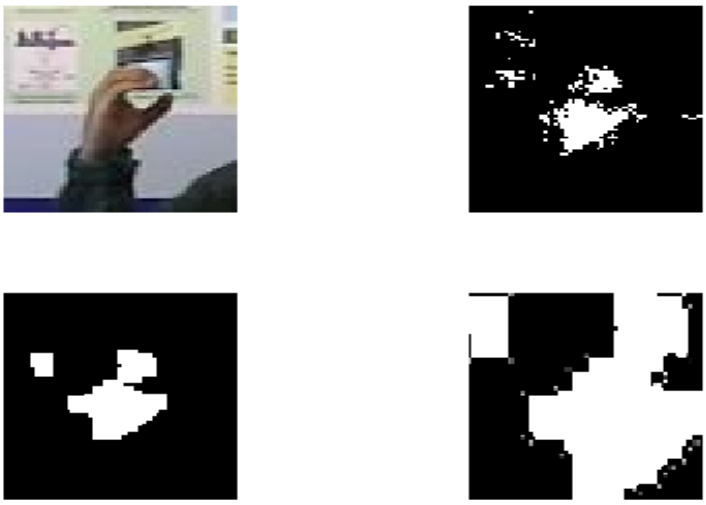

Figure 4.4: Hand region segmentation for gesture "C"

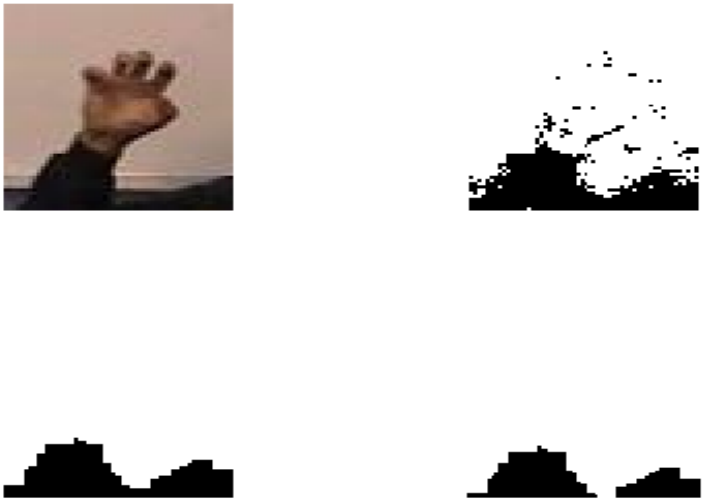

Figure 4.5: Hand region segmentation for gesture "Five"

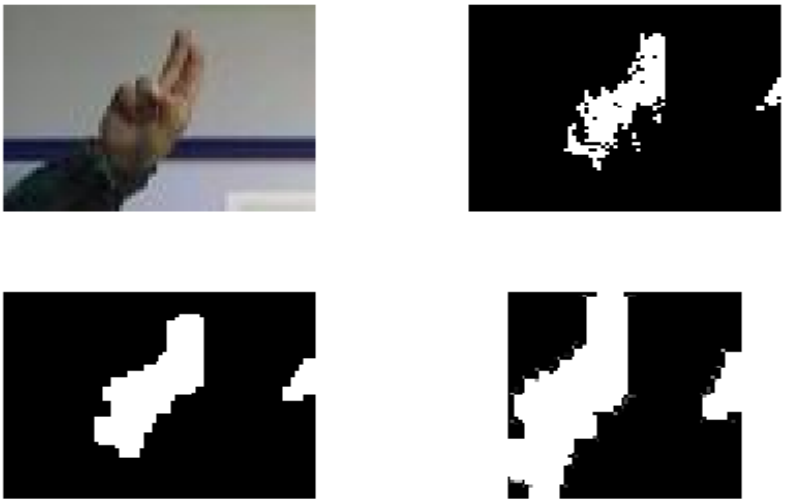

Figure 4.6: Hand region segmentation for gesture "Point"

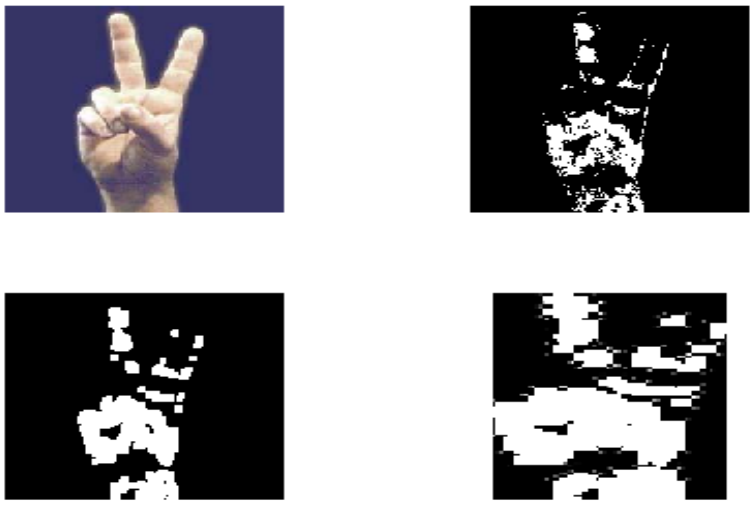

Figure 4.7: Hand region segmentation for gesture "V"

Figure 4.8: Testing Classification for gesture "B"

Figure 4.9: Testing Classification for gesture "C"

Figure 4.10: Testing Classification for gesture "Point"

Figure 4.11: Testing Classification for gesture "V"

Figure 4.12: Testing Classification for gesture "V"

affect the ultimate performance greatly. Increasing the number of PCs increases the recognition rate. But too much increase in the number of PC leads to a drop in the recognition rate because there is a trade-off between the recognition rate and the computational speed. When too many PCs are included, noise will affect the final result and thus cause a drop in recognition rate.

Chapter 5

CONCLUSION and RECOMMENDATIONS

5.1 Summary of Results

With a little over 900 images tested, the system achieved an overall 60% total accuracy in it's classification of the 6 finger spelling gestures.

The Mahalanobis classifier performed better; classifying correctly 50 more gestures than the Euclidean classifier.

5.2 Conclusion

A functioning system that can recognize finger spelling gestures was successfully developed. In the performance analysis, the system does fairly at classifying the 6 hand spelling gestures. The number of PCs selected greatly affects the accuracy of the classification. This comes with a trade-off between computational time and noise though. The single space Euclidean Classifier was tested alongside the single space Mahalanobis Classifier.

The single space Mahalanobis Classifier performed better than the single space Eu-

clidean classifier especially as the number of PCs increased.

5.3 Recommendation for Further Studies

5.3.1 Recommendation

The importance of the skin segmentation process cannot be understated in this system. Selecting of false skin pixels can greatly affect the accuracy of the classification. The segmentation of skin pixels greatly affects the selection of PCs. Based on the conclusion it is recommended that a more precise method of skin pixel segmentation be employed. The number of PCs selected should be optimized based on the type of classifier being employed to reduce the noise that affects the classification process.

5.3.2 Further Studies

1. A hierarchical decision tree combined with multi-scale theory to speed up the recognition procedure.

2. The use of HMMs to recognize dynamic hand gestures(hand gestures in videos).

3. A multiple subspace classifier with a decision tree to help improve the classification of the proposed gestures

REFERENCES

Argyros, A. A. and Lourakis, M. I. A. (2004a), "3D tracking of skin colored regions by a moving stereoscopic observer." in *IApplied Optics*, vol. 43, pp. 366–378.

Argyros, A. A. and Lourakis, M. I. A. (2004b), *Real-time tracking of multiple skin-colored objects with a possibly moving camera.*

Argyros, A. A. and Lourakis, M. I. A. (2006), "Vision-based interpretation of hand gestures for remote control of a computer mouse," in *ECCV Workshop on HCI*, p. 4051, Graz, Austria.

Baumberg, A. and Hogg, D. (1994), "Learning flexible models from image sequences." *Proc. European Conference on Computer Vision*, 1, 299–308.

Birk, H., Moeslund, T. B., and Madsen, C. B. (1997), "Real-time recognition of hand alphabet gestures using principal component analysis," in *Proc. Scandinavian Conference on Image Analysis*, Lappeenranta, Finland.

Blake, A., North, B., and Isard., M. (1999), "Learning multi-class dynamics." *Proc. Advances in Neural Information Processing Systems (NIPS)*, 11, 389–395.

Bregler, C. (1997), "Learning and recognizing human dynamics in video sequences," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, p. 568574, Puerto Rico.

Chai, D. and Ngan, K. (1998), "Locating the facial region of a head and shoulders color image." *IEEE Int. Conference on Automatic Face and Gesture Recognition*, pp. 124–129.

Cootes, T. F. and Taylor, C. J. (1992), "Active shape models-smart snakes." in *British Machine Vision Conference*, pp. 266–275, Colorado.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995), "Active shape models - their training and applications." *Computer Vision and Image Understanding,*, 61, 38–59.

Cote, M., Payeur, P., and Comeau, G. (2006), "Comparative study of adaptive segmentation techniques for gesture analysis in unconstrained environments." *IEEE Int. Workshop on Imagining Systems and Techniques*, pp. 28–33.

Crowley, J., Berard, F., and Coutaz, J. (1995), "Finger tracking as an input device for augmented reality," in *International Workshop on Gesture and Face Recognition*, Zurich.

Cui, Y. and Weng, J. (1996), "Hand sign recognition from intensity image sequences with complex background." *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 88–93.

Cui, Y., Swets, D., and Weng, J. (1995), "Learning-based hand sign recognition using shoslf-m," in *Int. Workshop on Automatic Face and Gesture Recognition*, p. 201206, Zurich.

Derpanis, K., Wildes, R., and Tsotsos, J. (2004), "MRI segmentation: Methods and applications," *Hand Gesture Recognition within a Linguistics-Based Framework*, 3021, 282–296.

Dipietro, L., Sabatini, A. M., and Dario, P. (2008), "A survey of glove-based systems and their applications," in *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*, vol. 38, pp. 461–482.

Dour, S. (2000), "Skin Detection method for Indian Sign Language Recognition," techreport, Navarachana University, Vadodara.

Gao, W., Ma, J., Wu, J., and Wang, C. (2000), "Sign language recognition based on HMM/ANN/DP." in *International journal of pattern recognition and artificial intelligence*, vol. 14, pp. 587–602.

Garg, P., Aggarwal, N., and Sofat, S. (2009), "Vision Based Hand Gesture Recognition," in *Academy of Science, Engineering and Technology*, vol. 49, pp. 972–977.

Gavrila, D. and Davis, L. (1996), "3-D model-based tracking of humans in action: a multi-view approach." *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 73–80.

Goncalves, L., Di Bernardo, E., and Ursella, E. (1995), "Monocular tracking of the human arm in 3D," in *International Conference on Computer Vision (ICCV)*, p. 764770, Cambridge.

Gupta, L. and Ma, S. (2001), "Gesture-Based Interaction and Communication: Automated Classification of Hand Gesture Contours." *IEEE transactions on systems, man, and cyberneticspart c: applications and reviews,*, 31.

Hager, G. and Belhumeur, P. (1996), "Real-time tracking of image regions with changes in geometry and illumination," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 403410,, Washington, DC.

Horn, B. K. P. (1987), "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America*, 4, 629–642.

Imagawa, K., Lu, S., and Igi, S. (1998), "Color-based hands tracking system for sign language recognition." in *Conf. Face and Gesture Recognition,*, pp. 462–467, Zurich.

Jones, M. J. and Rehg, J. M. (2002), "Statistical color models with ap- plication to skin detection." *International Journal of Computer Vision*, 1, 81–96.

Kadous, M. (1996), "Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language," in *Proceedings of the Workshop on the Integration of Gestures in Language and Speech*, pp. 165–174.

Kalman, R. E. (1960), "A new approach to linear flitering and prediction problems," *Transactions of the ASMEJournal of Basic Engineering*, 82, 35–42.

Kjeldsen, R. and Kender, J. (1996), "Finding skin in color images." *Int. Conf. Automatic Face and Gesture Recognition*, pp. 312–317.

Kohler, M. (1997), "System architecture and techniques for gesture recognition inunconstrained environments," in *International Conference on Virtual Systems and MultiMedia*, vol. 10-12, p. 137146.

Krueger, M. (1991), *Artificial Reality II*, Addison Wesley, Reading, MA.

Krueger, M. (1993), "Environmental technology: Making the real world virtual." *Communications of the ACM*, 36, 36–37.

Lee, C. and Xu, Y. (1996), "Online, interactive learning of gestures for human robot interfaces," in *Carnegie Mellon University, The Robotics Institute*.

Lee, J. and Kunii, T. L. (1995), "Model-based analysis of hand posture." *IEEE Computer Graphics and Applications*, 15, 77–86.

MacCormick, J. and Blake, A. (199), "A probabilistic exclusion principle for tracking multiple objects." *Proc. International Conference on Computer Vision (ICCV)*, pp. 572–578.

MacCormick, J. and Isard, M. (2000), "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *European Conference on Computer Vision*, p. 319.

Martin, J. and Crowley, J. (1997), "An appearance-based approach to gesture recognition." *Int. Conf. on Image Analysis and Processing,*, pp. 340–347.

Martin, J., Devin, V., and Crowley, J. (1998), "Active hand tracking," in *IEEE Conference on Automatic Face and Gesture Recognition*, p. 573578, Nara, Japan.

Martinez, A., Wilbur, B., Shay, R., and Kak, A. (2002), "Purdue rvl-slll asl database for automatic recogition of american sign language." in *International Conference on Multimodal Interfaces*, pp. 167–172, Zurich.

Meyering, A. and Ritter, H. (1992), "Learning to recognize 3D-hand postures from perspective pixel images," *Artificial Neural Networks II*, p. 821824.

Murase, H. and Nayar, S. (1995), "Visual learning and recognition of 3- d objects from appearance." *International Journal of Computer Vision,*, 14, 5–24.

Nolker, C. and Ritterpages, H. (1998), "Illumination independent recognition of deictic arm postures," in *IEEE Industrial Electronics Society*, p. 20062011, Germany.

OHagan, R. and Zelinsky, A. (1997), "Finger Track - a robust and realtime gesture interface," in *Australian Joint Conference on Artificial Intelligence*, p. 475484, Perth, Australia.

Pavlovic, V. I., Sharma, R., and Huang, T. S. (2000), "Visual Interpretation of Hand Gestures for Human- Computer Interaction: A Review," in *IEEE Transactions on Pattern Analysis And Machine Intelligence*, vol. 19, pp. 677–695.

Quek, F. and Zhao, M. (1996), "Inductive learning in hand pose recognition," *IEEE Automatic Face and Gesture Recognition*, p. 7883.

Rehg, J. and Kanade, T. (1994), "Digiteyes: Vision-based hand tracking for human-computer interaction." *Workshop on Motion of Non-Rigid and Articulated Bodies*, pp. 16–24,.

Rehg, J. and Kanade, T. (1995), "Model-based tracking of self-occluding articulated objects." *Proc. International Conference on Computer Vision (ICCV)*, pp. 612–617.

Rockett, P. I. (2003), "Performance assessment of feature detection algorithms: Amethodology and case study on corner detectors," *IEEE Trans. Image Process.,*, 12, 1668–1676.

Saxe, D. and Foulds, R. (1996), "Toward robust skin identification in video images." *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, p. 379384.

Schapire, R. (2002), "The boosting approach to machine learning: An overview," in *MSRI Workshop on Nonlinear Estimation and Classification*.

Sigal, L., Sclaroff, S., and Athitsos, V. (2004), "Skin color-based video segmentation under time-varying illumination." *IEEE Trans. Pattern Analysis and Machine Intelligence,*, 26, 862–877.

Sirovich, L. and Kirby, M. (1987), "Low-dimensional procedure for the characterization of human faces." *Journal of the Optical Society of America,*, 4, 519–524.

Soriano, M., Martinkauppib, B., Huovinenb, S., and Laaksonenc, M. (2003), "Adaptive skin color modeling using the skin locus for selecting training pixels," *The Journal of Pattern Recognition Society*, 36, 681–690.

Starner, T. and Pentland, A. (1995), "Visual recognition of american sign language using hidden markov models." in *IEEE International Symposium on Computer Vision,*.

Stenger, B., Mendonca, R., and Cippola, R. (2002), "Model-based 3D tracking of an articulated hand," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, p. pages 126133, Hawaii, Hawaii.

Terrillon, J., Shirazi, M., Fukamachi, H., and Akamatsu, S. (2000), "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images." *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 54–61.

Terzopoulos, D. and Szeliski, R. (1992), *Tracking with Kalman Snakes*, p. 320, MIT Press.

Triesch, J. and Malsburg, C. (1996), "Robust classification of hand postures against complex background," *IEEE Automatic Face and Gesture Recognition*, p. 170175.

Triesch, J. and Von der Malsburg, C. (1998), "A gesture interface for human-robot-interaction," in *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, p. 546551, Nara, Japan.

Utsumi, A. and Ohya, J. (1998), "Image segmentation for human tracking using sequential-image-based hierarchical adaptation." *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 911–916.

Utsumi, A. and Ohya, J. (1999), "Multiple-hand-gesture tracking using multiple cameras," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, p. 473478, Colorado.

Vogler, C. and Metaxas, D. (1999), "Toward scalability in asl recogni- tion: Breaking down signs into phonemes in gesture workshop." in *International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pp. 211–224.

Waldron, M. (1995), "Isolated asl sign recognition sytem for deaf persons." *IEEE Transactions on Rehabilitation Engineering,*, 3, 261–271.

Waldron, M. and Kim, S. (1995), "Isolated ASL sign recognition system for deaf persons." *IEEE Transactions on Rehabilitation Engineering,*, 3, 261–271.

Watson, R. (1993), "Gesture recognition techniques," in *Technical report, Trinity College, Department of Computer Science*.

Wu, Y. and Huang, T. S. (2000), "View-independent recognition of hand postures," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 2, p. 8494, Hilton Head Island, SC.

Yang, J., Lu, W., and Waibel, A. (1998), "Skin-color modeling and adaptation." *International Conference on Electrical, Communications, and Computers*, pp. 687–694.

Yang, M. and Ahuja, N. (1998), "Detecting human faces in color images," *Proc. International Conference on Image Processing (ICIP)*, pp. 127–130.

Yang, M. H., Ahuja, N., and Tabb, M. (2002), "Extraction of 2D motion trajectories and its application to hand gesture recognition." *IEEE Transactions Pattern Analysis and Machine Intelligence,*, 24, 1061–1074.

Yuan, Q., Sclaroff, S., and Athitsos, V. (1995), "Automatic 2D hand tracking in video sequences," *IEEE Workshop on Applications of Computer Vision*, pp. 250–256.

Zhu, X., Yang, J., and Waibel, A. (2000), "Segmenting hands of arbitrary color." *Proc. International Conference on Automatic Face and Gesture Recognition (FG),*, pp. 446–455.