5-2015

Face modeling for face recognition in the wild.

Eslam AbdelFattah Mostafa
University of Louisville

FACE MODELING FOR FACE RECOGNITION IN THE WILD

By

Eslam AbdelFattah Mostafa
B.S., Alexandria University, Egypt, 2006

A Dissertation
Submitted to the Faculty of the
J. B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering
University of Louisville
Louisville, Kentucky

May 2015

FACE MODELING FOR FACE RECOGNITION IN THE WILD

By

Eslam AbdelFattah Mostafa
B.S., Alexandria University, Egypt, 2006

A Dissertation Approved on

April 13,2015

by the Following Reading and Examination Committee:

Aly A. Farag, Ph.D.

John F. Naber, Ph.D.

Robert W. Cohn, Ph.D.

Prasanna K. Sahoo, Ph.D.

Roman V. Yampolskiy, Ph.D.

DEDICATION

My beloved country

EGYPT

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Aly Farag, for his guidance and patience over the past five years. I would also like to thank my committee members, Dr. John Naber, Dr. Roman Yampolskiy, Dr. Prasanna Sahoo, and Dr. Robert W. Cohn, for their valuable comments and great support. I would also like to thank all the CVIP lab members for their help and support. I would like to thank the Speed school of Engineering and the ECE department for their continuous support. I also would like to thank the graduate school of the University of Louisville for their continuous support.

ABSTRACT

FACE MODELING FOR DIFFERENT FACE UNDERSTANDING APPLICATIONS

Eslam AbdelFattah Mostafa

April 13, 2015

Face understanding is considered one of the most important topics in computer vision field since the face is a rich source of information in social interaction. Not only does the face provide information about the identity of people, but also of their membership in broad demographic categories (including sex, race, and age), and about their current emotional state. Facial landmarks extraction is the corner stone in the success of different facial analyses and understanding applications. In this dissertation, a novel facial modeling is designed for facial landmarks detection in unconstrained real life environment from different image modalities including infra-red and visible images.

In the proposed facial landmarks detector, a part based model is incorporated with holistic face information. In the part based model, the face is modeled by the appearance of different face part(e.g., right eye, left eye, left eyebrow, nose, mouth) and their geometric relation. The appearance is described by a novel feature referred to as pixel difference feature. This representation is three times faster than the state-of-art in feature representation. On the other hand, to model the geometric relation between the face parts, the complex Bingham distribution is adapted from the statistical community into computer vision for modeling the geometric relationship between the facial elements. The global information is incorporated with the local part model using a regression model. The model results outperform the state-of-art in detecting facial landmarks. The proposed facial landmark detector

is tested in two computer vision problems: boosting the performance of face detectors by rejecting pseudo faces and camera steering in multi-camera network.

To highlight the applicability of the proposed model for different image modalities, it has been studied in two face understanding applications which are face recognition from visible images and physiological measurements for autistic individuals from thermal images. Recognizing identities from faces under different poses, expressions and lighting conditions from a complex background is an still unsolved problem even with accurate detection of landmark. Therefore, a learning similarity measure is proposed. The proposed measure responds only to the difference in identities and filter illuminations and pose variations. similarity measure makes use of statistical inference in the image plane. Additionally, the pose challenge is tackled by two new approaches: assigning different weights for different face part based on their visibility in image plane at different pose angles and synthesizing virtual facial images for each subject at different poses from single frontal image. The proposed framework is demonstrated to be competitive with top performing state-of-art methods which is evaluated on standard benchmarks in face recognition in the wild.

The other framework for the face understanding application, which is a physiological measures for autistic individual from infra-red images. In this framework, accurate detecting and tracking Superficial Temporal Arteria (STA) while the subject is moving, playing, and interacting in social communication is a must. It is very challenging to track and detect STA since the appearance of the STA region changes over time and it is not discriminative enough from other areas in face region. A novel concept in detection, called supporter collaboration, is introduced. In support collaboration, the STA is detected and tracked with the help of face landmarks and geometric constraint. This research advanced the field of the emotion recognition.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

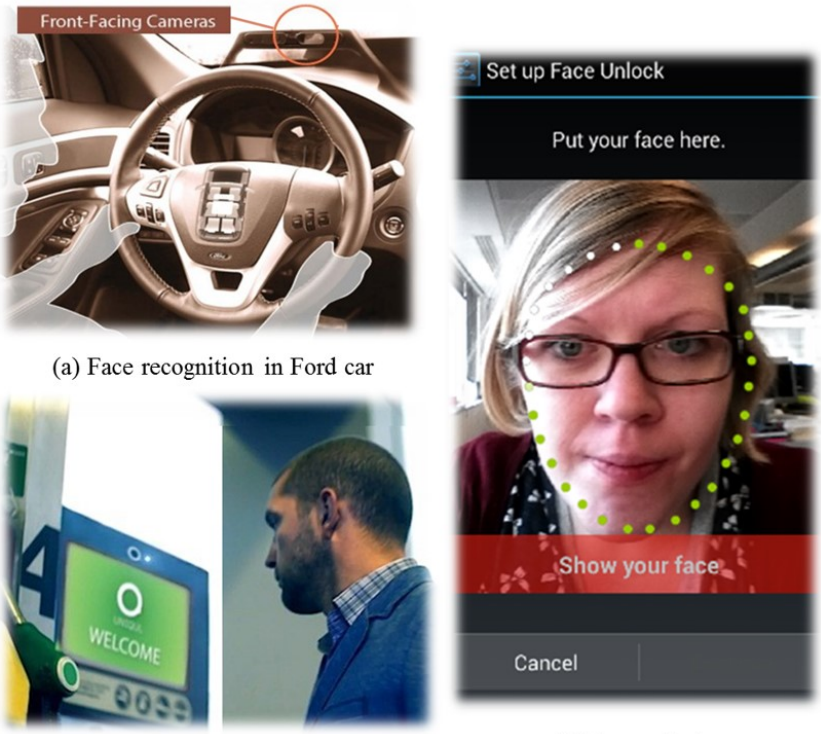
CHAPTER 1
INTRODUCTION

Visual perception is probably the most important sensing ability for humans to enable social interactions and general communication. Many researches have attempted to mimic human visual perception by computer-based methods. This new field is called computer vision which is the intersection of artificial intelligence, machine learning, image processing, graphics and cognitive science. Although a large number of applications are explored using this field's approaches, these approaches only try to mimic the first layer of human visual perception and going beyond the first layer requires more complicated techniques

Face understanding is considered one of the most important topics in the field of computer vision. The face is a rich source of information in social interaction. Not only does the face provide information about identity of people, but also their membership in broad demographic categories of humans (including sex, race, and age), and about their current emotional state.

Facial recognition technology(FRT) has emerged as an attractive solution to address many contemporary needs for identification and verification of identity claims. It brings together the promise of other biometric systems, which attempt to tie identity to individually distinctive features of the body, and the more familiar functionality of visual surveillance system without requiring physical contact to the sensor and active cooperation from the target may not be required. Since the middle of the last century, face recognition is noted to be attractive to intelligent agency and government applications: police station, airport, metro station, etc.

In the last five years, face recognition has become more popular in many other applications such as the car industry, cell phone industry, advertising, and social networking.



(a) Face recognition in Ford car

(c) Pay with your face from Uniqul

(b) Face unlock

FIGURE 1: Application for Face recognition in non-government applications. (a)automatic personalization of your car, (b)unlock your mobile phone, (c)pay with your face.

In the automotive industry, Ford has incorporated a new system in its cars that recognizes the driver as shown in Figure 1-a. It was used to automatically adjust the mirrors, seat and steering wheel, as well as turn on a favorite radio station. The system was also used to modify performance settings such as throttle response, gear shift patterns and suspension stiffness. In the smart phone industry, Android smart phones use facial recognition to unlock phones. This feature is called Face Unlock as shown in Figure 1-b. In the finance industry, Uniqul has launched the first ever payment platform based on face recognition. The system enables customers to pay without having a wallet, card or mobile phone. Paying is as easy as giving the camera a nod and pressing OK on a point-of-sale tablet as shown in Figure 1-c. Moreover, Millennial ATM use facial recognition as its primary security method. Nowadays, auto tagging is popular feature in social networking sites like Facebook and personal photo organizers like Picasa which enables users to add metadata

about an image that include the names of the people in the image.

Recently, the interest in another face understanding application called facial expressions recognition, or emotion recognition, has increased with the emergence of the Human Computer Interaction(HCI) field. The facial expression is divided into two categories micro and macro expression. Micro-expressions are very brief facial expressions lasting only a fraction of a second. They occur when a person either deliberately or unconsciously conceals a feeling. Psychological research has classified six facial expressions which correspond to distinct universal emotions: disgust, sadness, happiness, fear, anger, and surprise [159] as shown in Figure 2. Facial expression is a visible manifestation of the affective state, cognitive activity, intention, personality, and psychopathology of a person. It conveys non-verbal communication cues. There are many application areas that could benefit from the ability to detect affect. These include interfaces that do not interrupt their users when they are stressed, online learning systems that adapt the teaching if the student is confused, and video games that adapt their difficulty based on player engagement. Further applications include: assisted living environments that can monitor the user's state and report to medical professionals if the patient is feeling pain; assistive technologies for diagnosing conditions such as depression; and systems that monitor drivers or pilots for boredom.

1.1 Motivation and Contribution

1. Facial Feature Points detection

Face understanding applications face recognition systems, expression recognition, gender recognition, measure of beauty, ethnicity recognition are based on the shape of the face. Detecting face shape is considered to be the corner stone in these applications. Face anthropometry provides a set of meaningful measurements that allow the most complete description the shape of the face. The measurements are taken between the landmark points, facial feature points, defined in terms of visually-identifiable or palpable features on the

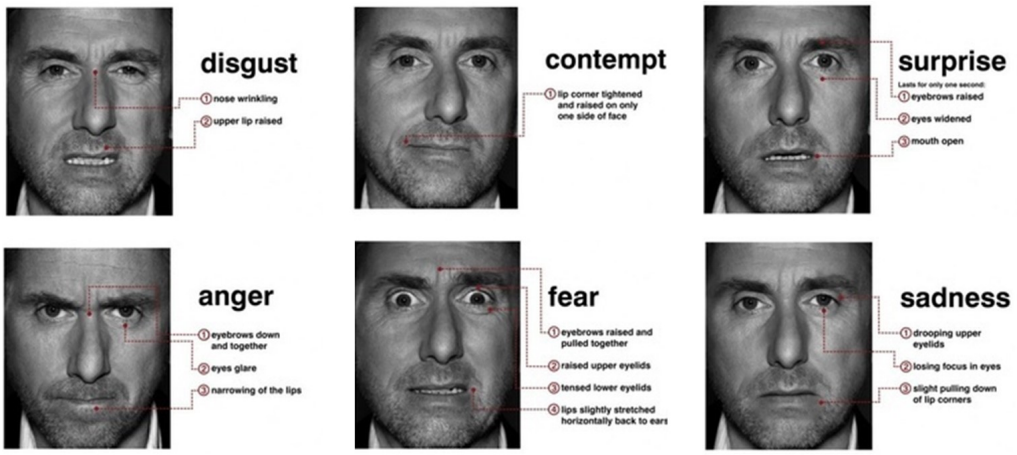


FIGURE 2: Six emotions which have universal signals are disgust, contempt, surprise, anger, fear, and sadness.

subject's face using carefully specified procedures and measuring instruments.

Automatically detecting these face facial points, face landmarks, for describing the face shape is an essential and challenging task. It is an essential task since the accuracy of different face understanding applications such as face recognition, expression recognition, and age recognition depends mainly on the accuracy of detecting these facial points. The challenging side is that detection of these points need to be achieved in images captured indifferent environments: indoor, in the wild outdoor environments, dimly lit rooms, in presence of harsh shadows, and various other noisy environments. Also, the face image can be partially occluded by external objects,such as scarf and sunglasses, or self-occluded due to pose. Moreover, it has to be computationally efficient, especially if large scale monitoring, or analysis of large databases is needed. These requirements combine to present an extremely challenging task for computer vision.

In the computer vision community, the face has been modeled using different approaches to detect the face landmarks. These models can be broadly divided into two main categories global based models and part based models [24, 25, 44, 46].

In the global models, the appearance of the face is described by the holistic appearance of the face using either parametric or non-parametric model. The active appearance

model (AAM) [26–29] is a prominent example of these approaches where the appearance of face is represented using a weighted linear combination of face basis, eigen faces. Detecting the facial point using AAM model tends to fail with illumination problem and bad initial for model parameters. Many extensions has been proposed for AAM [30–33, 35–37].

The part based model [38, 43–46] has recently attracted the attention of computer vision researchers. It describes the face as a collection of parts with connections between certain pairs of parts. The model is quite general, in the sense that it is independent of the specific scheme used to model the appearance of each part as well as the type of connections between parts. A natural way to express such a model is in terms of an undirected graph $G = (V, E)$, where the vertices $V = (v_1, \cdots, v_n)$ correspond to the n parts, and there is an edge $(v_i, v_j) \in E$ for each pair of connected parts v_i and v_j. An instance of the face is given by a configuration $L = (l_1, \cdots, l_n)$, where each li specifies the location of part v_i. Sometimes L is simply referred as the object location, but "configuration" emphasizes the part-based representation. The part based model has an advantage over global models since it robust for small pose and illumination variation. Active shape model [38–41] is the prominent example of these part based models. The part based models describe the shape prior using point distribution model which belongs to constellation family as shown in Figure 3. The appearance of each part is described using gradient. Detecting the facial landmarks using active shape model is suboptimal since the texture and shape are not combined together in the solution. The shape prior used as filter to constrain the output.

This dissertation proposes a new face model that incorporates advantages of the two families of global and part based models for detecting face landmarks. The proposed part based model is demonstrated using both theoretical and experimental results advantage to have advantages over existing models. Moreover, the proposed facial landmark detector has been a part of two different face analysis and understanding applications that are presented in this work which are face recognition and emotion recognition for children with autism spectrum disorder.

The proposed model incorporates the part based model with holistic face information. The The part based model is based on soft combining of a texture classifier with

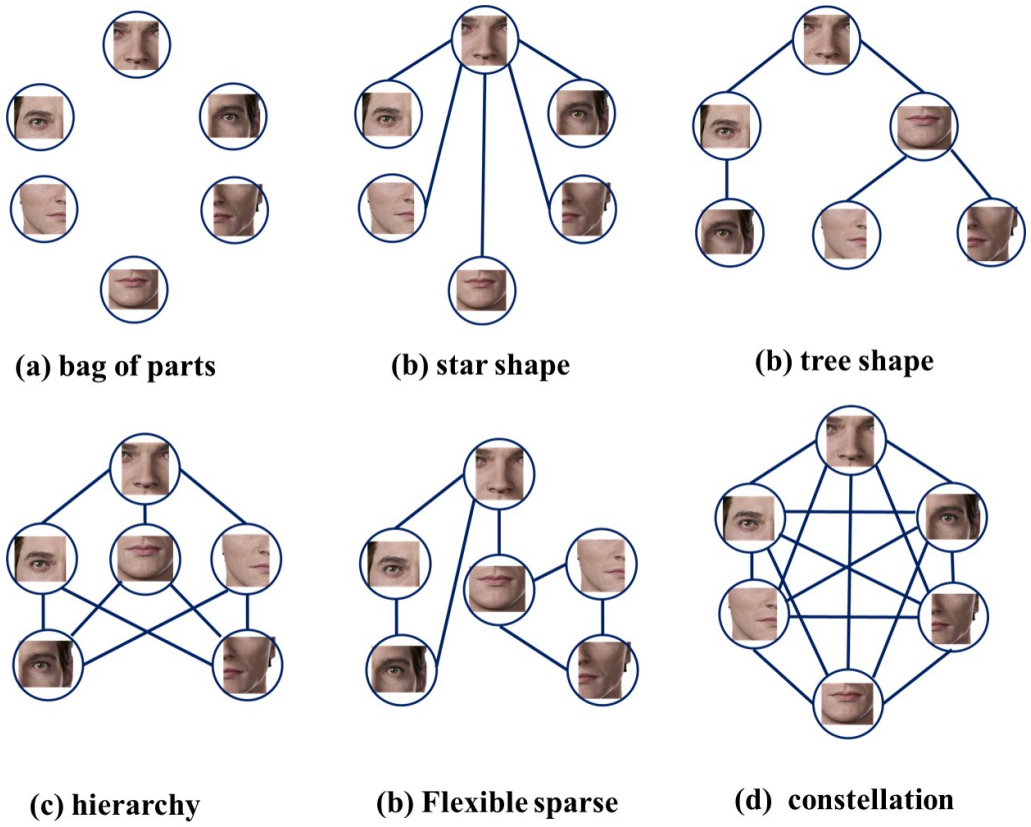
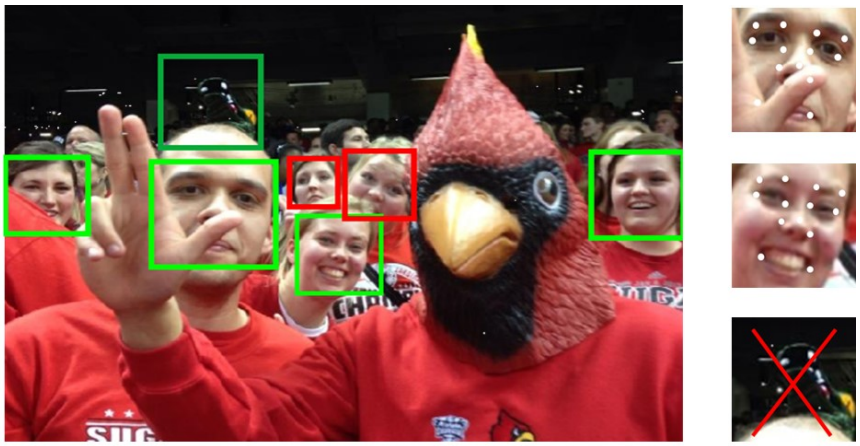


FIGURE 3: Graphical geometric models of shape constraint among face parts.

complex Bingham distribution as shape representation. The texture classifier is built by a support vector machine classifier that uses novel feature representation called pixel difference. The complex Bingham distribution is adapted from statistical community into computer vision for face shape representation since it is invariant to in-plane rotation giving this model superiority with respect to existing shape models. Energy minimization function is formulated to incorporate information from both texture classifier and shape models simultaneously. In the final stage, the global information is used to improve the results of the part based model by using regression model that does not penalize the outliers of human face shape due to extreme expression, occlusion, and different ethnicity.

The proposed facial landmarks detectors is used in two problems related to computer vision. The first one is camera steering in multi camera surveillance systems. One of main challenges in multi camera surveillance system is that a subject of interest in the filed of view of one of the cameras may not be in the field of view of other cameras. Facial



(a) (b)

FIGURE 4: Rejecting pseudo faces using the facial points detector (a) Face detection output where the undetected faces in red and detected faces in green. (b) Facial points detector declare the black object is not face

feature points detector aims to help in solution for steering other cameras to the subject of interest for unified anaylsis. The second application is rejecting pseudo faces for robust face detection. Face detection produces errors which are false positive and false negative. The false positive means the detector declares there is a face despite lack of facial presence while the false negative simply fails to detect the face. The proposed facial feature points detector can also reject false positive faces as shown in Figure 4.

2. Face Recognition in the wild

Face recognition is considered the standard face understanding and analysis application. The field of facial biometrics is vibrant and its applications cross various domains. Impressive theory and algorithms have been developed under each component during the past two decades; which have been explored rapidly in recent years with advances in machine learning, computing and availability of novel sources of facial information, e.g., social media. The face recognition pipeline usually consists of three main modules: face detection, face representation and face matching. Face detection is the first step in this

process since it segments the facial region from the background before further processing is performed. Face representation provides useful low-level information from face image. Face matching measures the similarity between two face representations to achieve one of these tasks: 1- Choose one of the gallery faces that matches a probe face which is called closed set face recognition. 2- Choose one of gallery faces that matches a probe face or identify there is no match which is called open set face recognition. 3-Indicate whether the probe face belong to certain person based on the gallery face which is face verification.

Despite the enormous successes of facial biometrics, still a fully adaptive and fast functional robust system for facial recognition in the wild is far from being achieved. The obstacles for achieving this goal stem from the uncertainties in modeling ”age”, ”pose”, ”illumination” and ”expression”, individually, and much more simultaneously.

Pose invariant face representation was identified as one of the prominent unsolved problems in the research of face recognition [3] and it has attracted great interest in the computer vision and pattern recognition research community. As the viewpoint varies, the 2D facial appearance will change because the human head has a complex non planar geometry. Magnitudes of variations of innate characteristics, which distinguish one face from another, are often smaller than magnitudes of image variations caused by pose variations. Directly matching and comparing two faces of different poses is quite difficult since a pose varies in 3-D space, but there is only the information of 2-D appearances in the face images. There is a strong connection between solving the pose problem and three dimensional construction of the human head.

This dissertation proposes two different approaches for solving the pose problem. The first one is called dynamic weighting of facial features [73]. In this approach, the similarity measure between the face signature of the probe image (query image) and face signature of gallery images is the sum of similarity measures of feature vectors of the patches around facial feature points. Since some facial feature can be partially occluded with head pose angle, a dynamic weight for these facial features is proposed. Dynamic weights are assigned for each facial feature at each pose based on the overlapping scores which is based on the number of pixels in the patch in the frontal gallery image and captured

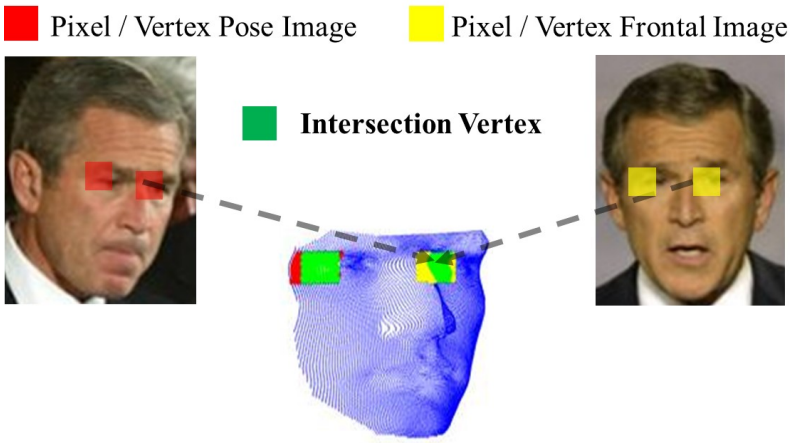


FIGURE 5: Corresponding vertices in the 3D face for the pixels in frontal which is indicated in yellow, and the pixels in pose image which is indicated in red. The green vertices correspond to the intersection vertices from red and yellow pixels

pose image that correspond to the same vertices in the 3D shape of the person, as shown in Figure 5.

The second approach is based on rendering face images at different poses for each subject from the enroll image [115]. The gallery in this approach consists of multiple images for the person at different poses that are generated from enroll image(s). Rendering face images needs information about 3D shape and texture for subject face. The information of texture is captured from the galley face image. The 3D shape is reconstructed using two algorithm. The first algorithm is statistical shape from shading where the input is single image [107]. While the second algorithm is stereo reconstruction where the input is two face images from two different camera and geometric information about the relation between two cameras in known. In this approach, the similarity measure between the face signature of the probe image (query image) and face signature of gallery images is the sum of similarity measures of feature vectors of the patches around facial feature points without weighting since both face images (query image and render gallery image) have approximately same pose angle.

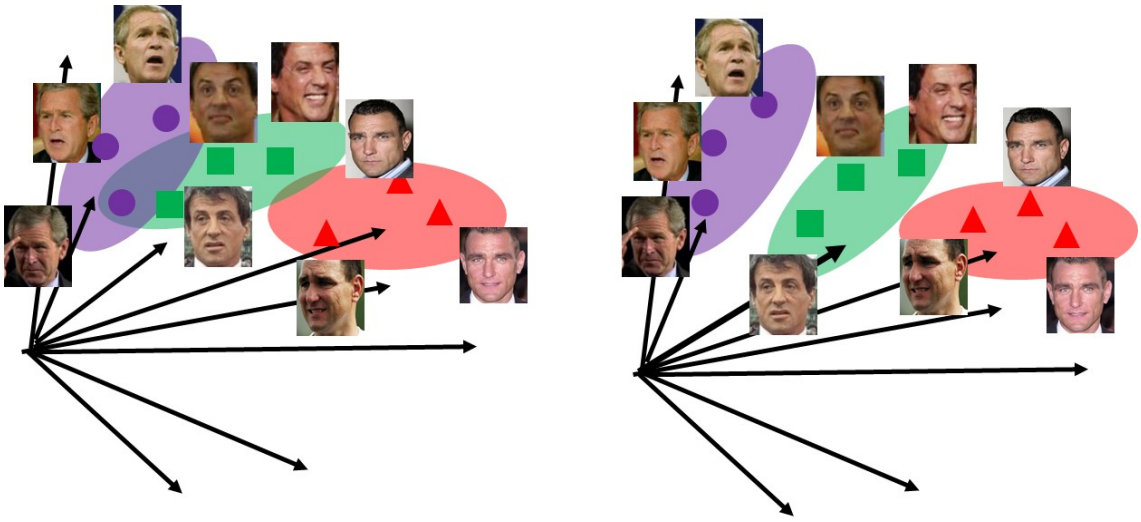
It is worth mentioning that face representation and face matching have the same goal. Face representation aims to convert the face image from pixel domain into a fea-

ture vector that invariant to intra-person variations such as pose. However, face matching aims to make distance between two face representations, two face feature vectors, which belong to the same identity, should relatively smaller than two face representations, two face feature vectors, which belong to different identity. Therefore, computing a similarity measure between a face representation and other face representation plays an important role in the success of face recognition. Standard distance measure i.e., Euclidean distance, treats all face representation features equally. However, certain image features could be more reliable than others. To overcome this drawback and to enhance the measure performance, prior information to discard bad features selectively in each individual matching circumstance should be used on computing the measure.

A novel similarity measure between two pose invariant face representations is proposed where the distance is small if the two face representation belong to same identity [129]. This similarity measure is based on a nonlinear combination of Mahalanobis distances which is determined by using equivalent constraints labeled data(the restricted setting). The proposed similarity measure maps data from its original feature space to a target space such that a simple distance will be adequate for the verification task. Original feature space is invariant to pose but it may be affected by many uncontrolled sources of variations e.g., changes in illumination, expression and camera properties. The target should be invariant to pose, illumination, and expression.

3. Steps towards Emotion Recognition for Children with Autism spectrum disorder

Emotion recognition is one of the most important pieces of information provide by face understanding. Emotions play an important role in human survival and adaptation as they affect the way people perceive their surroundings, interpret them, and act upon these perceptions. According to the national research council, children with Autism spectrum disorder (ASD) have major difficulties in expression and emotion recognition. They show a reduced verbal and nonverbal communication facility. In other words, they have a problem in revealing their expressions and emotions, and in understanding others emotions and



(a) Before similarity measure (b) After Similarity measure

FIGURE 6: Multiple face image representation under different capture conditions for different identities in feature space before and after similarity measure.

expressions. Therefore, it is very hard to understand their emotions based on gesture and facial expressions.

An affective model to understand the emotion state of children with ASD using the physiological signal is established using contact sensors [177]. The contact sensors limit the daily application for emotion recognition. Farag et al. [176] propose a solution for measuring cardiovascular activity and breathing rate using thermal infra-red camera. However, their approach is limited to laboratory environment. Moving this work into real life situation is challenging since there is need for continuously tracking and detecting superficial temporal arterial (STA) branches from thermal camera while the subject is moving, playing, and participating in social communication. The challenges in detection and tracking STA area are as follows: the size of STA branches area is small (e.g., around 20-30 pixels height and width), which makes it easy to confuse with many other areas in the face, and the appearance of STA changes over time in response to cardiovascular activity.

Long term tracking and detection framework is proposed. The proposed framework

consists of three main modules: (1) an adaptive particle filter tracker for (STA) branches area which is used to overcome continuous change in the appearance, (2) online detector that used a new concept which is called supporters to avoid confusion that results of small size of STA branches area, and (3) an integrated learning and decision making unit.

Moreover, a data set consist of thermal and visible videos of children with autism spectrum disorder and controlled at age 6-8 years old has been collected at Kentucky Autism Training center and Medical School Campus, and Down Syndrome of Louisville. These data has been collected over fifteen months. The data were collected while the subjects are playing games at different levels of difficulty.

1.2 Structure of the dissertation

A novel human face modeling method for detecting facial feature points in multi-modality imaging (thermal- visible) is presented in Chapter 2, since facial feature points detection is critical step in different applications of face understanding such as face recognition, emotion recognition, age recognition, and gender recognition.

The proposed facial feature points detector is used to solve two different problems related to face analysis in Chapter 3. The two applications are rejection of pseudo faces for robust face detection and steering the cameras in multi-camera network system.

Chapter 4 and Chapter 5 discuss a novel face recognition approach in the wild. This approach is invariant to pose, illumination, and expression. Chapter 4 presents several new approaches for extracting pose invariant face representations. While, Chapter 5 shows a novel similarity measure between two pose face representation which makes the distance between two pose face representation with same identity but with different expression and/or illumination is relatively smaller than the distance between two pose face representation with different identities.

Chapter 6 shows a framework for tracking and detection area of superficial temporal arterial (STA) branches in thermal imaging and nasal tissue area for measuring vital signs for children with Autism spectrum disorder (ASD). These measurements signals can be

used with skin conductance as strong indicator for emotion recognition for ASD children.

Chapter 7 concludes the dissertation with insights toward future work to be explored in the field.

CHAPTER 2
FACIAL FEATURE POINTS DETECTION

Facial feature points, also known as facial landmarks or facial fiducial points, have semantic significance. Facial feature points are mainly located around facial components such as eyes, mouth, nose and chin. Facial feature point detection (FFPD) refers to a supervised or semi-supervised process using abundant manually labeled images. FFPD usually starts from a rectangular bounding box returned by face detectors [7, 51] which implies the location of a face. This bounding box can be employed to initialize the positions of facial feature points. Figure 7-a shows the face detection output which is the input for facial points detector, and Figure 7-b shows the output of proposed facial points detector.

Facial feature points can be reduced to three types: points labeling parts of faces with application-dependent significance, such as the center of an eye or the sharp corners of a boundary; points labeling application-independent elements, such as the highest point on a face in a particular orientation, or curvature extrema (the highest point along the bridge of the nose); and points interpolated from points of the previous two types, such as points along the chin. According to various application scenarios, different numbers of facial fea-

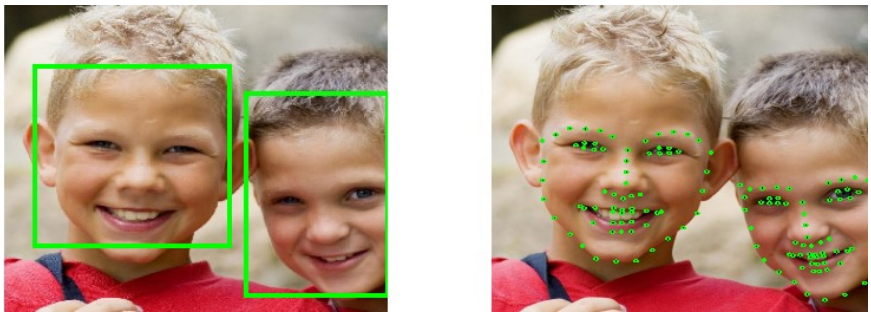


FIGURE 7: (a) Face detection output which is the input for facial feature points detector. (b) output of the proposed facial feature points detector

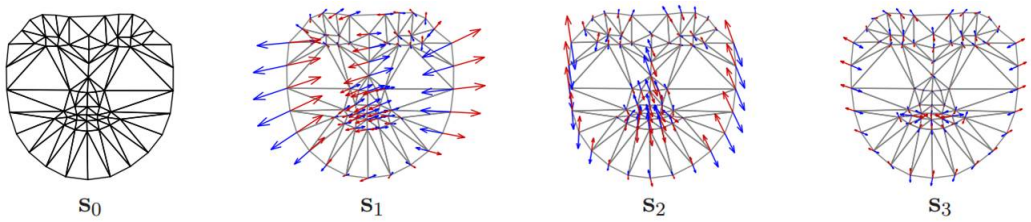


FIGURE 8: The linear shape model of an independent AAM. The model consists of a triangulated base mesh S_0 plus a linear combination of n shape vector S_i. The base mesh is shown on the left, and to the right are the first three shape vectors S_1, S_2, and S_3 overlaid on the base mesh. Image courtesy of Matthew and Baker [28]

ture points are labeled for example, a 17-point model, 29-point model or 68-point model. Whatever the number of points is, these points should cover several frequently-used areas: eyes, nose, and mouth. These areas carry the most important information for both discriminative and generative purposes. Generally speaking, more points indicate richer information, although it is more time-consuming to detect all the points.

2.1 Related Work

Detecting the shape of a facial image is a challenging problem due to both the rigid (scale, rotation, and translation) and non-rigid (such as facial expression variation) face deformation. Existing facial feature points detection methods can be grouped into three categories: constrained local model based methods, active appearance model based methods, and regression-based methods.

Active appearance model based methods model the appearance variation from a holistic perspective. In the training phase of these algorithms, principal component analysis (PCA) is applied to a set of labeled faces (manually annotated face) to model the intrinsic variation in shape, and texture. This results in a parameterized model that can represent large variations in shape and texture with a smaller set of parameters. Figure 8 and 9 show the set of models (eigenfaces, and eigenshapes) that results from applying PCA on shape and texture respectively. Eigenfaces should be free of shape variation and is referred

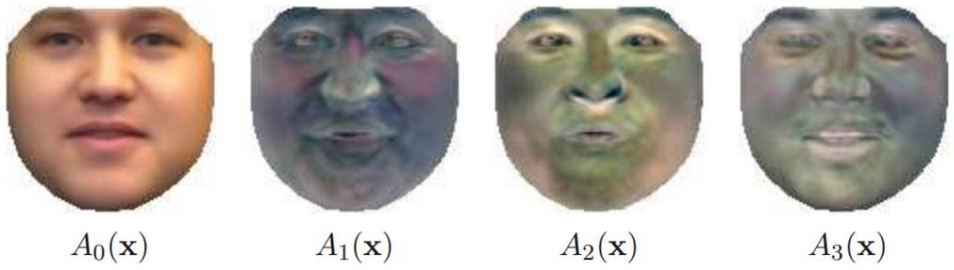


FIGURE 9: The linear appearance variation of AAM. The model consists of a base appearance image A_0 defined on the pixels inside the base mesh S_0 plus a m linear combination. The first three appearance images are shown A_1, A_2, A_3 and also defined on the same set of pixels. Image courtesy of Matthew and Baker [28]

to as shape-free textures. The coefficients of eigenshapes and eigenfaces can be used to synthesize different variations in shape and texture. The AAM algorithm aims to find the coefficient of eigenfaces, and eigenshapes model to minimize the difference between the texture as sampled from the testing image and the texture that synthesized by the model. The coefficients of model eigenfaces, and eigenshapes are defined over a high dimensional space, making it impossible to find its global maximum. Much work has been done to improve and extend AAM by Cootes [27]. Gao et al. [12] present a survey about the recent developments on AAM.

Regression-based methods estimate the shape directly from appearance without learning any shape model or appearance model. It learns a regression function which maps image appearance (feature) to the target output (shape). Zhou and Comaniciu [13] proposed a shape regression method based on boosting [14, 15]. Their method proceeds in two stages: first, the rigid parameters are found by casting the problem as an object detection problem which is solved by a boosting-based regression method; secondly, a regularized regression function is learned from perturbed training examples to predict the non-rigid shape. Haar-like features are fed to the non-rigid shape regressors. Cao et al. [16] proposed a two-level cascaded learning framework based on boosted regression [18]. Unlike the above method which learns the regression map for each individual facial feature, their

method directly learns a vectorial map that combine all landmarks. The main drawback of the regression methods is that they need huge amounts of memory as compared with the methods based on constrained local models and are very sensitive to the initialization.

In the constrained local model, the local texture and shape prior models are the main components. For the texture model, the local texture around a given facial feature is modeled, i.e., the pixels intensity in a small region around the feature point, while for the shape model, the relationship among facial features are modeled. Both models are learned from labeled images (labeled images).

Texture-based detectors aim to find the best suitable point in the face that matches the texture model. The texture model can be constructed using different descriptors such as Haar-like [8], local binary pattern (LBP) [5], Gabor [9], scale-invariant feature transform (SIFT) [10] features instead of using pixel intensity directly as a feature. The search problem can be formulated either as a regression or classification. For the classification problem, a sliding window runs through the image to determine if each pixel is a feature or non-feature. For the regression problem, the displacement vector from an initial point to the actual feature point is estimated.

Texture-based detectors are imperfect for many reasons; visual obstructions such as hair, glasses, and hands can greatly affect the results. The detection of each facial feature is also independent from others and it ignores the relation among these facial feature points. To overcome the disadvantages of texture-based detectors, constraints related to the relative location of facial features from each other can be established from the shape model. The shape model either is used to filter the output of texture model or they are combined together into single formula.

Cristinacce et al. [40] modeled the relative positions of facial features by a pairwise reinforcement of feature responses and the texture model around facial features points using PCA as in ASM. Valstar et al. [21] modeled shape using Markov Random Field (MRF) and the texture using Haar-like feature with boosting classifier. These two approaches use a single distribution for shape model, which is not suitable for modeling a wide range of poses and used the shape model to filter the output of texture model. Felzenszwalb et

al. [53] modeled the relation between facial featured in a graph tree where the relation between each two nodes is a gaussian distribution and the texture is modeled using iconic representation. Everingham et al. [19] extended the relation between facial feature points from a single Gaussian distribution into a mixture of Gaussian trees to handle different poses and used the Haar-like features with boosting instead of iconic representation to represent the appearance (texture) around facial feature points. Zhu et al. [46] built on [19] but they combined the texture and shape model into single formula and used HOG feature [54] to represent the texture around each facial feature points. Belhumeur et al. [44] used a non-parametric approach for shape modeling, using information from their large collection of diverse, labeled examples and represented the texture around each facial feature point using SIFT features. They used SVM [55] to classify each pixel as a candidate facial feature point or not. Their algorithm takes 17 seconds to detect 17 facial feature (1 second per feature).

2.2 Proposed Face Model

The proposed face model for detecting facial feature points combine advantages of the part based face model and the holistic face model. The face is modeled using part based model where the texture around facial points is modeled using pixel difference feature and complex Bingham distribution is used to model human face shape. The texture and shape model are combined together to detect human face shape using position of facial feature points in the image. The output is refined by a regression model that is built using non-parametric global information.

The following are the contributions of this work [22]: (a) the proposed pixel difference feature for modeling texture around facial feature point (b)using mixture of complex Bingham distributions to model human face shape from various viewpoints, (c) developing a new energy function for facial feature points detection combining two uncertainty terms related to (a) and (b), and (d) Incorporate non parametric holistic information to proposed face model for achieving few pixels accuracy.

The texture around facial feature point is represented by the difference in value of random pixel in the neighbourhood of this facial feature point. This new feature is the lowest computational complexity as compared with existing state of art while it has similar accuracy. The output of the classifier is regularized to handle false positives in the classification step of each pixel in a certain neighborhood as feature or non-feature. The output of the classifier should give a high response in the actual facial feature position and decrease smoothly going away from the actual position. If the neighborhood variance is low, it is certain that one pixel position is the actual feature point; otherwise, if the neighborhood variance is high, i.e., all classifier outputs in the neighborhood have combined high or low scores, the classifier is uncertain if a feature point exists in the area. Regularization is performed by dividing a normalization term to the classifier output related to the standard deviation of the output probability scores in the search neighborhood.

The complex Bingham distribution is more robust in modeling the joint probability of the location of facial features than existing models. Existing models need a preprocessing step before using the shape prior to filter out scale, translation, and rotation using least-squares approaches (e.g., Procrustes analysis), which can introduce errors to the system due to noise and outliers. Since the probability distribution function (PDF) of a complex Bingham has a symmetric property, there is no need to filter out rotation. Scale and translation can be easily removed by a simple mean and normalization step [48].

The facial feature point detection problem is formulated as energy minimization function that incorporates information from both texture and shape models simultaneously, while most of the state-of-the-art approaches use the shape model to filter the results of texture-based methods.

However, the parametric shape model helps in estimating the positions of facial points that construct face shape with avoiding outlier solutions. The parametric shape model has drawbacks since it penalizes the human face shapes that are far from the mean shape. Therefore, this work proposes adding a stage to refine the output that correspond to minimum energy by using regression model that estimate displacement to final face shape model, position of facial feature points. The regression model based on global texture to

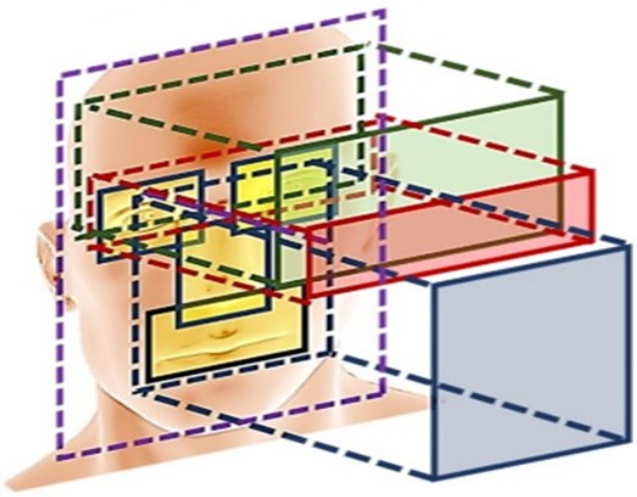


FIGURE 10: Searching space for locating candidate positions for different facial feature points.

give complementary information with local texture model in local texture model, first stage.

1. Local Texture Detector

Currently, a sliding window searching approach is the standard for object detection. In a sliding window approach, the image is divided into overlapped windows while trying to determine if there is an object in the present window or not. Deciding whether or not there is an object in this window requires two steps object representation and classification. Object representation aims to transform the pixel information from pixel domain into the feature domain where the object with different variation have close representation. Object classification aims to determine if certain representation is an object or not.

The seminal work of Viola and Jones [51] is considered the corner stone for many developments in the area of object detection. The object is represented by Haar-like features and adaboost is used for classification and feature selection.

The main idea behind the success of Haar-like feature with Adaboost classifier is the integral image. The integral image is an algorithm for a quick and efficient calculation for the sum of intensity values in a rectangular subset of an image. Calculation of Haar-like

features has constant time using the integral image.

Recently, many researchers [54] moved toward using histograms of gradient orientation for object representation since the histogram of a gradient orientation is invariant to illumination and small change in view point (affine transformation). However, they use support vector machine for classification since it shows significant improvement in many pattern recognition application such as voice recognition, hand writing recognition. The main drawback of using histogram of gradient orientation over Haar-like feature with adaboost is the computation time. Figure 10 shows histogram representation for some face parts.

In this work, two approaches have been compared for the application of facial feature points detection, face landmark detection. Moreover, a new representation is proposed that is as accurate as histogram of gradient orientation but with lower computational complexity. The feature is based on pixel difference at random positions.

In the facial feature points detection problem, each facial point is considered an object. Therefore, the sliding window approach is run over the face image n times where n is the number of facial points. However, the searching area can be limited to certain area relative to face detection which is smaller than the face region. The idea behind limiting the search area is removing ambiguity and speed up the running time for the algorithm.

Figure 10 shows the searching region for different facial feature points. The center of the searching area for facial point i is the mean position of feature point i in all training images after filtering translation, scale and rotation. The width and height of the searching area is based on the variance of the feature position from the mean.

Figure 11 shows the search region for tip of nose and left corner of right eye, their response map, and the candidate positions for each facial point using pixel difference feature representation. The response map represents the score at each pixel in the searching area is the tip of nose. The candidates are the peaks in corresponding response map. It estimated using non-maximal suppression technique.

The sliding window classifier for detecting facial point i scan the corresponding search area as shown in Figure 10 and the score is that the pixel at position z_i is the facial

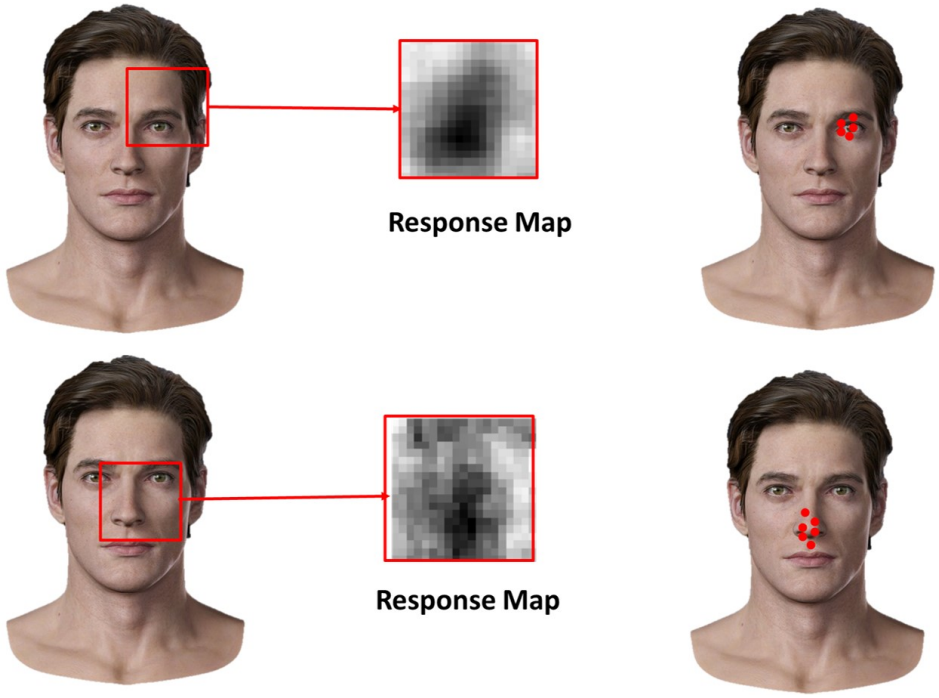


FIGURE 11: The score at each pixel in the searching area for the tip of nose and inner corner of right eye.

feature point $S(D_{z_i})$

Calculation the score is that the pixel at position z_i is the facial feature point $S(D_{z_i})$ is based on the classifier type. Support vector machine is used with histogram of orientated gradient and pixel difference feature representation, while adaboost classifier is used with Haar-like feature.

The score $S(D_{z_i})$ where the appearance of each facial point is represented by Haar-like feature within window w*w with adaboost classifier is given [22]

$$S(D_{z_i}) = \sum_{t=1}^{r} \alpha_{t_i} F_{t_i}(z_i) \tag{1}$$

where α_{t_i} is weight of weak classifier t for the feature i and F_{t_i} is the binary response of weak classifier.

On the other hand, the score $S(D_{z_i})$ where the appearance of each facial point is represented by either histogram of orientated gradient or pixel difference feature represen-

tation with support vector machine classifier is given by

$$S(D_{z_i}) = \sum_{t=1}^{r} \alpha_{t_i} \mathfrak{z}_{\mathfrak{d}\mathfrak{i}} \mathfrak{S}_{\mathfrak{i}} \tag{2}$$

where α_{t_i} is weight of each support vector t for the feature i, $\mathfrak{z}_{\mathfrak{d}\mathfrak{i}}$ is the extracted texture feature which is either random pixel difference or histogram of orientated gradient, $\mathfrak{S}_{\mathfrak{i}}$ are the support vectors [55].

In the case of a perfect texture-based detector, the response of classifier, response map, are homogenous as the probability of the pixel being feature is high at the true position and decreases smoothly going away from this position. Therefore, the output of classifier is regularized with a variance normalization factor by dividing the output probability of classifier with $\sigma_{\aleph(z)}$. $\sigma_{\aleph(z)}$ is the standard deviation of the output probability among the neighborhood $\aleph(z)$. Then, the probability of position z is feature i based on the texture detector $P(D_{z_i})$ can be written as

$$P(D_{z_i}) = \frac{K}{\sigma_{\aleph(z_i)}} S(D_{z_i}) \tag{3}$$

where K is the normalization constant.

Since for each facial feature, a sliding window classifier scan the corresponding search area, the output of each facial point texture detector can be considered independent from others. Therefore, the overall probability of $\mathbf{Z} = [z_1, z_2 ... z_N]$, the positions of N facial features based on the texture-based detector, is given by

$$P(D_{\mathbf{Z}}) = \prod_{i=1}^{N} P(D_{z_i}) \tag{4}$$

Figure 12, and 13 show the candidate positions for each facial feature point using histogram of oriented gradient and pixel difference feature representation respectively and the best shape $\mathbf{Z}$. The best shape is the facial feature points output where the only best candidate is chosen.

Figure 14 shows a comparison between different representation Haar-like feature, histogram of oriented gradient and pixel difference feature representation with respect to

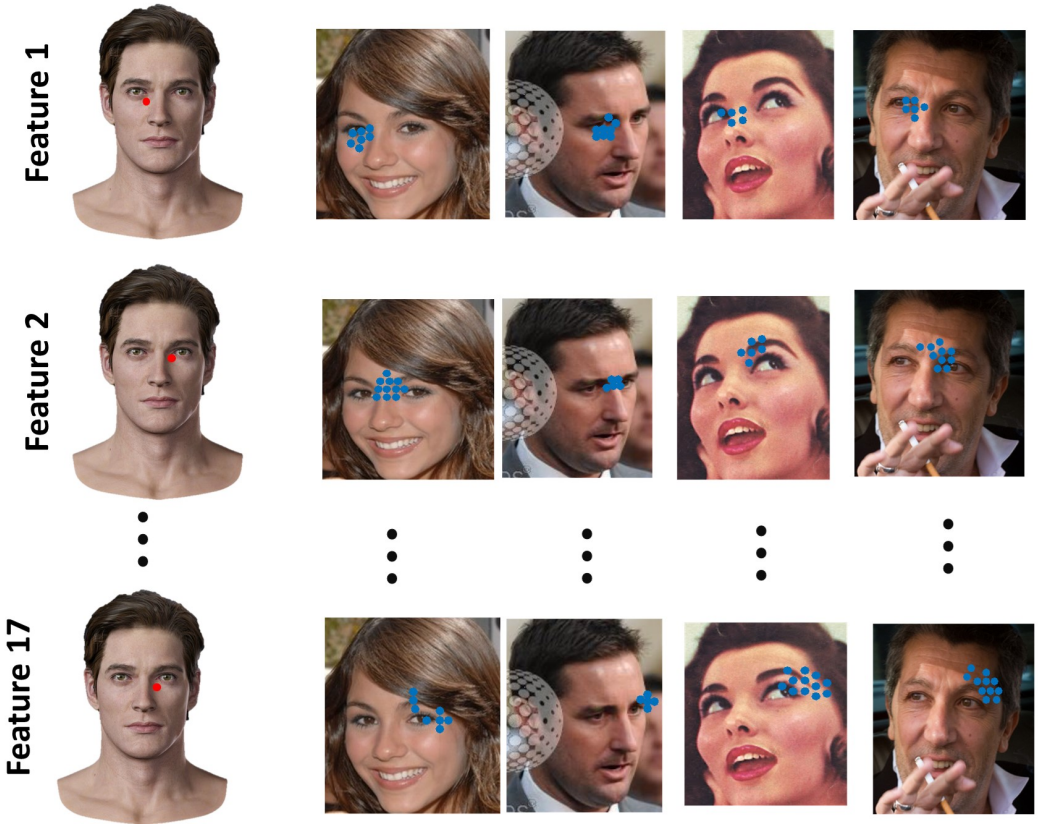


FIGURE 12: The candidate positions for different landmarks using Histogram of oriented gradient feature

detection accuracy. The histogram of oriented gradient and pixel difference feature representation shows similar detection accuracy. The main differences are: the window size that describe the facial point appearance in histogram of orientated gradient is half the window size in pixel difference feature representation. Increasing the window size captures more global information which may be needed but unfortunately the running time will increase dramatically using histogram of oriented gradient However it is constant with pixel difference feature representation.

It is worth describing some implementation details involving training. The face detection box in the training image is rescaled to 50×50. The patch size around a given facial feature position has been empirically determined to be 13×13 for optimum running time and accuracy. Positive samples are taken at manually annotated locations. Negative samples are taken at least 10 pixels away from the annotated locations. Figures 15, 16,

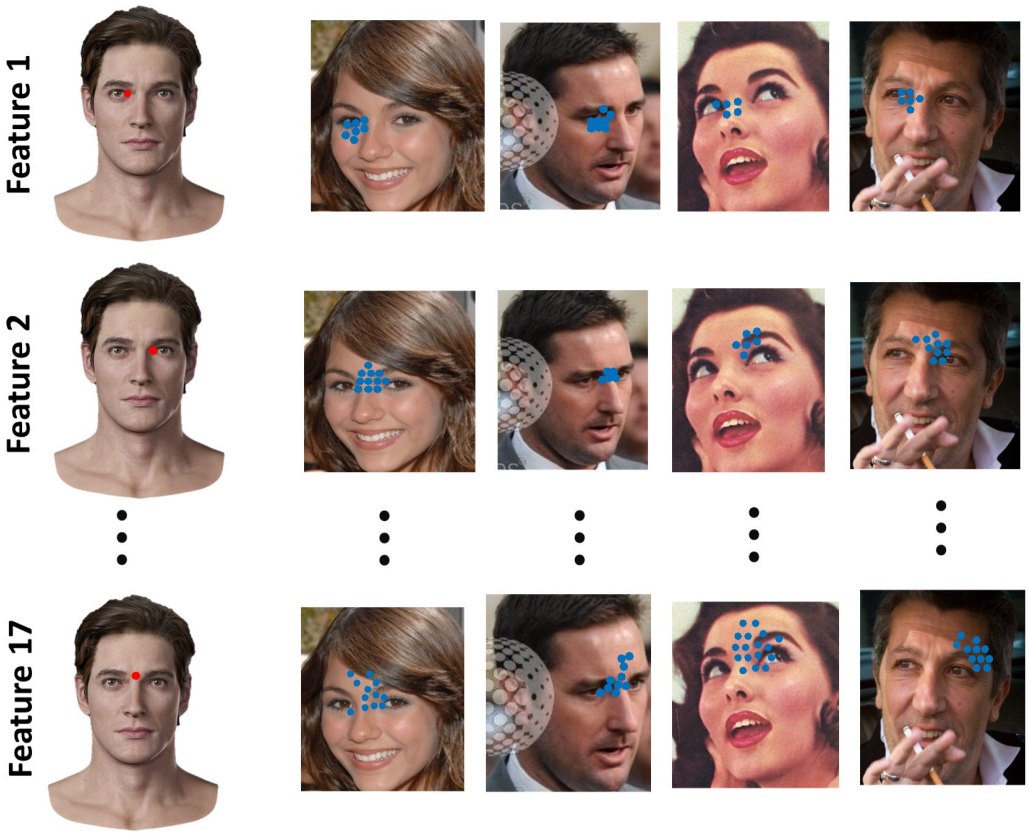


FIGURE 13: The candidate positions for different landmarks using pixel difference feature

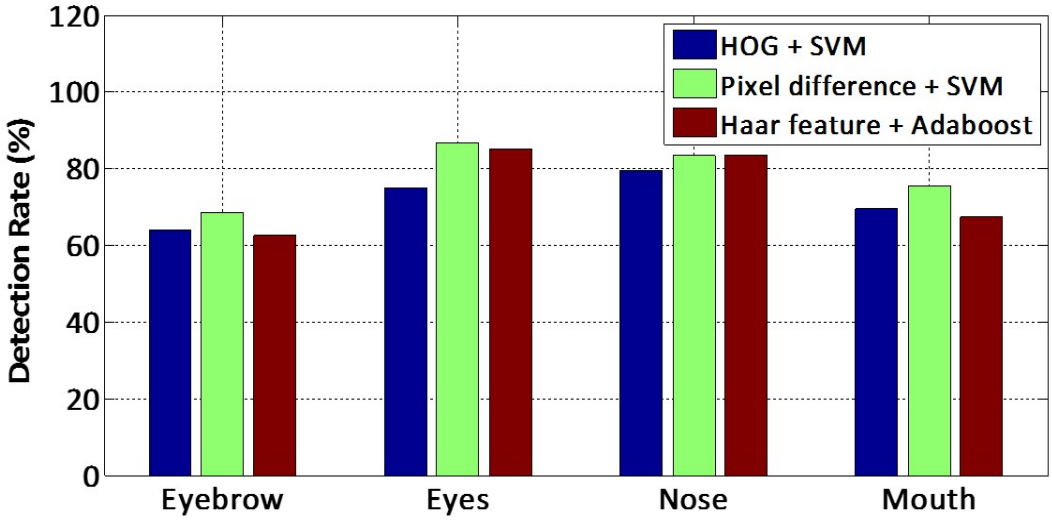


FIGURE 14: Detection rate comparison of three different appearance features: Histogram of oriented gradient feature, pixel difference feature, and Haar-like feature

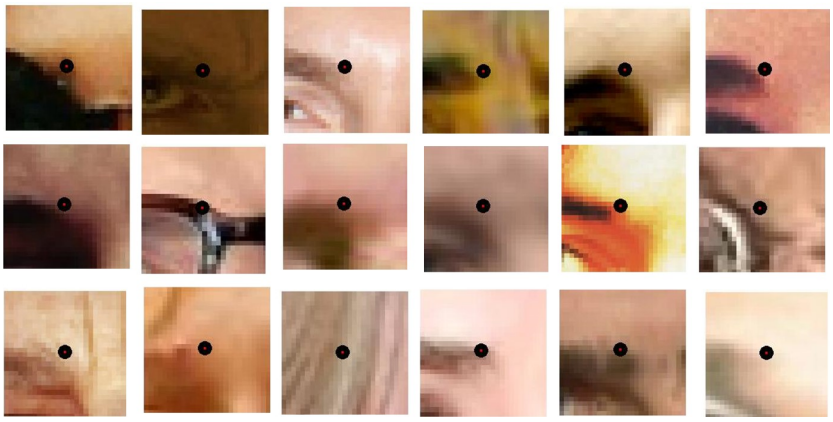


FIGURE 15: Illustration of intrinsic variation in the appearance of the eyebrow corner

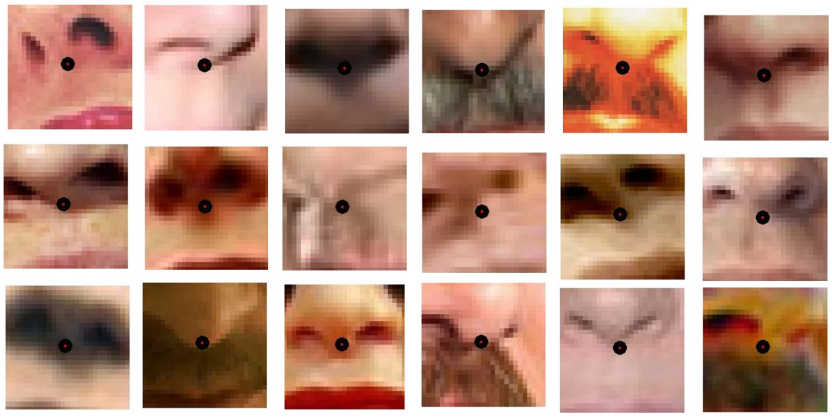


FIGURE 16: Illustration of intrinsic variation in the appearance of the nose tip

17 show subset of positive samples for the left corner of the eyebrow, the tip of the nose, and the left corner of the mouth. These figures show high variation within the appearance positive samples.

2. Shape Prior Model

Faces come in various shapes due to differences among people, pose, or facial expression of the subject. However, there are strong anatomical and geometric constraints that govern the layout of facial features. The representation of shape, i.e., joint distribution between facial feature points, is described by various models in the literature. The active shape model (ASM) is one example based on a single Gaussian distribution.

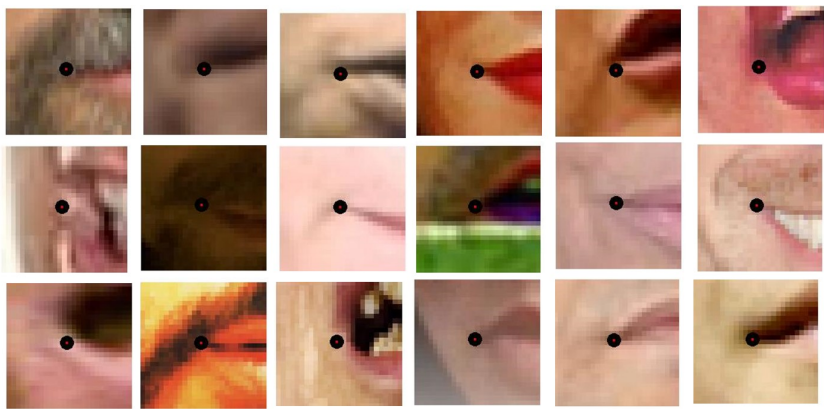


FIGURE 17: Illustration of intrinsic variation in the appearance of the mouth corner

Typically, one would like to have a shape representation that is invariant to translation, scale and rotation. A common way to address this problem is to use least-squares (LS) fitting methods, e.g., Procrustes analysis [48], where misalignments due to noise and outliers may happen [23]. Moreover, an iterative procedure is needed to align multiple shapes.

Complex Bingham distribution [48] for the proposed facial feature detection approach is used. The advantage of using this distribution is that shapes do not need to be aligned with respect to rotation parameters. The probability distribution function of the Complex Bingham is defined by

$$P(Y) = c(A)^{-1} exp(Y^* AY) \tag{5}$$

where $c(A)$ is a normalizing constant and Y^* is the complex conjugate of the transpose of Y.

Since the Complex Bingham distribution is invariant to rotation, it is suitable to represent shape in the pre-shape domain where shapes are zero-offset and unit-scale. In this work, the classical way of transforming is used to transform from the original shape vector to the pre-shape domain by simply multiplying to the original shape vector the (matrix) and then performing normalization [48]. H is given by

$$\begin{pmatrix} h_1 & -h_1 & 0 & \cdots & \cdots & 0 & 0 \\ h_2 & h_2 & -2h_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{N-1} & \cdots & \cdots & \cdots & \cdots & h_{N-1} & -(N-1)_{N-1} \end{pmatrix} \tag{6}$$

where

$$h_b = -(b^2 + b)^{-1/2}$$

Multiplying the original shape vector with the Helmert sub-matrix will project the original facial features position vector $Z \in C^n$ to C^{n-1}. Then, the shape representation using the complex Bingham is

$$P(Z) = c(A)^{-1} exp(\frac{HZ}{\| HZ \|}^* A \frac{HZ}{\| HZ \|}) \tag{7}$$

where A is a $(N-1) * (N-1)$ Hermitian parameter matrix, N is number of landmarks or facial feature points. The spectral decomposition can be written as $A = \mathbf{U}\Lambda\mathbf{U}^*$, where $\mathbf{U} = [U_1 U_2 \cdots U_{N-1}]$ is a matrix whose columns U_i correspond to the eigenvectors of A and $\Lambda = diag(\lambda_1, \cdots, \lambda_{N-1})$ is the diagonal matrix of corresponding eigenvalues.

The normalization constant $c(A)$ is given by

$$c(A) = 2\pi^{N-1} \sum_{i=1}^{N-1} a_i \exp(\lambda_i) \tag{8}$$

where $a_i^{-1} = \prod_{m \neq i}(\lambda_i - \lambda_m)$

The log likelihood of parameters is

$$L(\Lambda, \mathbf{U}) = \sum_{i=1}^{N-1} \lambda_i U_i^* S U_i - N \log c(A) \tag{9}$$

where the matrix S is a $N-1 \times N-1$ matrix denoting the auto correlation matrix for manually annotated shapes that have zero mean and unit scale. The maximum likelihood estimators are given by [48]

$$U_i = G_i \quad i = 1, 2, \cdots, N-1 \tag{10}$$

and the solution to

$$\frac{d \log c(\Lambda)}{d\lambda_i} = \frac{l_i}{N} \tag{11}$$

where $\mathbf{G} = [G_1 G_2 \cdots G_{N-1}]$ denotes the corresponding eigenvector of S and $L = diag(l_1, l_2 \cdots l_{N-1})$ is the diagonal matrix of corresponding eigenvalues.

Since no exact solution exists, λ is estimated by minimization of function F

$$F_i = \frac{d \log c(\Lambda)}{d\lambda_i} - \frac{N}{l_i} \tag{12}$$

This function is linearly approximated and solved iteratively using gradient descent. The update equation is given by

$$\lambda_i^{t+1} = \lambda_i^t - \kappa \frac{a_i + \lambda_i^t \sum_{m=1}^{N-2} \prod_{i \neq m \neq k} (\lambda_i - \lambda_k)}{\sum_{i=1}^{N-1} a_i \lambda_i^t} \tag{13}$$

Since the deformation of shape due to different poses is large and cannot be handled by a single distribution [19], [45], the training annotated shapes is divided into M classes. Each class carries a small range of poses and a Bayesian classifier rule is used to estimate the class of testing shape. The index of class is given by

$$m^* = \arg_m \min \frac{HZ}{\| HZ \|}^* A_m \frac{HZ}{\| HZ \|} + \log(c_m(A)) \tag{14}$$

3. Combining Texture and Shape Model

In facial feature detection problems, a numbers of hidden variables (position of facial features) is estimated based on observable variables (image gray level). This problem can be formulated as a Bayesian framework of maximum a-posteriori (MAP) estimation.

The probability model of the input image and the facial feature positions is given by the joint distribution, $P(I, Z) = P(I|Z)P(Z)$, where $P(I|Z)$ is the conditional distribution of the original image given the facial feature positions and $P(\mathbf{Z})$ is the distribution of the facial feature positions, i.e., the shape prior model. The maximum-a-posteriori estimate of facial feature positions given the image I is expressed as

$$Z^* = argmaxP(I|Z)P(Z) \tag{15}$$

The problem of facial features detection is formulated in Bayesian framework of maximum-a-posteriori. The goal is to find the vector Z, which maximizes the response probability for the texture model and shape model.

$$\hat{Z} = \arg\max P(I|Z)P(Z). \tag{16}$$

$P(I|Z)$ represents the probability of similarity between the texture of the face to off-line model given the facial feature vector which is given by Equation 17

$$P(I|Z) = P(D_Z) = \prod_{i=1}^{N} P(D_{z_i}) \tag{17}$$

Therefore, the maximum-a-posteriori estimate of facial features can be formulated as energy minimization of function $E(Z)$

$$E(Z) = -\frac{HZ^*AHZ}{\| HZ \|^2} - \sum_{i=1}^{N} \log P(D_{Z_i}) \tag{18}$$

This energy function is non-linear and not amenable to gradient descent-type algorithms. It is solved by a classical energy minimization technique, which is simulated annealing.

2.3 Non-parametric global information for detection refinement

Random fern regression is used to find the displacement from the position of detected facial points that corresponding to minimum energy as shown in the previous section to more accurate position with few pixels accuracy. The regression model learns the relation between the appearance around these detected points corresponding to minimum energy and displacement to the ground truth position of facial feature points, face shape. A single regression model is not sufficient since the relation is very complex, therefore boosted regression model is used.

In boosted random ferns regression, T random fern regressors($\mathfrak{I}^1$,$\mathfrak{I}^2$,...,$\mathfrak{I}^T$) are combined in an additive manner. Given a face image I and detected facial feature points,

face shape, corresponding to minimum energy Z^0, each random ferns computes a shape increment δZ from the appearance representation around these points and updates the detected facial feature points in a cascaded manner:

$$Z^t = Z^{t-1} + \Im^t(I, Z^{t-1}), \qquad t = 1, 2, ..., T, \tag{19}$$

where Z^{t-1} is the position of facial feature points that is the output of previous random fern regressor stage, while Z^t is the output of the current random fern regressor stage.

Each fern is learned by minimizing the sum of alignment error in the training set. Alignment error is difference between the detected positions for facial feature points, face shape, and the ground truth positions of facial feature points. In the training stage, the regression function $\Im$, is learned which minimizes the alignment error that is estimated from as following

$$\Im^t = \arg\min \|\hat{Z} - (Z^{t-1} + \Im^t(I, Z^{t-1})\| \tag{20}$$

where $\hat{Z}$ is the ground truth positions of facial feature points which is manually annotated.

The regression function in each fern is estimated by dividing the training data into b bins based on face appearance represented by pixel difference feature. Each bin is associated with regression output δZ^t that minimizes alignment error of the training samples falling into this bin. Figure 18 shows an illustration for testing fern using pixel difference feature.

Dividing the data into b bins to build regressor fern requires selection of $\log_2 b$ features and their threshold from B pixel difference feature that represent the holistic appearance of the face image. A good fern should satisfy two properties: (1) each feature in the fern should be highly discriminative to the regression target; (2) correlation between features should be low so they are complementary when composed. To find features satisfying such properties, correlation-based feature selection method is used [16]

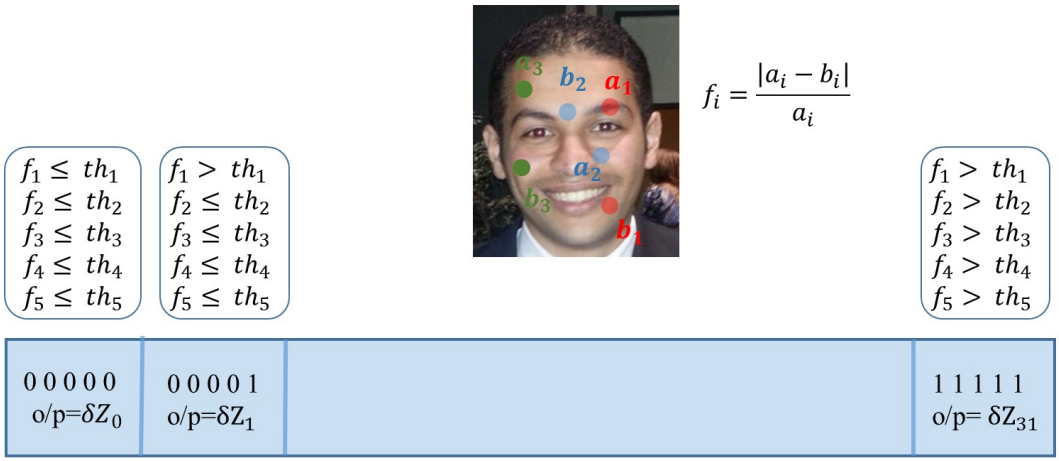


FIGURE 18: Illustration of regression output using pixel difference feature in random tree regression

1. Project the regression target to a random direction to produce a scalar where the regression target is the difference between position of current facial feature points and the ground truth.

2. Among B features, select a feature with highest correlation to the scalar.

3. Repeat steps (1) and (2) $\log_2 b$ times to obtain $\log_2 b$ features.

4. Construct a fern by $\log_2 b$ features with random thresholds.

The random projection serves two purpose. The first purpose is preserving proximity such that the features correlated to the projection are also discriminative to delta shape. The second purpose is the selected features are likely to be complementary.

2.4 Experiments

The proposed model for facial feature detection is evaluated in visible [22] and thermal images [52]. The same procedure is applied for thermal and visible. The only difference is the number of facial features in thermal image is chosen to be six points instead of sixty eight points in visible images. In the thermal image, most facial points cannot even be detected manually since the iris is hardly visible and there is no contrast with the sclera. The eyebrows are not consistently visible since this depends on their density. Also, the

lips are in many cases undistinguishable and therefore the mouth is hardly distinguishable if it is closed. Therefore, the number of detected facial points in thermal is small which are around $3-6$ points and most of the algorithms for detection facial points in visible images cannot be used in thermal images. The facial feature points detector is evaluated using cumulative distribution of the relative error. The relative error is distance between the detected facial feature point and manual annotated point (ground truth) divided by the ground truth distance between the two eyes. At every point in the curve, the x-axis shows the relative error, and the y-axis is the percentage of facial feature points that have relative error less than or equal the value of x-axis.

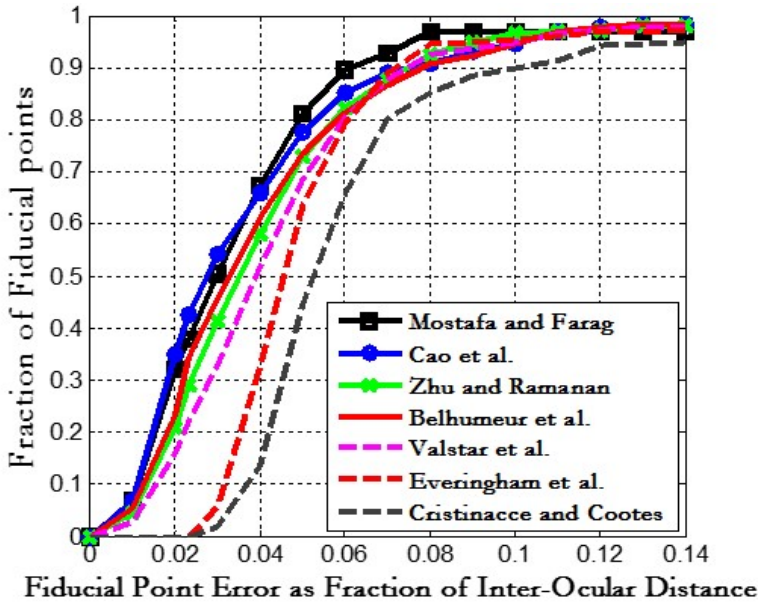


FIGURE 19: A comparison of the cumulative error distribution measured on BIO-ID dataset.

1. Visible Images

The performance of the introduced facial features detector is evaluated on BIO-ID dataset, Labeled Face Parts in the Wild (LFPW) dataset [44], and Helen dataset [66]. Most of the researchers about facial features detection in the literature reported results on the BIO-ID database, therefore it is included here as a testing dataset. The BIO-ID

dataset contains 1521 images, each showing a near frontal view of a face in controlled indoor environments with no illumination and occlusion problems for 23 distinct subjects. On the other hand, Belhumeur et al. [44] released LFPW as a challenging uncontrolled dataset. It consists of 1432 faces from images collected from the web. The dataset contains different challenges pose, existence of shadow, presence of occlusion objects as sunglasses or subject's hand, existence of in-plane rotation, and blurred images. Recently, Vuong et al. [66] released Helen dataset consisting of 2,330 faces in 2,330 high resolution images collected from Flickr with a broad range of appearance variations.

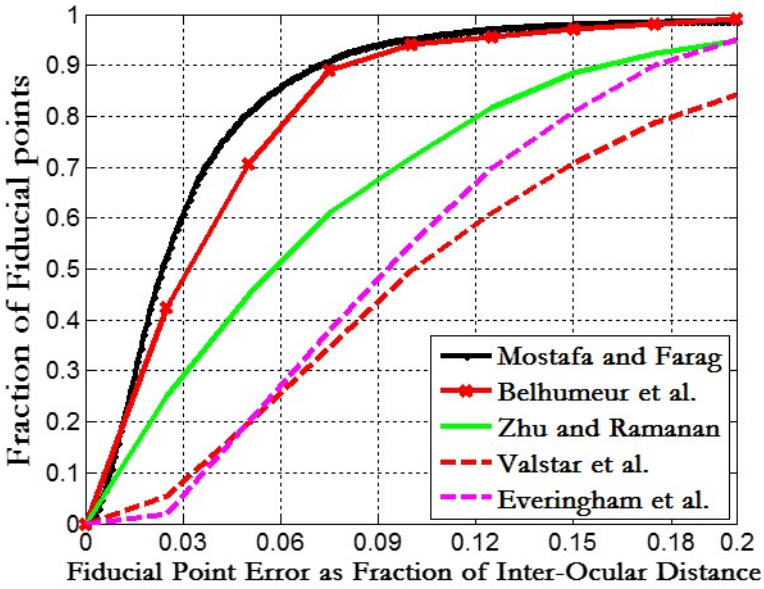


FIGURE 20: A comparison of the cumulative error distribution measured on LFPW dataset.

Figure 19, and Figure 20 show the cumulative error distribution for the proposed detector versus those reported by [154], [46], [44], [21], [19], and [40] on BIO-ID, and LFPW dataset respectively. The introduced detector and detectors in [16], [46], [44], [21], and [19] have comparable performance on BIO-ID since this database includes images with near frontal view of a face in controlled indoor environments with no illumination and occlusion problems. In LFPW database, the introduced detector and detectors by Belhumeur et al. [44] showed similar performance. This performance has the highest accuracy

compared with other approaches. Cao et al. [16], and Burgos et al. [17] reported their performance using mean error as percentage of interocular distance instead cumulative error distribution. Figure 21 (a,b) show the mean error as percentage of interocular distance for Belhumeur et al. [44], Cao et al. [16], Burgos et al. [17], and the proposed approach on Helen , and LFPW dataset, while Figure 21 (c,d) show the other two factors in comparison which are model size, and running time.

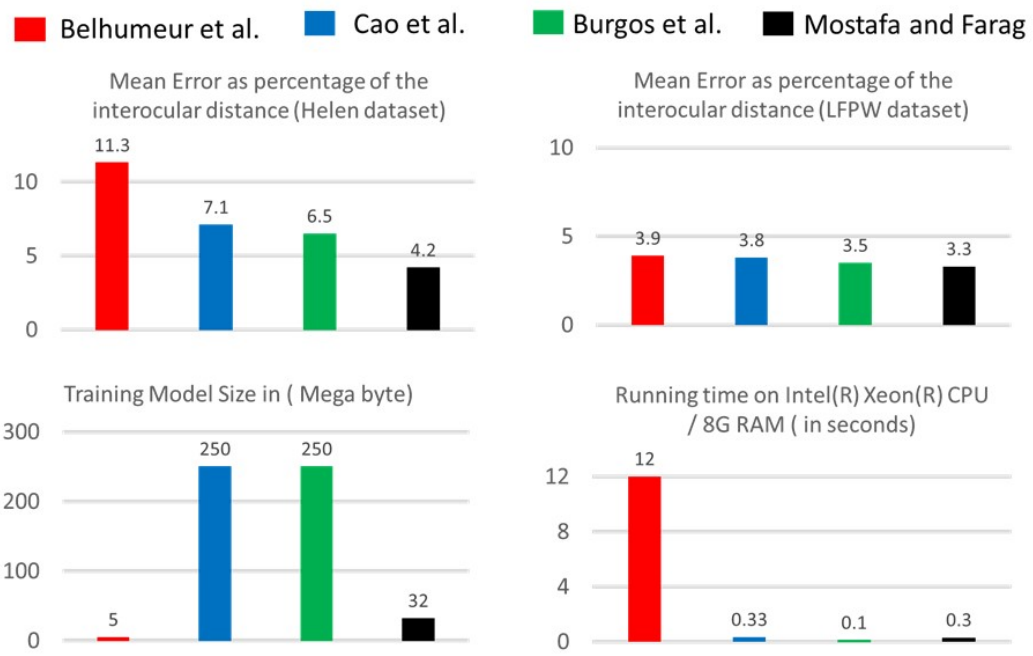


FIGURE 21: A comparison of proposed detector against the state-of-art according to accuracy, running time, and memory usage.

Detector model by Belhumeur et al. [44] takes a long time since it is based on SIFT feature which is extensive feature in extraction. The model utilizes more memory as compared with most of the existing algorithms since it needs to save shapes instead of the parameters of the shape model. While the proposed shape model is a mix between parametric and non-parametric shape models. The parametric shape model with local texture detector is used for finding approximate positions for facial feature points which it does not need a lot of memory. On the other hand, regression random ferns are used to find accurate facial feature points starting from the output of the parametric model. This stage needs

more memory but is still less demanding than detectors by either Cao et al. [16] or Burgos et al. [17]. Their detectors need many stages of cascade regression trees to achieve good results since they start with random initialization which may be far from the true position.

FIGURE 22: Samples of results of the proposed facial feature detector on Labeled Faces Parts in the Wild (LFPW) dataset.

Figure 24 shows the accuracy of each stage in the proposed detector using three experiments. The first one is denoted as local texture detector where the pixel difference feature with support vector machine is used to find best candidate for each facial feature point without any shape constraint. The second experiment is denoted as parametric shape model. In this experiment, the facial points are detected based on optimized energy function that combines complex Bingham distribution for shape modeling with texture model. The last one is the fine tuning stage based on the ferns regression model for shape relaxation

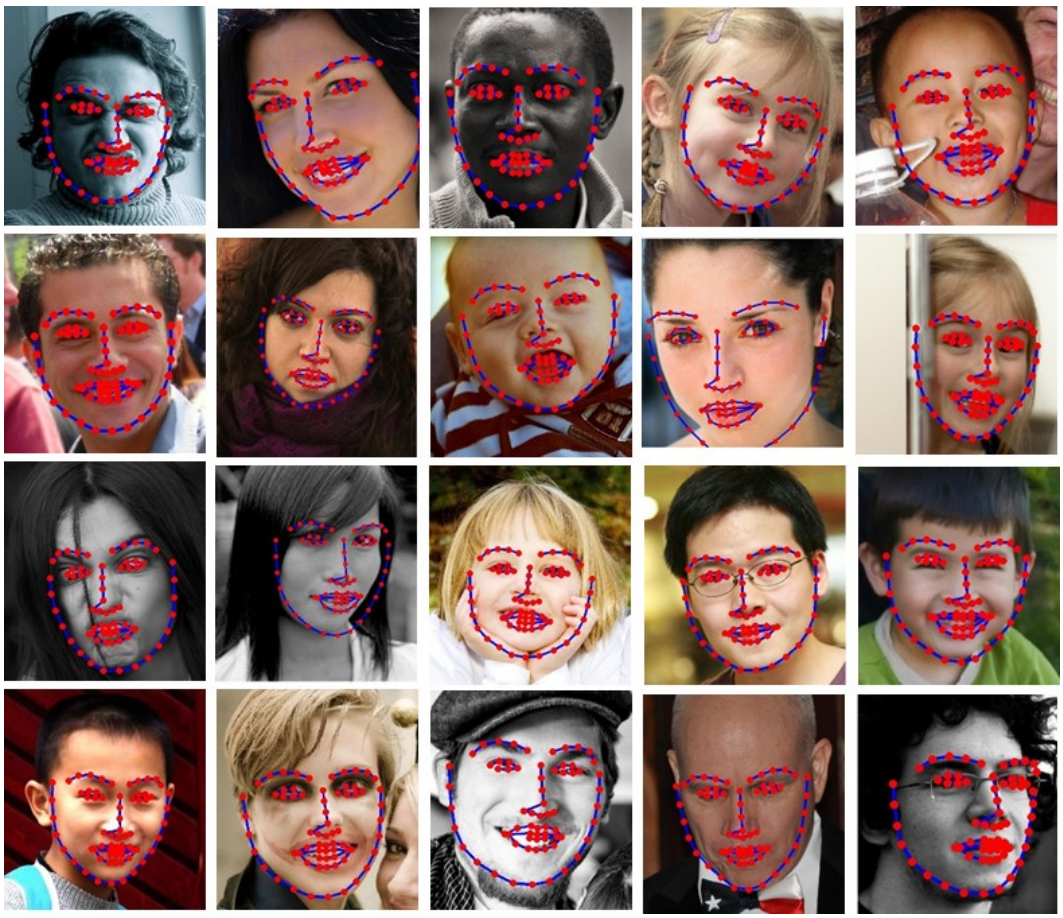


FIGURE 23: Samples of results of the proposed facial feature detector on Helen dataset.

and incorporates global information.

2. Thermal Images

The performance of the newly facial features detector is evaluated on a subset of the UND dataset. This subset consists of 328 thermal images. This subset contains 82 subjects with four different images (i.e., different illuminations and expressions). Figure 25 shows the cumulative error distribution for the proposed detector compared to [49, 50]. Trujillo et al. [49] detected the two eyes and mouth for expression recognition. Their method is based on applying the Harries algorithm to extract critical points in the face thermal image. Then k-means clustering is performed under the assumption that the cluster will be coincident with the facial component. Martinez et al. [50] used Haar features and the GentleBoost

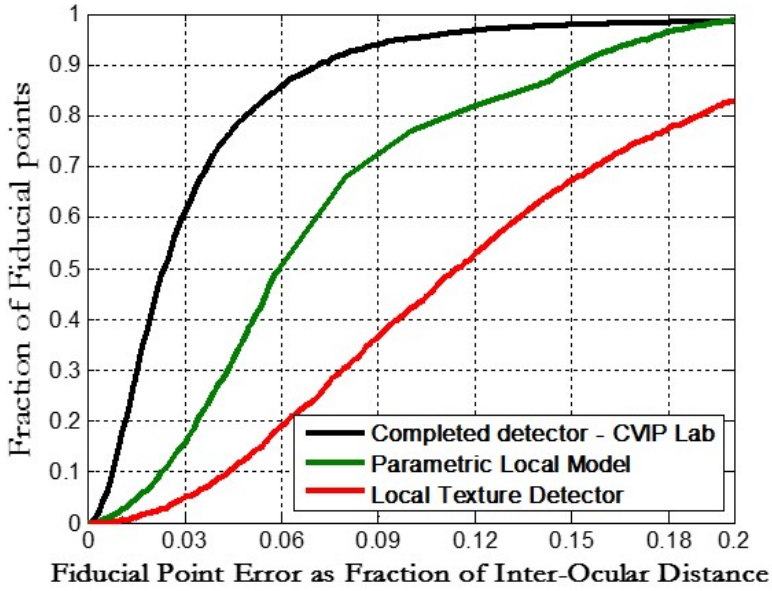


FIGURE 24: Effects of each component in the proposed approach: local texture detect only, local texture detector with shape constraint, and the full proposed approach.

algorithm to detect the two eyes and the nostrils. The classifier has many false outputs because the search for a feature is done in the whole image. These outliers have been filtered by using Gaussian distribution as a shape constraint.

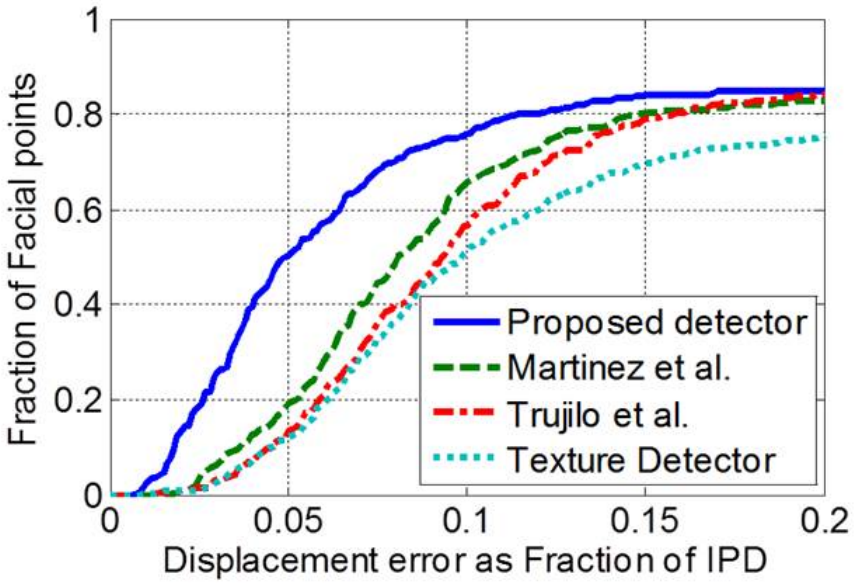


FIGURE 25: A comparison of the cumulative error distribution measured on Notre Dame dataset.

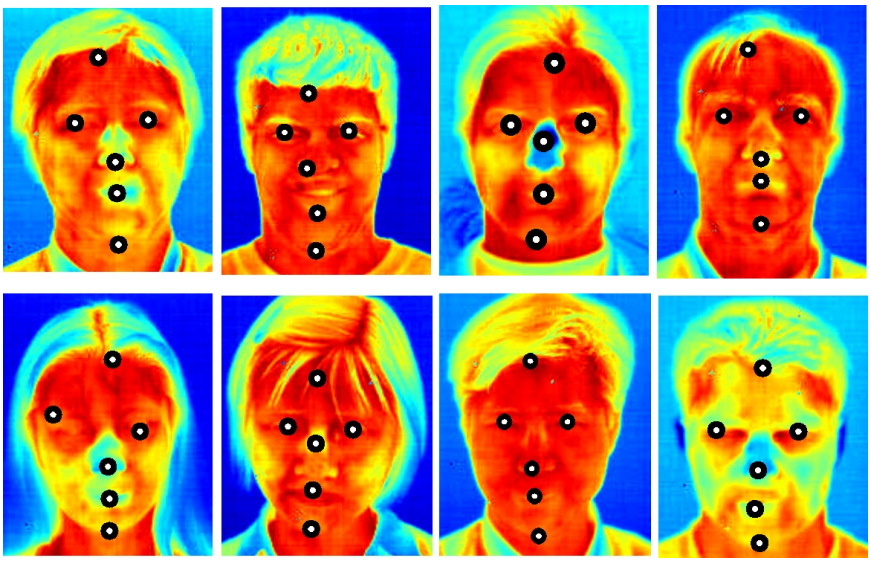


FIGURE 26: Samples of the results of the proposed facial features detector on Notre Dame dataset (Thermal Imaging)

CHAPTER 3
FACIAL FEATURE POINTS DETECTOR: APPLICATIONS

Facial Feature points are important components in many face analysis and understanding application such as face recognition, emotion recognition, and gender recognition. This chapter focuses on two new applications for facial feature points that are not related to recognition task which are camera steering in multi camera surveillance systems and rejecting pseudo faces for robust face detection.

Multi camera surveillance systems exist in many transit stations, shopping stores, grocery, parks, private and government buildings, and many streets all over the world. In the last decade, the main application for multi camera surveillance systems is monitoring and recoding. Due to advances in technology and higher security demands, there is a noticeable upsurge in application of biometrics to be a part of multi-camera surveillance system. There is a particular interest in biometric systems which are capable of acquiring multi view images for integrated surveillance/identity tasks since active cooperation from the target may not be required. Fixing many cameras to the same scene for capturing multi-view images is very expensive. Therefore, each camera is usually mounted on pan/tilt unit that allows the camera to rotate in 3D space. When a suspicious action is happening in the field of view of the camera, the other cameras should be steered to the same action/subject. Figure 27 shows an illustrated example using two cameras where a suspicious subject is detected by one of the cameras while this subject is not in the field of view of the second camera. The goal is to automatically steer the second camera such that it captures the subject of interest. A passive approach (i.e., no need for active sensing devices) without adding an extra wide field view camera for solving steering problem is discussed in this chapter. This approach for steering the other cameras is based on utilizing information from the detected facial feature points.

FIGURE 27: A multi-NFOV camera surveillance system: the cameras are constantly moving to cover the whole area (1st row). Once a suspicious subject is detected by one camera (2nd row, left). The other camera can be imaging a completely different area (2nd row, middle). The goal is to steer this other camera to get the same target subject in its field-of-view (2nd row, right).

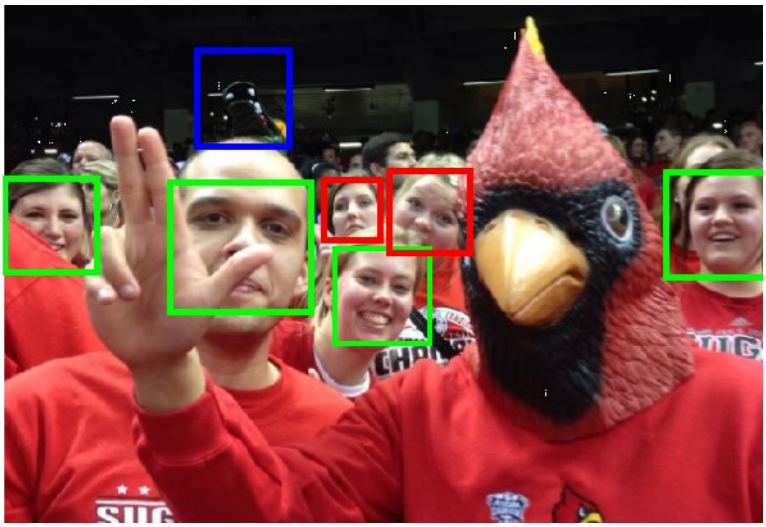


FIGURE 28: Illustration of face detection errors. The red rectangular is false negative (undetected faces) while the blue rectangular is false positive (pseudo faces)

Rejection of pseudo faces for robust face detection is the second application for the proposed facial feature points detector. Face detection can be defined as the automatic process of isolating faces from their background. Figure 28 shows examples of the two error types in face detection: missing detection and false positive detection. There is a strong connection between these two errors since decreasing one error leads to an increase in the other error. This chapter presents an approach for decreasing the false positives without increasing the missed detections. The approach is based on utilizing information from facial feature points detector. The facial feature points detector extracts the facial features for all face candidates resulting from the face detector regardless of the nature of these candidates being true positives or false alarms but it generates a probability that can be used to reject false alarms.

3.1 Camera steering in multi camera surveillance system

1. Problem statement and Related work

Considerable effort has been made on management of camera network in surveillance applications. Prior work in this field focusing on the efforts related to the system configuration and how one camera or more is/are steered to a particular subject given the subject is captured by another camera for the purpose of facial image capturing and recognition will be reviewed.

Stillman et al. [118] developed a system for person recognition consisting of two static overlapped WFOV cameras and two NFOV cameras. The two overlapped WFOV cameras are used to determine 3D location in real world coordinates of the person using triangulation, then the two NFOV cameras are steered based on the calculated 3D position. Similarly, Hamppaur et al. [128] and Wheeler et al. [119] proposed a system from the point of view of how to locate a subject in 3D and they differ in the tracking method and how they detect the subject in WFOV cameras. Krahnstoever et al. increase the number of WFOV cameras from two to four for more coverage area and accordingly increasing the accuracy of locating a subject in 3D world coordinates. All these systems have one or more NFOV cameras that are steered based on the location of a subject in world coordinates to capture the facial image that is used for recognition.

Greiffenhagen et al. [117] developed a system based on only one static overhead WFOV camera mounted below the ceiling to capture the 3D location of a subject. They used in their algorithm the person's foot since the z-component of the person's foot is known. Given the distance between the camera position and the floor, the captured image from WFOV can be used to estimate the x-y component. The system is restricted for indoor application. Zhou et al. [127] used one WFOV camera for outdoor application to locate the subject. They used a regression model between the position of a subject (in pixels) in the image, which is captured from the WFOV, and the camera control parameters (pan and tilt angle) of the NFOV cameras. Marchesotti et al. [122] also used one WFOV camera. They

restricted the starting point of tracking and steering the camera at a certain distance, to coincide with the gate of the parking area. These restrictions help to find the 3D location of a subject in world coordinates using the image captured from a WFOV camera.

Elders' group [120, 124] developed an approach for 2D localization in the image plane based on a combination of the likelihood of three cues: 2-frame motion difference, background subtraction, and skin detection. No modality alone is sufficient. In their work, they did not explain how a subject is located in the 3D world coordinates and how NFOV cameras parameters, pan and tilt angle, are evaluated. Their system is also intended for indoor coverage.

Rother et al. [121] have developed a 3D prior for scene learning from a single view. They use the average height of the person to locate a subject in world coordinates. However, since the statistics of people's height has rather a high variance, the uncertainty of estimation depth is accordingly high. In addition, this approach would encounter difficulties when the subject's complete height is not visible in the image (e.g. subject sitting, bending, or even partially occluded). As a result the distance estimate can be way off.

2. Proposed approach

To solve the steering problem, a passive approach (no need for active sensing devices) is proposed . Unlike other approaches, the proposed approach does not need WFOV. In fact, the proposed approach uses facial biometric measures which are statistically more consistent. The contribution of the proposed approach is the use of human face biometric measures to infer an approximate estimate of the subjects distance to a camera. In particular, inter-pupil distance (IPD) of a human face is used for that purpose. The IPD has much more consistent statistics across different people with much smaller variance. The IPD statistics is shown in Table-1, and has similar mean across the male and female population, as well as having a low standard deviation. Therefore, IPD can lead to more accurate estimate of the subject distance. In addition, the distance from the mid-point between eyes to outer edge of the lips (ELD) is used. This additional biometric has a well-known statistic,

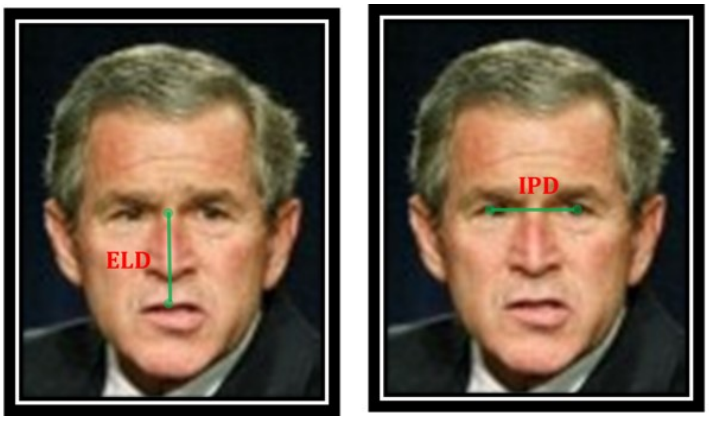


FIGURE 29: Biometric distances used in the proposed approach.

and more importantly is quite robust to subject pose. Hence, the subject's distance can be estimated away from the camera even if she/he is only visible from the side view (profile), sitting or even with occluded body. The distance from the mid-point between eyes to outer edge of the lips is approximately 1.2 times of inter-pupil distance.

Although, the proposed approach is applicable for multi-camera systems, its description in this section is restricted to a two-camera system. This approach can be summarized in three steps. First, from the statistics of distances IPD and ELD in centimeter and the measured distances in the image plane of one of the cameras in pixels, the approximate distance of the subject to this camera (reference) can be estimated. Then, given the distance of the subject to the reference camera, a trigonometric is used to estimate the required pan of the second camera in order to automatically steer it to the vicinity of the target. Afterwards, since the two cameras are assumed to have equal elevation above the mount platform, the tilt angle of the second camera is set equal to that of the reference camera. After the movement, two or more subjects may appear in the field of view of the steered camera, see for example the right image in the 2nd row of Figure 27. Therefore, a quick search for the target subject in the captured image of the second cameras is carried out using matched filter to localize the face of the target subject.

2..1 Distance from subject to the reference camera To estimate the distance of the target subject to the reference camera, assume that the camera is modeled as pinhole

TABLE 1: IPD values (mm) from 1988 Army Survey [126]

Gender	Population size	Mean	Std.	Min.	Max.
Male	1771	64.7	3.7	52	78
Female	2205	62.3	3.6	52	76

camera model, where the world coordinates system coincidences with the camera coordinates system originated at the camera's focal point. As such, any 2D point $\mathbf{x} = [x\ y\ 1]^\top$ in the image plane is related to the corresponding 3D point $\mathbf{X} = [X\ Y\ Z\ 1]^\top$ via $\mathbf{x} = K[I|\mathbf{0}]\mathbf{X}$, where K is the camera matrix that encompasses all the camera intrinsic parameters, such as the horizontal and vertical scales k_x and k_y. For two horizontal 3D points ($Y_1 = Y_2$) on a plane vertical at the camera optical axis (i.e., same distance from the camera: $\zeta_1 = \zeta_2 = \zeta$), their projections on the image plane will be displaced by

$$\Delta x = \frac{k_x}{\zeta}(X_2 - X_1) = \frac{k_x}{\zeta}\Delta X, \tag{21}$$

where Δx is the disparity between the two points in pixels. From this relationship, the distance from the reference camera to the target subject ζ can be estimated if there is two nearly horizontal points that are on the same fronto-parallel plane, for which the disparity Δx and the metric 3D distance ΔX are known. In this work, these two points are the center of the pupils of the eyes (and the distance between them is the inter-pupillary distance and denoted by D). The two points have the same distance to camera and can be considered on a fronto-parallel plane. The corresponding pixel disparity Δx, denoted by d, can be easily computed from the previous step of facial features extraction. k_x is easily determined beforehand from a camera calibration process done off-line (or known from camera specs). Based on the notation, $d = \frac{k_x}{\zeta}D$.

Similarly, for the two vertical points on the same fronto-parallel plane approximating the eye to lips distance (ELD), the vertical distance between them M is related to the their disparity in the image plane m via $m = \frac{k_y}{\zeta}M$), where k_y is often equal to k_x, i.e., $k_x = k_y = k$. From the known statistics on D and M, and the measurements of d and m from

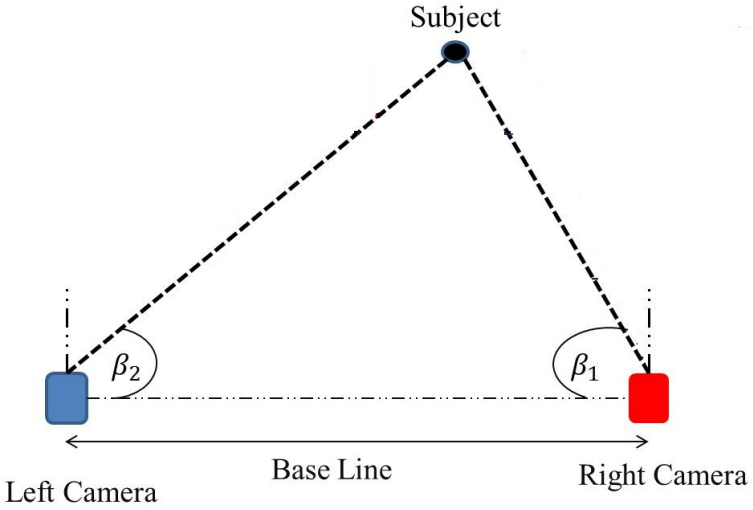


FIGURE 30: The setup geometry of the two cameras. The reference camera on the right is fixated on a target at a distance ζ with a pan angle β_1. The target is at a distance ζ_2 of the second camera on the left, which will be panned with an angle β_2. The base distance between cameras is B.

the extracted features, a least mean square estimate of the distance ζ can be found from

$$\zeta = \frac{k(d\ D + m\ M)}{d^2 + m^2}. \tag{22}$$

2..2 Steering the second camera After solving for the estimated distance (ζ) between the reference camera and the target subject from the previous step, the next stage is finding the parameters necessary to direct the second camera's pan angle β_2 such that the target subject should be in the field of view of the second camera. It is a trigonometry problem, as illustrated in Figure 30. The base distance B between the two cameras are assumed known.

Given the pan angle of the reference camera β_1, the estimated distance between the subject and the second camera can be found using the law of cosines

$$\zeta_2^2 = \zeta^2 + B^2 - 2B\zeta\cos(\beta_1). \tag{23}$$

Then, the pan angle β_2 of the second camera can be estimated using sine law

$$\sin(\beta_2) = \frac{\zeta}{\zeta_2}\sin(\beta_1) \tag{24}$$

From the previous equations, the error in pan angle of the second camera depends on the distance ζ, baseline distance B, and the pan angle of the reference camera. Now, it is interest to estimate the error in this angle due to the localization error in d and m from facial features extraction and the uncertainty in the two face biometrics M and D. To do this, the Jacobian matrix J of β_2 need to be estimated with respect to all these variables,

$$
\begin{aligned}
J &= \left[\frac{\partial\beta_2}{\partial D} \;\; \frac{\partial\beta_2}{\partial\Delta d} \;\; \frac{\partial\beta_2}{\partial M} \;\; \frac{\partial\beta_2}{\partial\Delta m}\right] \\
&= \begin{bmatrix}
\frac{k\sin(\beta_1)(\zeta B\cos(\beta_1)-\zeta^2+\zeta_2^2)}{\Delta x\zeta_2^2\sqrt{\zeta_2^2-\zeta^2\sin^2(\beta_1)}} \\
\frac{k\Delta X\sin(\beta_1)(\zeta^2-\zeta B\cos(\beta_1)-\zeta_2^2)}{(\Delta x)^2\zeta_2^2\sqrt{\zeta_2^2-\zeta^2\sin^2(\beta_1)}} \\
\frac{k\Delta Y\sin(\beta_1)(\zeta^2-\zeta B\cos(\beta_1)-\zeta_2^2)}{(\Delta y)^2\zeta_2^2\sqrt{\zeta_2^2-\zeta^2\sin^2(\beta_1)}} \\
\frac{k\Delta Y\sin(\beta_1)(\zeta^2-\zeta B\cos(\beta_1)-\zeta_2^2)}{(\Delta y)^2\zeta_2^2\sqrt{\zeta_2^2-\zeta^2\sin^2(\beta_1)}}
\end{bmatrix}^\top .
\end{aligned}
$$

After analyzing the proposed detector results, the localization error in d can be expressed as $N(0, 2^2)$, and the uncertainty in D (IPD) can be expressed as $N(63.5, 3.7^2)$ (from Table 1). One can propagate the uncertainty to the error in β_2 via $J\Sigma J^\top$ computed at the mean values of the variables with

$\Sigma = diag(\sigma_{\mathrm{d}}^2, \sigma_{\mathrm{D}}^2, \sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2)$, where $\sigma_{\mathrm{d}} = 2$, $\sigma_{\mathrm{D}} = 3.7$, $\sigma_{\mathrm{m}} = 2.8$ and $\sigma_{\mathrm{M}} = 4.44$

Figure 31 shows the pan angles of second camera β_2 and its error statistics for different pan angles of the reference camera at different distances and a baseline distance 7.5 meters.

3. Results and Discussion

The performance of steering the camera is evaluated using the collected dataset (UoFL-EWA). The pan and tilt angles of camera (1) and baseline distance are known and the objective is to estimate of the pan of camera (2). The mean, standard deviation and the maximum difference between the estimated angle and the ground truth is summarized in Table 2 at different poses.

The ground truth is defined when the subject's face is centered in the frame of steered camera. If the difference is less than half the field of view angle, the subject will

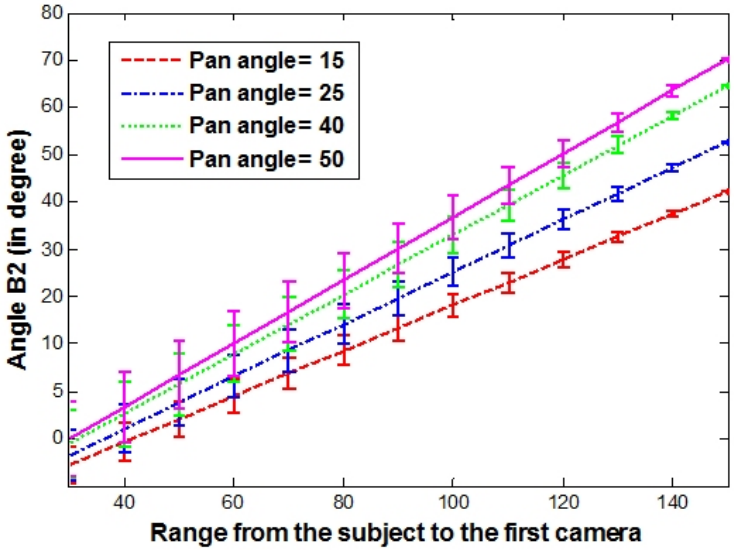


FIGURE 31: The pan angle of left camera (in degrees) given the baseline distance is 7.5 meters at different right camera pan angles of 15°, 25°, 40°, and 50°.

Distance	Maximum Difference			Mean Difference			Standard deviation of difference		
	Near Frontal	25°	45°	Near Frontal	25°	45°	Near Frontal	25°	45°
50*m*	1.31	1.52	1.80	0.62	0.69	0.72	0.45	0.73	0.82
80*m*	1.35	1.42	1.54	0.52	0.55	0.52	0.43	0.49	0.51
100*m*	0.77	1.33	1.34	0.36	0.40	0.41	0.23	0.28	0.32
150*m*	0.65	1.27	1.37	0.35	0.38	0.42	0.21	0.27	0.35

TABLE 2: The maximum, mean, and standard deviation of the difference between estimated pan angle of camera(2) and ground truth in degree

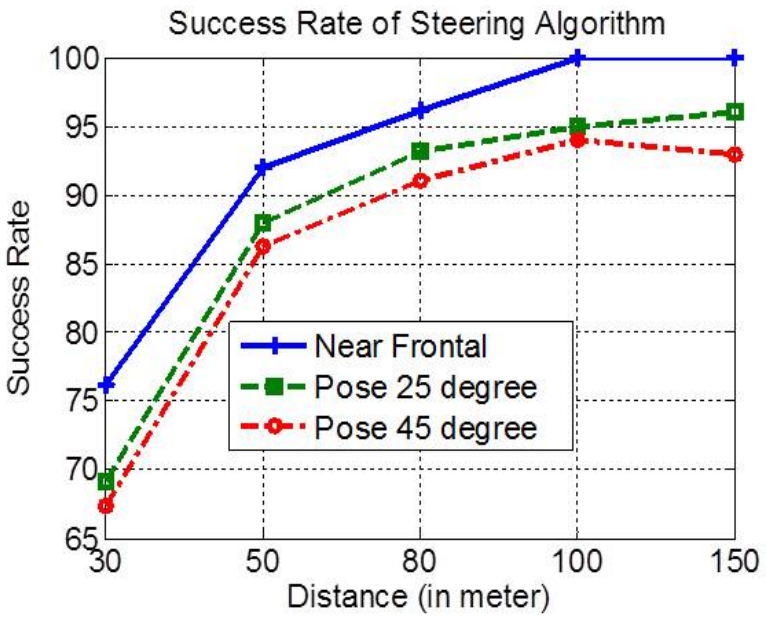


FIGURE 32: The success rate of steering algorithm at different ranges with different poses near frontal, 25°, and 45°.

appear in the image. Therefore, the matched filter can find the exact position in the image. If any part of target subject's face appears in the image plane of second camera, then the steering algorithm is successful. In other words, the steering is successful if the difference between estimated and ground truth pan angle is less than half the field of view angle. Figure 32 shows the success rate of the steering approach for subjects at different distances and have different poses. The lower success rate at distance (30m) is due to the fact that at this distance any small deviation in estimation the 3D location of the subject leads to high deviation in the estimated pan angle from the ground truth. So the target subject's face disappears from the image plane of the second camera completely. This can be overcome by a heuristic search in three different images captured at the estimated camera position, one step forward and backward from the estimated angle. The step angle is the field of view angle.

Figure 33 shows samples of the results of the steering approach in field test. The first column shows the scene as captured by the reference camera. The reference camera adjusts itself to make the subject in the middle of FOV. The second column shows the scene as captured by second camera before steering. The last column shows the scene after

steering the camera to the subject where the target's face subject is appeared in the image plane.

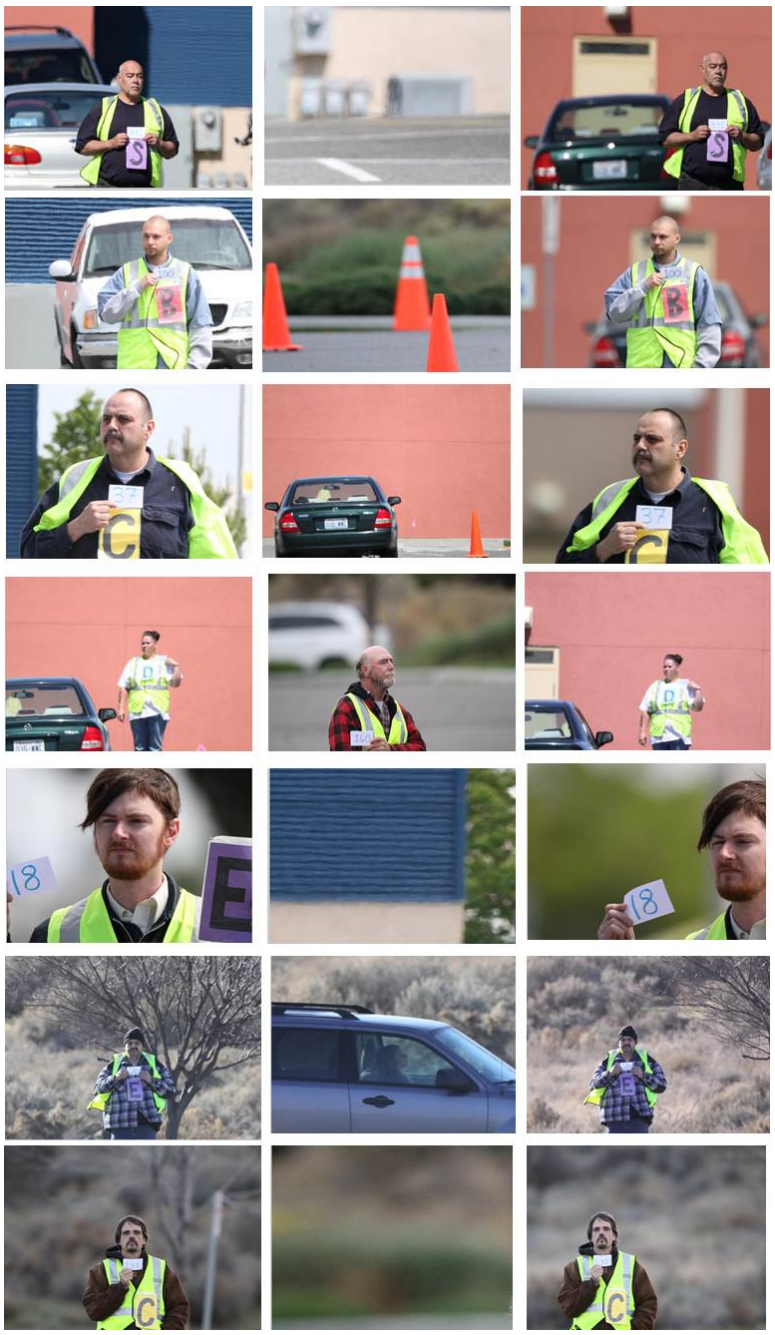


FIGURE 33: A sample result of the algorithm for steering a second camera to a subject given a single image for the subject from the first camera indoor at range 5 meters.(a) The left camera image with the subject in its FOV. (b) Locating the facial feature of the subject of interest. (c) The right camera image after steering, using our proposed algorithm. A bounded region is marked on the subject of interest, using the matched filter results.

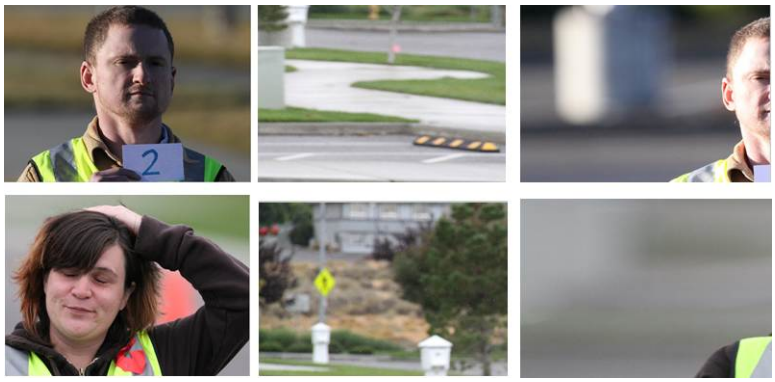


FIGURE 34: A sample of failure of proposed algorithm. The first column shows the scene as captured by first camera. The second column shows the scene as captured by second camera before steering. The last column shows the scene after steering.

3.2 Rejecting Pseudo faces for Robust Face Detection

1. Related Work

The input for a face detector is simply a digital image that may have multiple faces, a single face or even no faces at all. The output of the detector is the location and extent of each face in the image if any [60]. Despite the maturity of face detection algorithms, the problem is still open-ended in uncontrolled environments. There are several challenges in face detection their combined contribution makes the problem even difficult. Some of these problems include partial occlusion (by sunglasses or hair), pose (in-plane or out-of-plane rotation), expression (laughing, smiling), illumination, skin color, and out-of-focus faces. The two errors in face detection are false positive and false negative errors. The False Negative (FN) rate counts the number of faces in the image that was not detected and the False Positive (FP) rate counts the number of false detections in the image where the detector labeled a wrong region in the image as a face. Recently, a database for face detection in unconstrained settings called Face Detection Database and Benchmark (FDDB) was released [65] with the performance of several common algorithms in the face detection literature, e.g., Viola-Jones (VJ) [51], Mikolajczyk [68], and Kienzle et al. [71] detectors.

The results of these algorithms show that the current algorithms suffer from high false positive and false negative rates. Jain et al. [67] showed significant improvement by online adaption of the trained VJ cascade classifiers for decreasing the false negative rate. However, their detector is based on the assumption that there is more than one face in the image. On the other hand, Erden et al. [57] proposed utilizing the color information of the image by using skin color detector for false reduction after VJ detector. Many approaches utilize the skin with VJ detector as pre-filtering to limit the search space of VJ detector.

2. System Description

The input to the proposed framework is a digital image. This image is processed by a face detector based on Haar cascades in this experiment. The detector gives us all face candidates in the image which contains false alarms. Each face candidate passes through a facial feature detector that detects facial feature points. Moreover, it also gives the probability of the candidate being a false alarm. This probability is based on the texture around each detected facial feature point and the relation among the facial features. In the meantime, the colored version of each face candidate passes also through a skin detector that gives the probability of being a false alarm utilizing the complimentary information in color images. Finally, the probability of false alarm from the two detectors are combined for deciding face or non-face. If the image is not colored, the decision will only depend on the facial features. These blocks will be illustrated in details in the following subsections. This proposed framework aims to decrease the false positive numbers while it does not affect the false negative error.

2..1 Facial Features Points probability OpenCV implementation of face detector is used which can be replaced by any other face detector without affecting the proposed framework. The facial feature points, landmarks, has been detected using the proposed detector in the previous chapter. It is important to realize that the landmark detector assigns a probability for each landmark using the texture model given by $P(W(Z_i)|Z_i)$ that describes how much texture around this point is similar to off-line model. It also has another proba-

bility given by $P(Z)$ that describes its relation to other features. The combined probability of each landmark is given by $P(W(Z_i)|Z_i)P(Z)$.

The facial feature points detector detects the facial features for all face candidates resulting from the face detector regardless of the nature of these candidates being true positives or false alarms. Although the proposed framework gives a solution for the location of landmarks in both true positive and false alarm cases, the combined probability of the landmark has a low value in false alarm cases which can be used in rejecting false alarms. The probability of the candidate face being false alarm based on facial feature detector is given by

$$P(f_f) = 1 - \prod_{i=1}^{n} P(W(Z_i)|Z_i)\ P(Z_{m^*}) \tag{25}$$

The advantage of using the facial landmarks in rejecting false alarms is that it is accurate and it adds no overhead time to the system since the stage of landmark detection is usually required after face detection for further processing.

2..2 Skin Detection Probability Skin segmentation in color images can be summarized into choosing the suitable color space for image representation then selecting a satisfactory classifier. The HSV (hue-saturation-value) color space is used since it has several interesting properties [62]. The classifier is created by maximizing the agreement between the three channels of HSV for estimating the boundaries on each subspace. This method for creating the classifier removes the need for supervised annotation and allows rapid adaption of the classifier to different data which are the drawbacks of most of the existing classifiers [64]. Let $L = l_1, l_2, l_3, l_4, l_5, l_6$ be the boundaries of the subspace. A pixel (i) is classified as a skin if its color components in HSV space conform to : $\ell_1 < H_i < \ell_2$, $\ell_3 < S_i < \ell_4$, and $\ell_5 < V_i < \ell_6$

An estimate of the boundaries on each subspace given the channels HSV of the color image I is expressed as

$$\mathbf{L}^* = argmax\ (\tau) \tag{26}$$

where τ is the Kendall's agreement given by [64]

$$\tau = \frac{P(H \in I_1, S \in I_2, V \in I_3)\ P(H \in I_1^c, S \in I_2^c, V \in I_3^c)}{\sqrt{P(H \in I_1)P(H \in I_1^c)P(S \in I_2)P(S \in I_2^c)P(V \in I_3)P(V \in I_3^c)}} \tag{27}$$

where $I_1 = [\ell_1\ \ell_2]$, $I_2 = [\ell_3\ \ell_4]$, $I_3 = [\ell_5\ \ell_6]$.

A Dynamic programming-based solution is used to optimize two of the parameters at a time, iterating among channels until a solution is converged.

The skin class is modeled as multivariate gaussian distribution and the probability of the pixel C(i,j) being a skin is given by

$$P(C(i,j)) = \frac{1}{2\pi^{3/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(C-\mu)\Sigma^{-1}(C-\mu)^T) \tag{28}$$

where $C = [H\ S\ V]$, $\Sigma = diag(\sigma_H, \sigma_S, \sigma_V)$, $\sigma_H = \frac{l_2 - l_1}{6}$, $\sigma_S = \frac{l_4 - l_3}{6}$, , $\sigma_V = \frac{l_6 - l_5}{6}$, $\mu = [\frac{l_1+l_2}{2}\ \frac{l_3+l_4}{2}\ \frac{l_5+l_6}{2}]$.

The probability of the face candidate resulting from the face detector is being false alarm based on skin model is given by

$$P(f_s) = \sum_{i=1}^{i=N} \sum_{j=1}^{j=M} \frac{1 - P(C(i,j))}{N * M} \tag{29}$$

2..3 Combining probabilities for rejecting pseudo faces Based on the facial feature and skin model, each candidate face has a probability for being a false alarm from each model $P(f_f)$,and $P(f_s)$, respectively. The probability of false alarm based on combining the two models is given by

$$P(f_t) = P(f_s)\ w_s + P(f_f)\ w_f \tag{30}$$

where w_s, and w_f are the weight of skin model, and facial feature model. The optimal weights for the facial feature and skin models has been empirically determined to be 0.75 and 0.25, respectively in color images. In grayscale images, the weights are 1 and 0 respectively since there is no information about the skin in gray image.

3. Experimental Results

To evaluate the proposed system, the face Detection Database and Benchmark (FDDB) is used. This database was created in 2010 to act as a benchmark for face detection in unconstrained conditions. It consists of 2845 images with 5171 manually annotated faces. It contains photographs from several news sources under unconstrained environments with a wide range of challenges including partial occlusion, difficult poses, low resolution and out of focus faces. It contains mainly colored images with only 18 gray scale images. Some images, interestingly, are colored but contain both real colored faces and printed gray scale faces. The ground truth annotations of the faces are ellipses. To represent the degree of match between annotation and detection, the ratio of the overlapped area to the annotation area is calculated. If this ratio is greater than 0.5 then this detection is considered a true positive, otherwise it is considered a false positive.

The results of the proposed algorithm are compared with the following algorithms: The Open CV implementation of the Viola-Jones face detector which is used as the base face detection algorithm. Since it is used in the core of the proposed algorithm this makes it the natural baseline in the following comparisons. Mikolajczyk et al. [68] approach is also included which is considered one of the best performing public implementations of face detection algorithms [67]. The approach by Subburaman et al. [70] showed improvement in performance for a range of false positives while the performance of Jain et al [67] exceeded all the above algorithms. The performance of the proposed algorithm showed significant improvement over these algorithms. The performance curves for all of these approaches are shown in Figure 35.

Figure 36 shows a sample of the results acquired by the proposed face detector on the FDDB dataset. The originally annotated faces are displayed as red ellipses while the final results are displayed in green boxes. The blue dotted boxes show some of the false positives that were removed by proposed approach.

To show the effect of each block in the proposed block diagram, further experiments were conducted by removing one block at a time and measuring the resulting performance

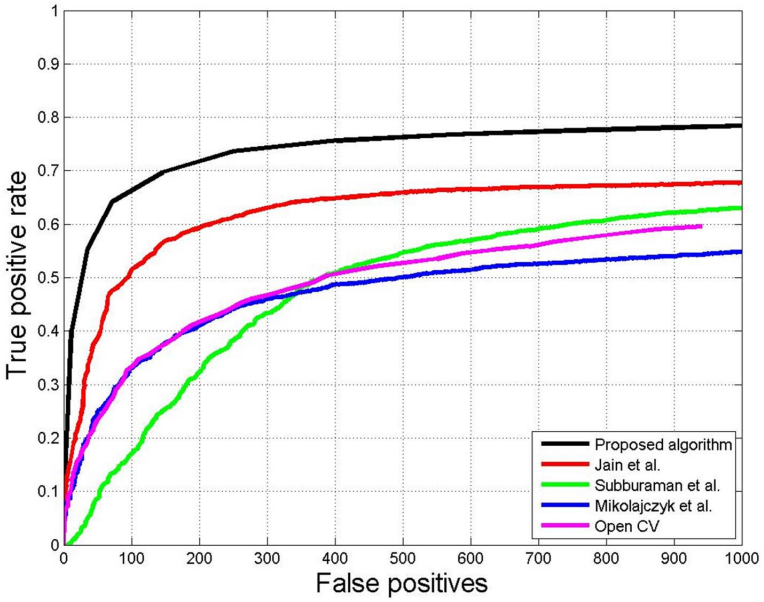


FIGURE 35: ROC curves for different approaches.

TABLE 3: Comparison of the effect of each component

Method	TP	FP	Time in sec
Proposed detector	0.77	403	0.82
Without facial Feature	0.67	479	0.37
Without skin	0.73	436	0.81

at specific operating point. Table 3 shows TP and FP rates with the associated execution time. Removing the facial feature step results in 76 false positive increase while removing the skin results in only 33 false positive increase. Also, the execution time shows that the added blocks did not add significant overhead time.

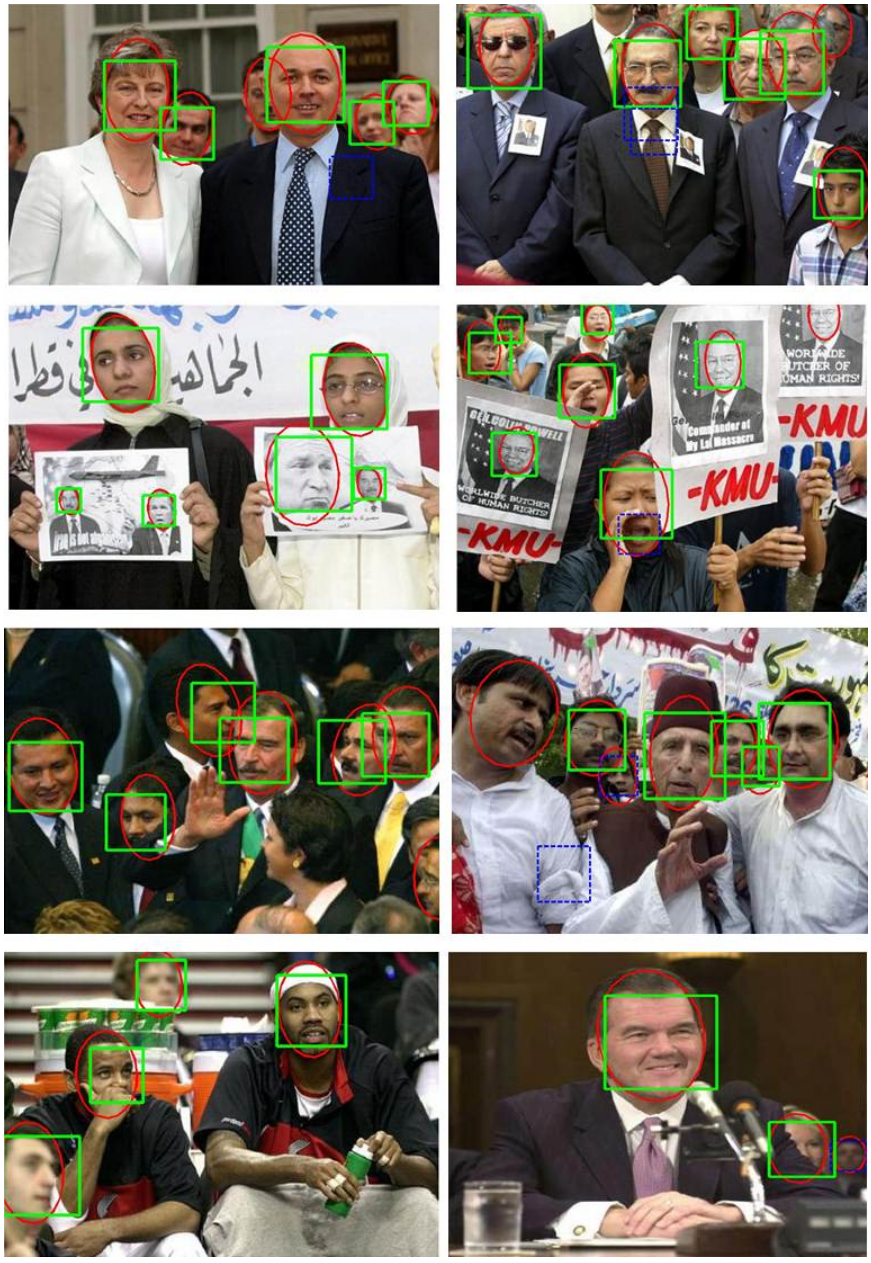


FIGURE 36: Sample results of the proposed face detector.

CHAPTER 4
POSE INVARIANT FACE REPRESENTATION FOR FACE RECOGNITION

The face recognition pipeline usually consists of three main modules: face detection, face representation, and face matching. Face detection is the first step in this process since it segments the facial region from the background before further processing is performed. Face representation provide useful low-level information from image. Face matching measures the similarity between two face representations to indicate whether the probe face belongs to a certain person based on his gallery face.

Face representation and face matching complete the framework to achieve same ultimate goal. Face representation aims to find a representation that is discriminative for inter-person difference and invariant to intra-person variations. The intra-person variation is a variation in capturing condition such as pose, lighting, and expression. While, face matching aims to make the distance between the two face representations, which belong to the same identity relatively smaller than the distance of the two face representations, which belong to different identity. In other words, face matching aims to have a similarity measure that is robust against intra-person variations and discriminative for inter-person difference.

This chapter focuses on extracting a face representation that is invariant to pose variation which is the most challenge in intra-person variations. The organization of the remaining of this chapter is as followed. First, a review on face representation is presented. Then, related work for pose invariant face representation is presented, followed by the two proposed approaches for pose invariant face representation. Finally, experimental results is discussed.

4.1 Face Representation

The face representation algorithms can be categorized into two broad categories.

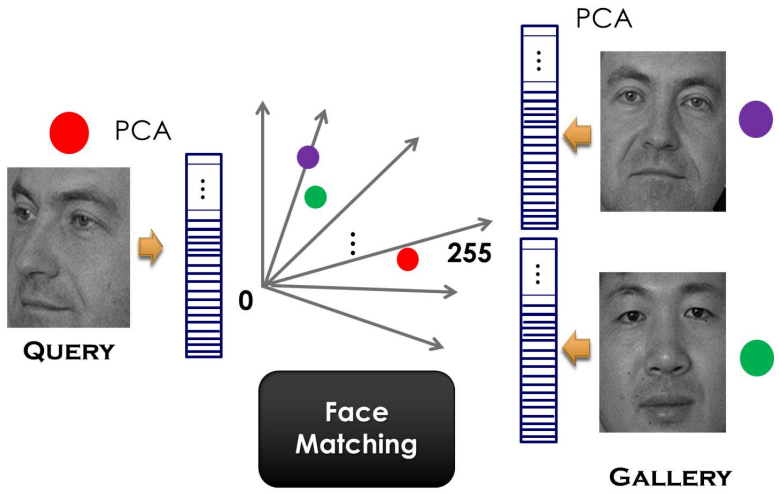


FIGURE 37: An example showing the distance between two frontal images of different persons is smaller than the distance between the same person under different view points using holistic approach

The first is the holistic category where the appearance of the whole face image is transformed into a vector in one step (e.g. Eigenfaces [2], and Fisherfaces [1]). These approaches usually have low recognition rates under pose changes as they do not take into account the 3D alignment issue when creating the feature vector. Figure 37 shows representation of the feature vector of two frontal images for two different persons and one pose image for one of this person. In particular, it shows the distance between two frontal images of different persons is smaller than the distance between the same person under different viewpoints.

The other category is local approaches where the face image is divided into blocks. The blocks are defined with grid over a face image. The appearance of each block is converted to a feature vector independently. The whole face representation, face signature, is a concatenation of the feature vectors of the different blocks. The local approaches are more robust for pose problem. However, they suffer from the problem of missing regions and region displacement as shown in Figure 38. To overcome the problem of missing regions and region displacement, the face image appearance is represented with the appearance of sparse patches instead of the patches on face grid. These patches are

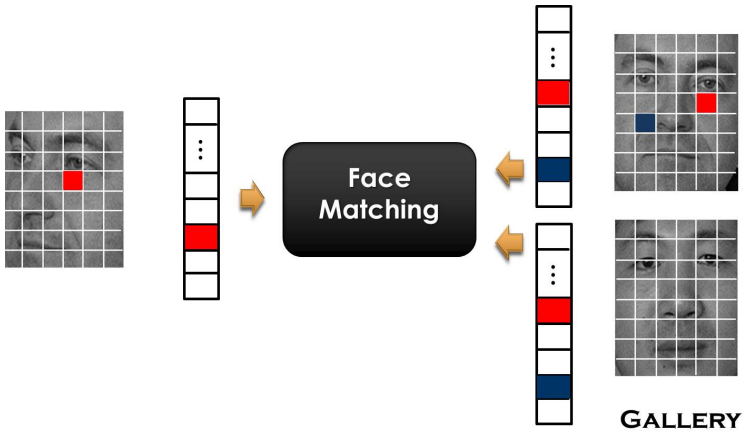


FIGURE 38: An example showing lack of correspondence due to missing regions and region displacement. Blue and red blocks indicate region displacement and missing region, respectively for traditional local approaches for face representation.

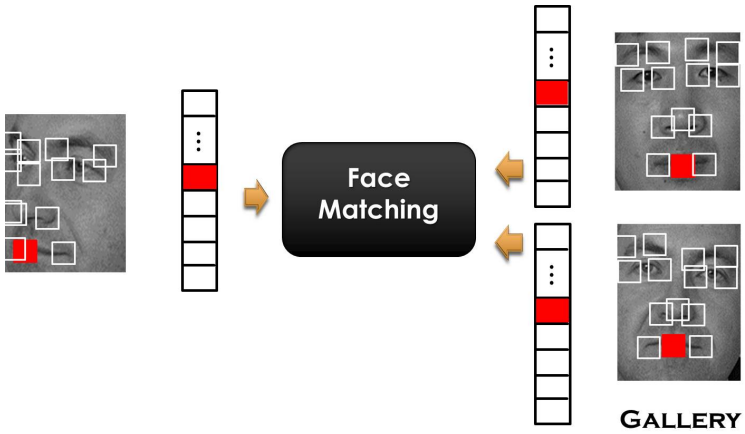


FIGURE 39: The feature based local approaches for face representation.

taken around facial feature points as shown in Figure 39. This representation is called the facial feature based face representation. This facial signature is state of the art, since it is more compact and efficient than traditional grid based face representation by preventing the region displacement. Moreover, it highlights the important area in the face since most of facial parts does not have distinguishing characteristics.

Gabor wavelets [9] and Local Binary Patterns (LBP) [5] are the most widely used algorithms for converting the appearance face into a feature vector. Due to its similarity to

FIGURE 40: Left, Traditional face signature using LBP [5]. Right, face signature using feature based LBP.

perception in human vision system, Gabor wavelets have been successfully applied to face recognition with many proposed variants in the literature [91]. The first and very notable method that used Gabor wavelets was elastic bunch graph matching (EBGM) [95]. The inherent disadvantages of Gabor based methods are heavy computation cost and very high dimension of feature vectors. In contrast, LBP based methods require lightweight computation and smaller feature vectors while providing very competitive recognition performance comparing with Gabor based ones. Many variations of LBP have been proposed, a related literature survey can be found in [6]. Many systems were formed by combining Gabor wavelets and LBP to capture the light computation from LBP and the high performance from gabor wavelets (e.g [94], [93], [92]).

In this work, the local binary pattern is adopted due to its efficacy. The used LBP signature is generated as follows. The LBP code of a given pixel (x_c, y_c) (a decimal value) is computed by comparing its intensity with the intensities of its surrounding pixels which are located on a circle, whose center is at the pixel itself. In details, with n neighboring pixels ($n = 8$), and radius $r = 3$, the formula for calculating LBP label of one pixel

$$LBP^{N,R}(x_c, y_c) = \sum_{i=1}^{N} s(g_i^{N,R} - g_c)2^{i-1} \tag{31}$$

where g_c is the gray level intensity of the given pixel and g_i is gray level intensity of i^th

neighbor. The $S(x)$is defined as

$$s(a) = \begin{cases} 1 & if a \geq 0 \\ \\ 0 & if a < 0 \end{cases} \tag{32}$$

LBP code is extracted from all pixels in the image patch 21×21 pixels around facial feature point (i), then LBP codes in the image patch are represented by the histogram (x_f). The histograms of image patches are concatenated into a single feature vector, which represents the face signature.

Following [5], the similarity between the signature of two face images $x, and\ \xi$ is measured by Chi square statistics (χ^2)

$$\chi^2(x, \xi) = \sum_{f=1}^{T} \sum_{j=1}^{M} \frac{x_{fj} - \xi_{fj}}{x_{fj} + \xi_{fj}}, \tag{33}$$

where T is the number of selected facial feature points (total number of image patches in the face image) and M is the number of bins in the histogram of LBP codes in each image patch.

4.2 Related Work for Pose Invariant Face Recognition

Pose invariant face representations can be categorized into multi-view based, poses normalization based, and invariant feature extraction based algorithms. In the first and second category, the challenge due to pose is solved in the image domain before transformation into feature vector. While in the last categories, the challenge is handled in the feature vector domain.

The simplest approach in multi-view based is recording each subject at each possible angle. A related approach is to take several images of the subject and use these to build a statistical model that can interpolate to unseen views. Other approaches of research interest under this category involve rendering 2D images for each subject under different poses from a 3D model of the head. The 3D model of the face can be directly measured or can be constructed from multiple images (e.g geometric stereo, photometric stereo) or video (structure from motion). These methods are valid, and some produce high-quality results.

However, they require either multiple images, or 3D sensors. Zhang et al. [110] introduced an automatic texture synthesis (ATS) approach to synthesis rotated virtual face views from a single frontal view for recognition using a generic face shape model. This face shape was generated by averaging 40 3D face shapes in range data format which were aligned using two eyes' locations. Liu and Chen [111] introduced a probabilistic geometry assisted (PGA) face recognition algorithm to handle pose variations. In their algorithm, human heads were approximated as an ellipsoid whose radiuses, locations, and orientations were estimated based on universal mosaic model. Their assumption that all faces have same 3D geometry is weak. The generic face model does not capture all 3D faces variances.

Many efforst have addressed the issue of 3D reconstruction of human face for recognition from single image. Atick et al. [76] introduced the first statistical SFS method by parameterizing the set of all possible facial surfaces using principal component analysis (PCA). Smith and Hancock [109] embedded a statistical model of surface normal within a shape from shading framework. Blanz and Vetter [77] introduced a face recognition system based on 3D morphable models that depend on image-based reconstruction and prior knowledge of human faces. The prior knowledge of face shapes and texture was learned from a set of 3D face scans. Then, shape and texture information in the forms of vertices and diffuse reflectance coefficients were spanned into different eigenspaces where principal component analysis was performed to construct a 3D morphable model. However, the identity-related shape and texture coefficients may be affected during cost function minimization [3]. Castelan et al. [78] developed a coupled statistical model, which is a variant of the combined AAM [27] that can recover 3D shape from intensity images with a frontal pose. The shape and intensity models in Castelan's work are similar to that of the AAM model. Note that in the shape recovery literature, albedo can be used, interchangeably, with the term intensity. The primary difference in Castelan's approach is that the 2D shape model in AAM is replaced with a 3D shape (height map) model.

The second category of the approaches that handle pose variation is the pose normalization-based approach. These aim to generate a virtual frontal image from a captured head pose image. One of these approaches is Active Appearance Model (AAM) [27], which was pro-

posed as a 2D model-based approach for face alignment. Once the model is fitted to an input image, the optimized shape model parameters are used to estimate the pose angle. Then, a frontal view of an input face image can be synthesized by configuring the shape parameters that control the pose using the optimized appearance model parameters. Instead of synthesizing a frontal face using the texture of the optimized model, Guillemaut et al. [82] warped the texture inside the shape of the fitted model to the pose corrected shape. The latter method is evaluated by Gao et al. [81]. In their work, they evaluated the effect of different texture wrapping techniques. They conclude that texture warping approaches have the advantage of preserving the textural information such as moles and freckles contained in the original image, which are lost in the synthesis-based approaches where the model parameters only represent the principal components of the appearance.

On the other hand, Chai et al. [80] introduced a local linear regression algorithm. In their algorithm, they divided the face image into patches. The appearance of each patch in the probe image was represented as a linear combination of the appearance of corresponding patches in the training images at pose angle of the probe image. The coefficients of the linear combination are used to combine the appearance of corresponding frontal patches in the training images to generate the virtual frontal view for the patch. However, this approach handles the pose problem in discrete domain (e.g., it solves pose 15, 45). Also, it requires manual detection of the center of two eyes and knowing the head pose angle of the probe image. Recently, Ho et al. [88] overcome the latter problem by using a Markov Random Field and an efficient variant of the Belief Propagation algorithm to estimate the head pose of the input image. However, this solution is computationally intensive i.e., it takes two minutes to estimate a pose.

Blanz and Vetter [77] recovered a 3D model from an input consisting of 2D face image using the idea that the variations in appearance caused by pose are closely related to the 3D face structure since 3D face information extracted as shape and texture features remain the same across all poses. This is the core of 3D Morphable Model (3DMM). Hence, in their work, given a 2D image they estimated the corresponding 3D model, i.e., the 3D shape and texture space, which is used in the matching process. Although this

method performs better than other algorithms for pose invariant face recognition, it heavily depends on the accurate extraction of 3D information from the 2D image. However, to learn the 3D shape and texture space, it requires 11 fiducial points and 3D face models during training. Thus estimation of 3D information is a difficult problem and computationally intensive, which makes this method is it too slow to be used in real-time application. To overcome these disadvantages, Prabhu et al. [112] showed that 3D depth information is not discriminative for pose invariant face recognition. Therefore, Asthana et al. [75] generated a virtual 2D image at frontal pose from a generic 3D face shape that is aligned with the input image using 2D detected facial feature points. Liu et al. [89] used a similar idea to the work in [75], but they improved the accuracy of recognition by filling occluded regions using facial symmetry property and they also enhanced the facial feature detection by using the multi-view random forest embedded active shape model.

The last category of pose invariant face representations are based on invariant feature extraction. Kanade et al. [83] learned a probabilistic model of the distance between two feature vectors of two face images. They modeled the distance between the two feature vectors in a similar (i.e., same identity but may differ in pose angle.) or a dissimilar (i.e., different identities) group as Gaussian distribution. Prince et al. [120], [113] used generative models to synthesize face images of a person across different poses from a common latent variable, which is called Latent Identity Variable (LIV). At the time of recognition, the images are transformed to the LIV space using a pose-specific linear transformation and recognition is carried out in that space. To learn the model parameters, they used the EM algorithm, which is prone to local minima and is computationally expensive. Moreover, the assumption that a single LIV can be used to faithfully generate all the different poses of a person seems to be over simplified. This is clear from poor performance even for small poses angles with simple intensity features. To improve the performance, they used 14 to 21 manually annotated points on face images to extract Gabor filter responses, which are more discriminative than raw pixels. However, this approach cannot be used in many applications since locating fiducial points automatically and accurately in non-frontal images is still an open problem.

Castillo and Jacobs [79] used the cost of stereo matching between a gallery face image and a probe face image to recognize faces. Since the approach is purely image based, it does not consider appearance change due to pose variation. Sarfraz et al. [84] assumed that there's a linear transformation between a face features representation in each pose and the face features representation in the frontal view. Then they represented any face using a gradient location-orientation histogram (GLOH) [56]. Although, there is no need for manual annotation in their algorithm, the solution is in the discrete domain of poses and information about the head pose of the probe is needed. Sharma et al. [86] used partial least square to learn a linear transformation from gallery and probe sets to correspondence latent space (CLS) where direct comparison can be applied. Their comparison in CLS is done using Linear Discriminant Analysis (LDA). The proposed approach solves the problem in discrete domain of poses and manual annotation of 4 points are needed for the solution while the head pose has to be known. Li et al. [87] represented the face image in each pose as a linear combination of face images in a training set. In each pose, they used a linear regression to formulate this representation. Also, two poses can be coupled by identity so joint face representation across two poses can be formulated as coupled regression problem. They used a similarity measure based on the correlation between the regressions parameters of a probe and a gallery images and a reconstruction error of the probe and the gallery images from the training set. Similar to other approaches, their approach had the same disadvantages that the head pose has to be known, it solves the problem in discrete domain of poses and manual annotation of five points are needed for the solution.

4.3 Rendering Posed face images for Pose Invariant Face Recognition

The first proposed approach for pose invariant face representation is based on rendering face images at different poses for each subject from the enroll image. The gallery in this approach consists of multiple images for the person at different poses that are generated from enroll image/images. Rendering face images requires information about 3D shape and texture for a subject face. The information of texture can be captured from the

FIGURE 41: Samples from the gallery. Columns from left to right are: the frontal captured image, synthesized images at poses 40°, 20°,−20°, and −40°.

galley face image, while the 3D facial shape can not be easily inferred from the image. This work focuses on two algorithm for 3D reconstruction. The first algorithm is statistical shape from shading where 3D shape is estimated using single image and prior model. The second algorithm is stereo reconstruction where the input is two face images from two different camera where geometric information about the relation between two cameras is known. Figure 41, and Figure 42 show the render images at different poses while the 3D is reconstructed using statistical shape from shading and stereo, respectively.

FIGURE 42: Samples from the gallery. Rows from upper to lower are: the left captured image, the right captured image, synthesized images at poses 40°, −40°, −20°, and 20°.

1. 3D Face Reconstruction from a Single Image

This subsection discusses the model-based approach for 3D facial shape recovery using a small set of feature points from an input image of unknown pose and illumination. The methods discussed here need only the 2D feature points from a single input image to reconstruct the 3D shape. Specifically, the input is a 2D image with detected feature points and the output is a 3D shape. This algorithm formulates the shape recovery problem into a regression framework, i.e., it uses Principal Component Regression (PCR) to reconstruct the 3D shape.

The USF database [104] used in this work contains both albedo and dense shape, where they are expressed as Monge patches, i.e., $(x, y, Z(x, y), T(x, y))$. The image data of the USF database samples are manually annotated with 68 points, which are the same facial points of the face alignment step. Since both image and dense shape data are in correspondence with each other, the annotation points can also be applied to the height maps, which results into 3D sparse shapes. Since the USF dataset has multiple subjects, a series of dense shapes together with corresponding sparse shapes exist. This series of dense and sparse shapes is integral to the proposed method in this work.

Suppose the input is a 2D sparse shape, which is the output of facial feature points detector, and the goal is to find the camera projection matrix $\mathbf{C}$ from its unknown (and yet to be solved) actual 3D sparse shape. A good substitute for this unknown 3D shape is the mean shape. A camera projection matrix can be computed between the mean 3D sparse shape and the input 2D sparse shape. Further, this projection matrix can be used to project a sample USF 3D sparse shape to the 2D space. The projection matrix $\mathbf{C}$ can be used to project all USF database samples to the 2D space.

The next step is to build two models related to the 3D USF dense shapes and the projected 2D shapes, projected facial features, of USF dataset , i.e., $\mathbf{s}_{3D} = \bar{\mathbf{s}}_{3D} + \mathbf{P}_{s_{3D}}\mathbf{b}_{s_{3D}}$ and $\mathbf{s}_{2D} = \bar{\mathbf{s}}_{2D} + \mathbf{P}_{s_{2D}}\mathbf{b}_{s_{2D}}$.

Principal Component Regression (PCR) is used to model the relationship between the dependent and independent data in the combined model. The basic idea is to decompose

both 2D and 3D dense shapes into a low-dimensional subspace, i.e., replace $\mathbf{x}_i$ and $\mathbf{X}_i$ by their respective PCA coefficients $b_{s_{2D},i}$ and $b_{s3D,i}$. Standard multivariate linear regression (MLR) is then performed between the low-dimensional representations of $\mathbf{x}_i$ and $\mathbf{X}_i$.

Let $\mathbf{T} = [b_{s_{2D},1}, \cdots, b_{s_{2D},m-1}]$ and $\mathbf{U} = [b_{s_{3D},1}, \cdots, b_{s_{3D},m-1}]$ be the low-dimensional representations of $\mathbf{x}_i$ and $\mathbf{X}_i$, respectively. Performing MLR yields

$$\mathbf{U} = \mathbf{T}\mathbf{C}_R + \mathbf{F} \tag{34}$$

where $\mathbf{C}_R$ is the matrix of regression coefficients and F is the matrix of random noise errors. The least squares method then provides

$$\tilde{\mathbf{C}}_R = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{U} \tag{35}$$

There are two remaining steps before the 3D dense shape can be recovered. The shape coefficient of the 2D input feature points need to be solved, i.e., $\mathbf{b}_{s_{2D},inp} = P_{s_{2D}}^T(\mathbf{x}_{inp} - \bar{\mathbf{s}}_{2D})$. Using the PCR model above, the 3D dense shape coefficient can be inferred with the following equation, $\tilde{\mathbf{b}}_{s_{3D}} = \mathbf{b}_{s_{2D},inp}\tilde{\mathbf{C}}_R$. The solved shape coefficient $\tilde{\mathbf{b}}_{s_{3D}}$ can be substituted to the 3D shape model, i.e., $\mathbf{x}_r = \bar{\mathbf{s}}_{3D} + \mathbf{P}_{s_{3D}}\tilde{\mathbf{b}}_{s_{3D}}$, to get the 3D dense shape. Algorithm **1** below summarizes these steps.

To quantify the reconstruction accuracy, the 3D shape for 80 out-of-training USF samples is recovered. The input images are generated with a random pan angle within the range of (-20 to 20), where the face moves left-to-right or right-to-left, sideways. For each reconstructed shape, the following measures are used: (a) Height Error - the recovered height map is compared with the ground truth height and the mean absolute error is computed as

$$S_{err} = \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{|s_i - s_{GT,i}|}{s_{GT,i}} \tag{36}$$

where N_p is the number of pixels, s_i and $s_{GT,i}$ are height values at the ith pixel position for the recovered shape and the ground−truth shape, respectively, and (b) Surface Orientation Error calculated by examining the directions of the recovered normals vectors

Algorithm 1 Principal Component Regression (PCR) Framework for 3D Dense Shape Recovery

INPUT: (a) Input image feature points, $\mathbf{x}_{inp}$ (b) USF dense $(\mathbf{X}_1^d, \cdots, \mathbf{X}_n^d)$ and sparse shape samples $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ (c) Sparse mean shape, $\mathbf{X}_m$

OUTPUT: (a) Recovered 3D dense shape, $\mathbf{x}_r^d$

1: **Solve for the camera projection matrix**: Determine $\mathbf{C}$ such that $\mathbf{x}_{inp} = \mathbf{C}\mathbf{X}_m$.

2: **Project all 3D sparse shapes to the 2D space using the computed projection matrix**: Solve for $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$, such that $\mathbf{x}_i = \mathbf{C}\mathbf{X}_i$

3: **Build the 3D dense shape model from the USF samples using PCA**: Construct $\mathbf{s}_{3D} = \bar{\mathbf{s}}_{3D} + \mathbf{P}_{s_{3D}}\mathbf{b}_{s_{3D}}$.

4: **Build the 2D sparse shape model from the projected 2D USF samples** $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$: Construct $\mathbf{s}_{2D} = \bar{\mathbf{s}}_{2D} + \mathbf{P}_{s_{32}}\mathbf{b}_{s_{2D}}$.

5: **Replace the 3D dense shape samples** $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ **with its coefficients**: Solve for $\mathbf{b}_{s_{3D},i} = P_{s_{3D}}^T(\mathbf{X}_i - \bar{\mathbf{s}}_{3D})$

6: **Replace the projected 2D shape samples** $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ **with its coefficients**: Solve for $\mathbf{b}_{s_{2D},i} = P_{s_{2D}}^T(\mathbf{x}_i - \bar{\mathbf{s}}_{2D})$

7: **Setup matrices for Principal Component Regression (PCR)**: Let $\mathbf{T} = [\mathbf{b}_{s_{2D},1}, \cdots, \mathbf{b}_{s_{2D},m-1}]$, and $\mathbf{U} = [\mathbf{b}_{s_{3D},1}, \cdots, \mathbf{b}_{s_{3D},m-1}]$

8: **Build the PCR model**: Construct $\tilde{\mathbf{C}}_R = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{U}$

9: **Solve for the shape coefficients of the 2D input feature points** $(\mathbf{x}_{inp})$: Solve for $\mathbf{b}_{s_{2D},inp} = P_{s_{2D}}^T(\mathbf{x}_{inp} - \bar{\mathbf{s}}_{2D})$

10: **Solve for the shape coefficients**: Get $\tilde{\mathbf{b}}_{s_{3D}} = \mathbf{b}_{s_{2D},inp}\tilde{\mathbf{C}}_R$

11: **Solve for the recovered 3D dense shape**: $\mathbf{x}_r^d = \bar{\mathbf{s}}_{3D} + \mathbf{P}_{s_{3D}}\tilde{\mathbf{b}}_{s_{3D}}$

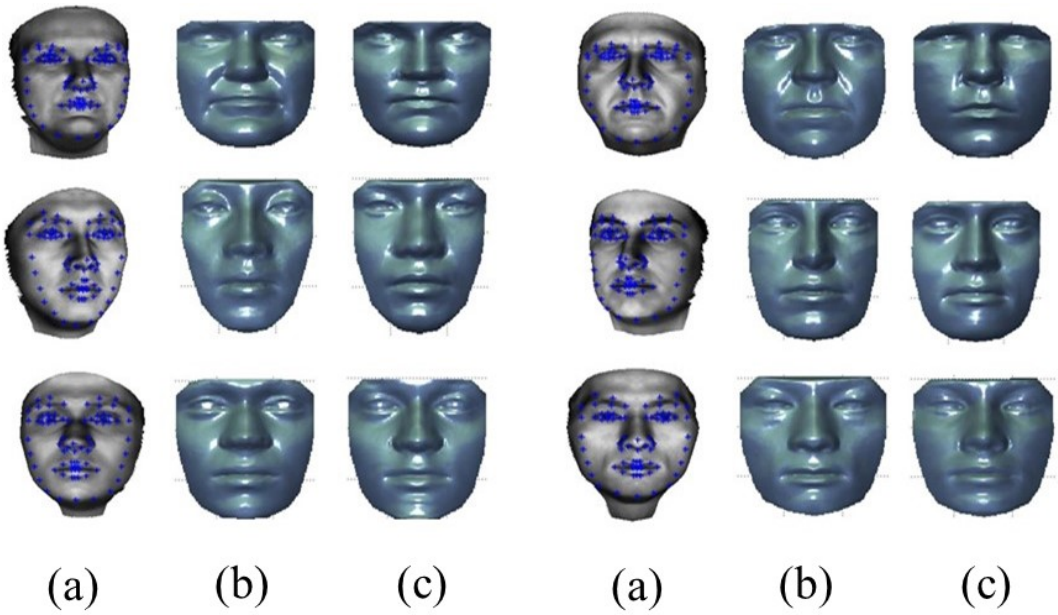


FIGURE 43: Recovered shapes, together with the input image and ground-truth(GT) shape, for the 3D shape recovery from 2D detected facial feature points

are compared with the ground truth data. The average of the difference angle is computed as

$$\theta_{err} = \frac{1}{N_p} \sum_{i=1}^{N_p} \cos^{-1}\left(\frac{n_i \cdot n_{GT,i}}{\|n_i\| \cdot \|n_{GT,i}\|}\right) \tag{37}$$

n_i and $n_{GT,i}$ are normal vectors at the ith pixel position for the recovered shape and the ground-truth shape, respectively.

Figure 44 is a side-by-side visualizations of the mean height and surface orientation to compare the $3D$ shape recovery accuracy using the manual annotation facial feature against the detected facial feature points using proposed detector. Average mean height error is 2.71% and 3.09% and average mean surface orientation error is 0.044 rad and 0.050 rad, across all 80 out-oftraining sample, for the case of manual annotated facial features and proposed facial feature points detector, respectively.

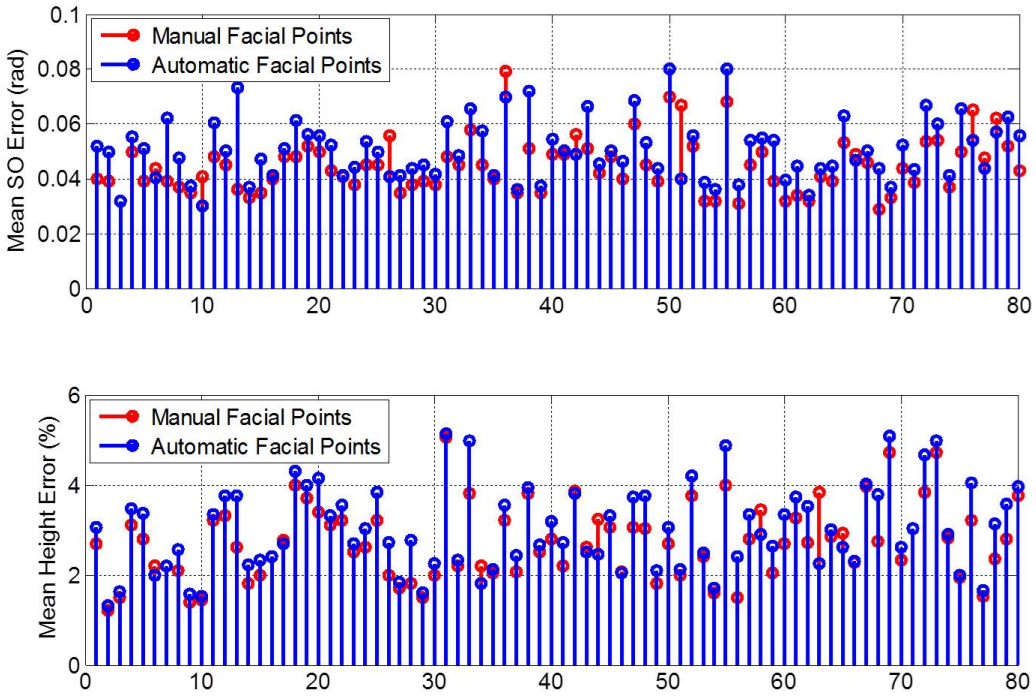
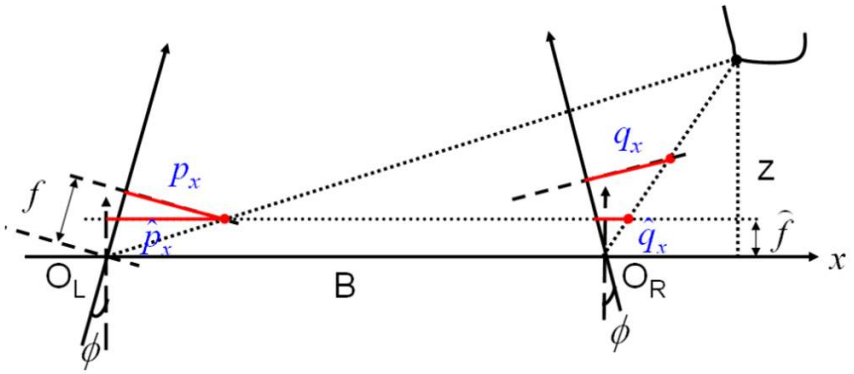


FIGURE 44: The 3D reconstruction accuracy, mean height error and mean surface orientation error, using manual annotated facial feature points verses using detected facial feature points from proposed algorithm.

2. 3D Reconstruction from stereo Imaging

This subsection discusses stereo reconstruction for 3D facial shape recovery from two face images captured from two different cameras where geometric information about the relation between two cameras is known. In the classical stereo matching problem, the objective is to find the pairs of corresponding points p and q that result from the projection of the same scene point into the two images. As shown in Figure 45, the distance from the scene point to the cameras is determined by difference in image locations of points p and q. This difference is called the disparity. To reconstruct the 3D shape of an object, one needs to determine the disparities of the correspondences between pixels of the images.

Finding the disparity map $\mathbf{f}$ for a stereo pair is an image labeling problem. Where, $\mathbf{I}$ and $\tilde{\mathbf{I}}$ represent the left and right images, respectively. The set of label $\mathcal{L}$ is the disparity range $\{\partial_x^1, \cdots, \partial_x^K\}$. To correctly solve this problem, the constraints of the visual



$$\hat{p}_x = \sqrt{p_x^2 + f^2}\,\sin(\phi + \tan^{-1}\frac{p_x}{f})$$

$$\hat{f} = \sqrt{p_x^2 + f^2}\,\cos(\phi + \tan^{-1}\frac{p_x}{f})$$

$$\hat{q}_x = \hat{f}\,\tan(\phi + \tan^{-1}\frac{-q_x}{f})$$

$$Z = \frac{\hat{f}\,B}{\hat{p}_x + \hat{q}_x}$$

$$X = Z/\tan(\frac{\pi}{2} - \varphi - \tan^{-1}(\frac{p_x}{f}))$$

$$\hat{Z} = \sqrt{X^2 + Z^2}\,\cos(\frac{\pi}{2} - \varphi - \tan^{-1}(\frac{Z}{X}))$$

$$Y = \hat{Z}p_y / f$$

FIGURE 45: General stereo pair setup. The relation between the depth and the disparity.

correspondence should be satisfied. The uniqueness is one of these constraints, where each pixel in $\mathbf{I}$ corresponds to at most one pixel in $\tilde{\mathbf{I}}$. The occlusion is another constraint, where some pixels do not have correspondences. To overcome these problems, Kolmogorov's approach [96] is used in this work, which treated the two images symmetrically by computing the disparities for both images in the same time. In this case, $\mathcal{P}$ represents the set of pixels of both images and $\mathbf{f}$ is the labeling of both images.

The framework for this problem can be the search for MAP configurations in a MRF model. The MAP problem is formulated as minimizing an interaction energy for the model. Therefore, the desired disparity map $\mathbf{f}$ is equivalent to minimizing the following energy function:

$$E(\mathbf{f}) = \sum_{\{p,q\}\in\mathcal{N}} V(f_p, f_q) + \sum_{p\in\mathcal{P}} D(f_p). \tag{38}$$

Most of stereo matching approaches are based on an assumption that corresponding pixels have a similar color value. Thus the data term $D(f_p)$ is usually chosen as a similarity measure (e.g., [97]) that handles slight variation between corresponding pixels' colors.

However, colors of the real scene are transformed nonlinearly to another colors in the stereo pair images, which violates color consistency assumption. In this work, the color normalization approach described in [98] is used to convert the transformation between the pixels' colors in two images from non-linear to a liner transformation. Then, the Normalized Cross Correlation (NCC) [99] is used in data term since (NCC) is invariant to linear transformation. The execution time of the NCC is reduced by reducing the calculation of the means of the pixels in the windows using an integral image. To enforce the visibility constraints, the approach [96] compares only pixels that have the same disparity in both images. The smoothness term is chosen to be piecewise smooth prior to allow smooth variations in the disparity map.

$$V(f_p, f_q) = min(|f_p - f_q|, \mathtt{M}), \tag{39}$$

where $\mathtt{M}$ is a constant. Note, $\mathtt{M} > 1$ leads to piecewise smooth prior.

After finding the disparity map, the occluded regions is filled by interpolating between the correctly reconstructed pixels of each scan line using a cubic Splines interpolation model. The cloud of 3D points, which are estimated using the disparity map and system geometry, is denser than required for reproducing the amount of actual detail present in the face. So first, these points are downsampled. Then to remove some artifacts of the reconstruction, an additional surface fitting step is done. The reconstructed scattered data is approximated in a least squares sense to generate a smoothed surface. Finally, triangular mesh from the smoothed and downsampled points is generated.

The stereo matching approach is used to reconstruct human faces in a 3D face recognition framework. Figure 46 illustrates the setup that is used to capture images. The setup parameters are shown in Table (4). Figure 47 shows the complete flow chart of the face reconstruction process.

To evaluate 3D reconstructions, the reconstructed 3D from stereo is compared to a laser scanner. In this comparison, the distance, which are shown in Figure 4.3.2(a), is measured between specific points in the 3D faces. These specific points and distances are selected depending on the most discriminatory anthropometric facial proportions. Since

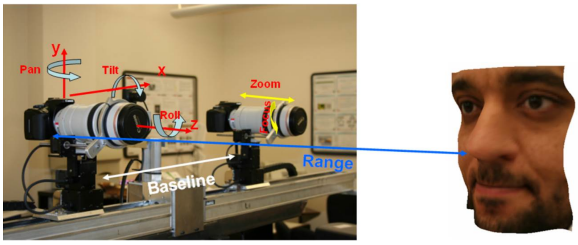


FIGURE 46: The system setup.

TABLE 4: Stereo setup parameters

Range (m)	Baseline B (m)	Zoom f(mm)	Focus	Pan ϕ (degree)	Tilt (degree)	Roll (degree)
3	0.6	150	Range	5^o	0	0

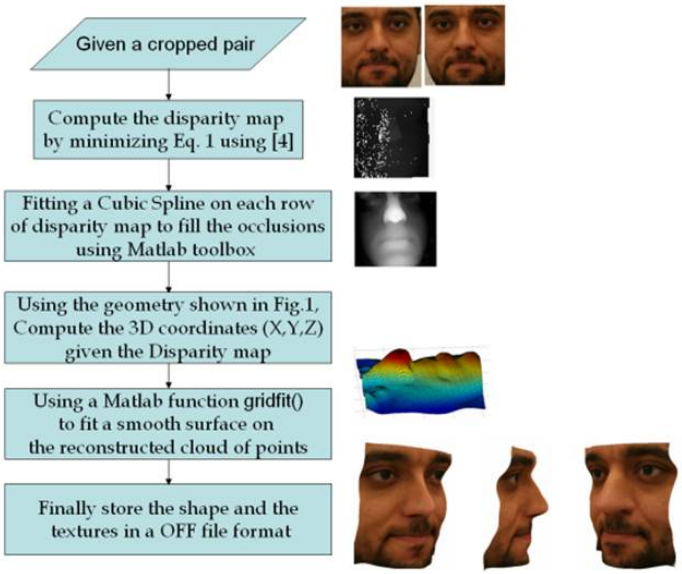
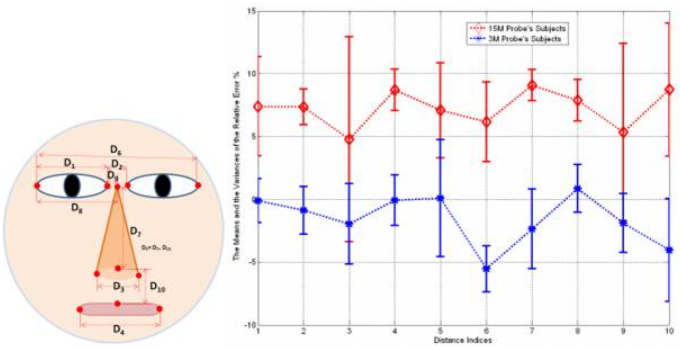


FIGURE 47: The stereo matching-based human faces reconstruction flowchart.



(a) (b)

FIGURE 48: (a)Distances that are used in comparison. (b) The means and variances of the relative error between the distances from the proposed results and the laser scanner's outputs.

the final goal is to generate a 3D face that will be used in recognition, measuring the errors in these distances is a good evaluation to the reconstruction. To do the comparison, the differences between these distances are computed that are measured from the reconstructed 3D faces in this work and the correspondence distances that are measured from laser scanner's 3D faces. For the most 10 discriminant distances, shown in Figure 4.3.2(a), Figure 4.3.2(b) shows the means and variances of relative errors (for 12 shapes for which the scanner's 3D faces is existed). The results in Figure (b) illustrate that, for 3M probe's reconstructions, the variances is less than $\pm\%10$, which means less than a ± 1 millimeter error in the centimeter. For the 15M probe's results, the variances is less than $\pm\%15$, which means less than a ± 1.5 millimeter error in the centimeter.

4.4 Weighting of Facial Features for Pose Invariant Face Recognition

The main drawback of the rendering approach either based on statistical shape from shading or geometric stereo is increasing the gallery size since each enroll image/stereo not represented by a single face signature, however, it is represented by number signatures corresponding to different pose angles that corresponding to render images. Facial feature based face representation, as shown in Figure 39, seems to be a good solution for pose

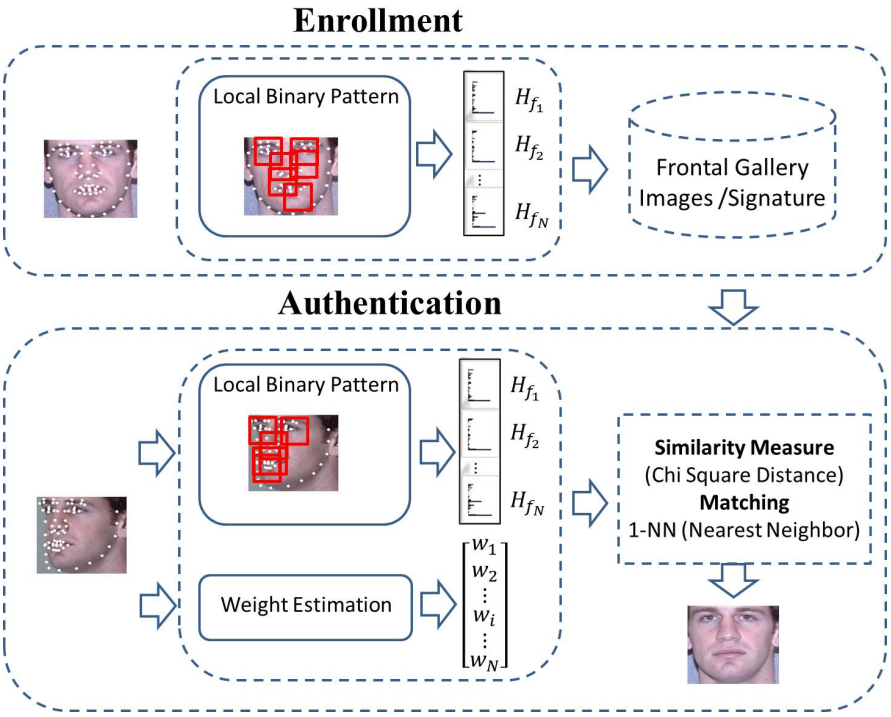


FIGURE 49: A schematic diagram for dynamic weighting of facial features approach.

invariant face representation since it generated feature vectors in patches around facial feature points in the captured pose image and the frontal gallery image. The question arise is that pixels in the patch in the frontal gallery image and captured pose image correspond to the same vertices in the 3D of the subject.

Since the vertices that correspond to pixels in a patch around a certain facial feature in the frontal gallery image may be visible, partially occluded, or completely occluded in the posed probe image, the pixels in a patch around a certain facial feature in a captured probe image may not correspond to the same pixels in the frontal gallery image.

To illustrate that pixels in the patch in the frontal gallery image and captured pose image are not correspond to the same vertices in the 3D of the subject, Figure 50 shows 3D dense face shapes with colored vertices at different poses angles. The red vertices correspond to pixels in a patch around a certain facial feature in a frontal galley image. The yellow vertices correspond to pixels in a patch around the same facial feature in a posed probe image. Overlapped vertices have green color. As shown in the Figure 50, the green vertices around facial features, which are located in the right half of the face, decrease

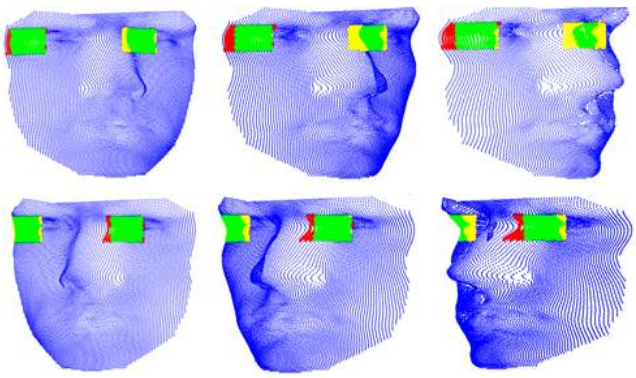


FIGURE 50: The effect of pose on the corresponding patches overlapping. Green vertices increases and decrease as the head moves left and right.

dramatically as the face rotates to the right and vice versa. This decrease in the overlapped vertices means that the distance between signature vectors, which are extracted from the corresponding patches in the frontal gallery image and posed probe image, increases. This is clearly because each signature represents a different pixels patch in the corresponding images.

To solve this problem, the first scenario is that the signature should be extracted around each facial feature, in the enroll and probe image, at the common pixels only, i.e., pixels correspond to green vertices. The main disadvantage of this solution is that in the offline stage for each enrolled image, instead of extracting a single signature from a fixed patch around a certain facial features, for each facial feature many signatures should be extracted from different variable patches corresponding to common areas at different head pose angles. This increases the gallery size i.e., each gallery image can not be represented by a single signature, however, it is represented by number signatures corresponding to pose angles you deal with. This method will have the same problem of rendering images at different pose angles. Otherwise, the image should be saved in the gallery instead of the extracted signatures. In the latter case, signature extraction is done online. This means that signatures will be extracted for all gallery images and a probe image at the time of recognition. This solution is time consuming.

The second scenario, which is proposed in this work, is assigning weights for each

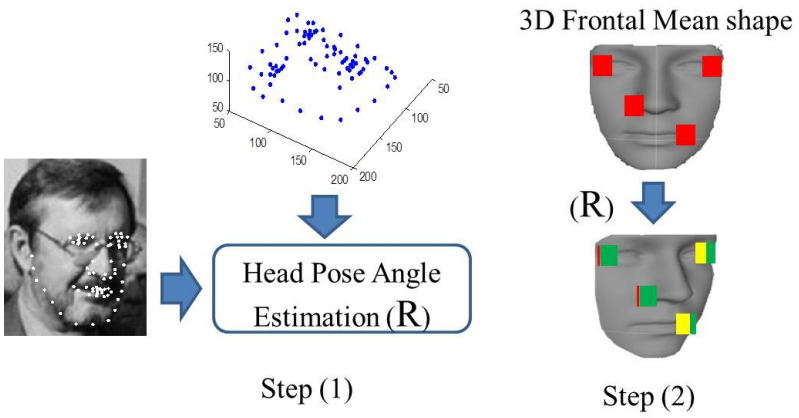


FIGURE 51: The two steps for estimation the weight of each facial feature. Step 1 (left) is estimation the pose. Step 2 (right) rotate 3D dense and find overlap between the patch around facial feature point at the frontal and captured pose angle.

facial feature point at each pose based on the overlapping score.

These weights are estimated from the captured head pose image and a 3D dense mean shape. This is done in two steps as shown in Figure 51. The first step is estimating head pose angle from posed image. The second one is rotating the 3D dense mean shape and find the overlap score in the window around each facial feature point.

A 3D point in world coordinate is related to its corresponding 2D point in the image plane by a projection matrix. Based on the perspective camera model assumption, the projection matrix has 11 degrees of freedom; five intrinsic camera parameters, three rotation parameters, and three parameters representing the translation of the camera center with respect to the world coordinate system.

Estimation of the projection matrix needs at least 6 (3D-2D) correspondences. In this work, many 2D points are known, these are the output of the face alignment step, however, their corresponding 3D points are unknown. A good substitute for these unknown 3D points is the 3D mean shape. The difference between the actual value and the mean can be neglected in this application. Using these data, the rotation parameters (the unknown pose) can be estimated.

In the second step, the weight for each facial feature is estimated. First, each vertex

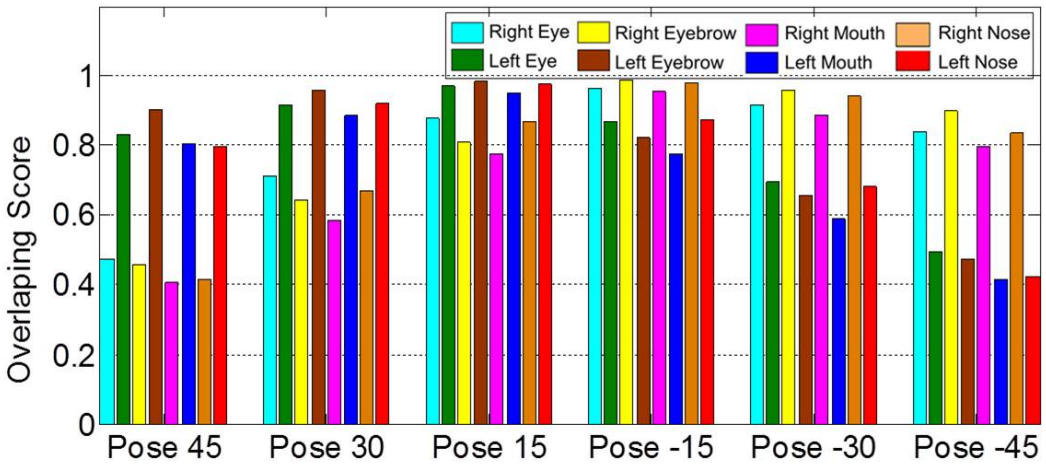


FIGURE 52: Proposed scheme for generating virtual frontal image (pose normalized image) from captured pose image

in the frontal 3D dense mean shape is rotated to the same pose of captured image using the estimated rotation parameters. Then the z-buffer test is carried out to determine the set of visible and occluded vertices in the rotated 3D dense mean shape. According to the z-buffer test, a vertex belongs to the occluded set, if there exists one or more vertex possessing the same x and y components and has z component less than the z component of this vertex. A subset of vertices, which correspond to a subset of pixels in a patch around a certain facial feature, is visible, if it is visible in both the frontal and the rotated 3D dense mean shapes. The overlapping score at a feature i is calculated using these visible vertices as follows:

$$O_i(V_{f_i}, V_{r_i}) = \frac{V_{f_i} \bigcap V_{r_i}}{V_{f_i} \bigcup V_{r_i}} \tag{40}$$

where V_{f_i}, and V_{r_i} are the subsets of visible vertices around the facial feature i in the frontal and rotated 3D dense mean shapes, respectively.

In this work, the facial feature points are divided into eight groups: parts of right eye, left eye, right eyebrow, left eyebrow, right part of mouth, left part of mouth, right part of nose, and left part of nose. Figure 52 shows the overlapping percentages of different facial parts at different head pose angles. At each pose, the scores are normalized to give the dynamic weights at that pose.

4.5 Experimental Results

1. Databases for pose invariant face Recognition

Experiments are performed on three indoor data sets where each subject is captured using one camera at different angles, i.e., the CMU-PIE [103], FERET [102] and the Multi-PIE [105], and one outdoor in-house dataset, UoFL-EWA, where the subject is simultaneously captured by two camera, stereo imaging at different angles.

CVIP-EWA dataset is collected in outdoor with different illumination conditions. It is collected within one year with laps three months. It consists of 773 sessions taken at distance ranges of 30, 50, 80, 100 and 150 meters and at different head pose angles from -40° to 40°. Each session consists of a pair of images from two cameras where the face is centered in both images with additional ground truth information related to cameras, pan and tilt angles and the baseline distance (the distance between two cameras). The image pairs are captured at different ranges using Canon 7D cameras, with 800mm telephoto lenses, FOV(2.5^{o}). Figure 56 shows samples images from UoFL-EWA database. The frontal neutral expression session at distance 50 meter for each subject is used in the enrollment stage, while the other pose images used as query.

In the CMU-PIE database, all 68 persons with one frontal image for enrollment and 6 non-frontal poses images with yaw angles from -45° to $+45^{\circ}$ with pose difference $+22.5^{\circ}$(*Pose ID c11, c29, c05, c37*) and pitch from -22.5° and 22.5° (*Pose ID c07, c09*) for each subject as query images (testing images) are used in these experiments. For the FERET dataset, all 200 subjects at 6 different non frontal poses with yaw angles $-40^{\circ}, -25^{\circ}, -15^{\circ}, +15^{\circ}, +25^{\circ}$, and $+40^{\circ}$ (*Pose ID bh, bg, bf, be, bd, bc*) are used. For Multi-PIE, one hundred and thirty seven subjects (Subject ID 201 to 346) captured in four sessions with neutral expressions and frontal illumination are used at 6 different non-frontal poses with yaw angle from -45° to $+45^{\circ}$ with pose difference $+15^{\circ}$ as query images against neutral expressions, frontal illumination, and frontal pose as enrolled image.

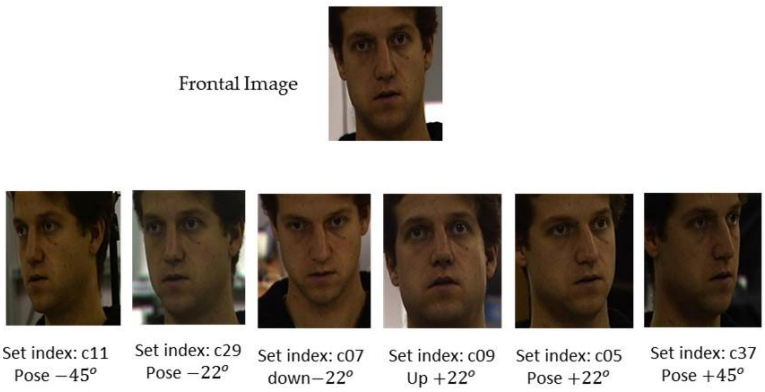


FIGURE 53: Example subject from CMU-PIE database. The top row is gallery image at frontal pose. The bottom row is probe (query) images at different pose angle.

FIGURE 54: Example subject from FERET database. The top row is gallery image at frontal pose. The bottom row is probe (query) images at different pose angle.

FIGURE 55: Example subject from Multi-PIE database. The top row is gallery image at frontal pose. The bottom row is probe (query) images at different pose angle.

FIGURE 56: Example subject from CVIP-EWA database. The middle column is gallery image at frontal pose. The other columns are probe (query) images at different pose angle. Thus data base is stereo, therefore there is left and right image for each subject in each session.

2. Weighting facial features for pose invariant face representation

In the "Enrollment" stage, for each subject, the near frontal neutral expression session is used. The face represented by local binary pattern signature [5] extracted around detected facial feature points.

In the online stage, the input is a probe session, which is a 2D image that is captured under unknown pose. The face is detected [51], then the system automatically detects the facial features. Afterward, the LBP signature is extracted from patches around these facial features. Then, the face representation is compared with the signatures of the gallery while the weight for each facial points is different based on pose.

First, the effect of facial feature weighting using manual annotated facial feature points is presented. Fifteen facial feature points are manually annotated. These 15 facial features are used to align mean $2D$ shape of 51 facial feature points to the face image. The recognition rate using each group from facial features at frontal and each pose angle from $-45°$ to $45°$ are shown in Figure 58. Figure 58 also shows that recognition rate of the right group of features (right eye, right eyebrow, and right part of mouth and nose) is higher than left group of features at poses $45°$, $30°$, and $15°$. Moreover, the difference between the right and the left group of the features in the recognition rate increases with the increasing head pose angle since the right group of features get more occluded and vice versa at poses

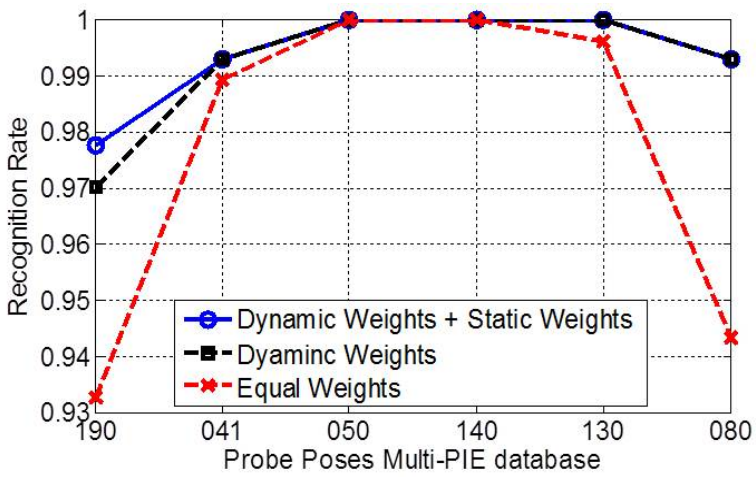


FIGURE 57: First one recognition rate for studying the effect of dynamic and static weights using manual annotated facial feature points on Multi-PIE

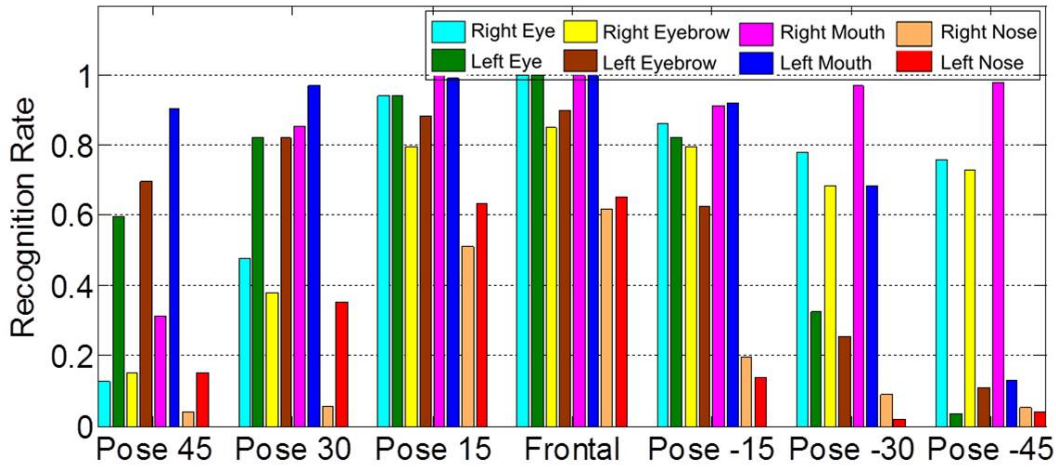


FIGURE 58: First rank recognition rate using different face parts at different head pose angle using manual annotated facial feature points on Multi-PIE database.

$-45°$, $-30°$, and $-15°$. The results emphasis the importance of assigning a dynamic weight for each facial feature based on the pose.

Since each group of features have different recognition rate as compared with the other facial features. A weight for each facial feature is assigned based on discriminatively of the group feature as compared with other groups. This weight is static since it does not change with pose angle. This weight is calculated in the frontal pose. The static weight for the feature is the recognition rate using this feature only in the recognition divided by the

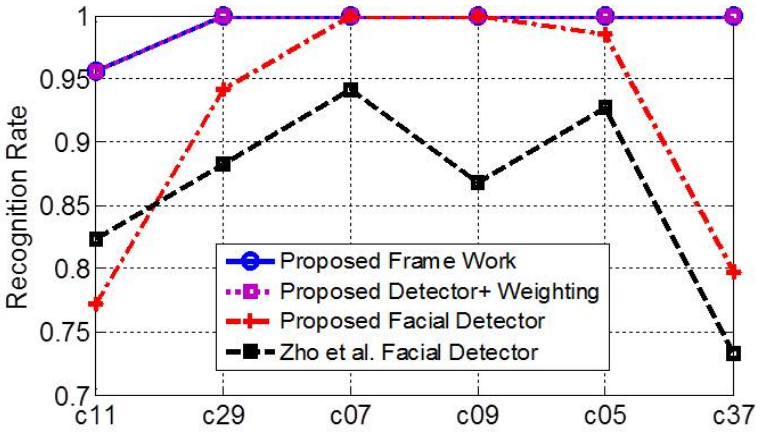


FIGURE 59: Performance evaluation for studying the effect of facial feature detector and proposed weights on CMU-PIE database.

sum of the recognition rate of all other features.

Secondly, an evaluation of the effect of different facial feature point detectors and proposed weighting scheme with automatic facial points detection is presented. Four experiments are done. First, the facial feature detector in [46] is used and without using the proposed weights is evaluated (baseline performance). Second, the proposed detector is used and also without using the proposed weights. These two experiments show the importance of proposed detector. Third, the proposed detector is used with using dynamic weights (DW) for facial features. The last one, the proposed detector is used with using both dynamic weights (DW) and static weights (SW) for facial features. Figures 59, 60, and 61 show the performance of these four experiments on CMU-PIE, FERET, and Multi-PIE database respectively. Moreover, the effect of the weighting on the recognition rate, regardless the effect of facial feature detector, is studied by using manual annotated facial feature points as shown in Figure 57.

3. Rendering Pose face images for Pose Invariant Face Representation

In the ”Enrollment” stage, for each subject, the near frontal neutral expression session is used to reconstruct a 3D face either using stereo based or statistical shape from shading. Ray-tracing techniques are used to render synthetic images under different poses

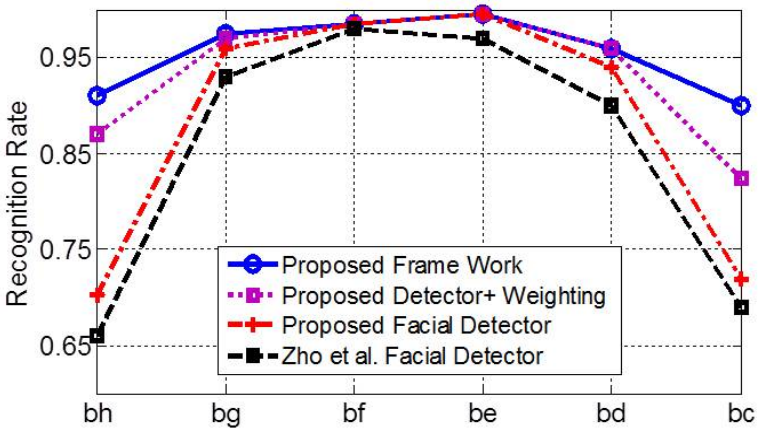


FIGURE 60: Performance evaluation for studying the effect of facial feature detector and proposed weights on FERET database.

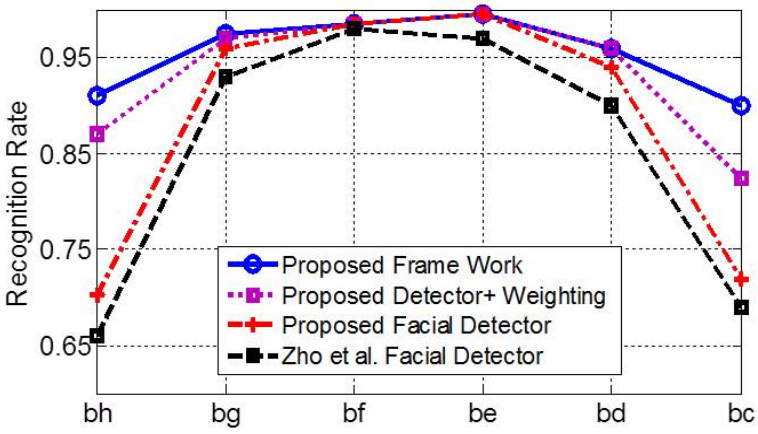


FIGURE 61: Performance evaluation for studying the effect of facial feature detector and proposed weights on Multi-PIE database.

using the reconstructed 3D shape for subject's face. A gallery entry of each subject consists of five images: the captured image plus four synthesized images at poses of yaw angles ±20° and ±40°. Finally, LBP technique [5] is used to generate five signatures from the five images around facial feature points that was part in 3D reconstruction.

In the online stage, the input is a probe session, which is a 2D image that is captured under unknown pose. The face is detected [51], then the system automatically detects the facial features. Afterward, the LBP signature is extracted from patches around these facial features. Finally, the probe pose is estimated from the projection matrix [75], which is determined by using the mean shape, the first step in 3D reconstruction. Then, its signature is compared with the signatures of the gallery subset, which has the closest pose to probe pose.

The proposed framework is evaluated using four experiments. First, the gallery consists of the captured images and no synthesis images are enrolled. The signature is LBP around facial feature points that are detected by Zhu and Ramanan [46]. This experiment is denoted as "Facial Detector [46]+ No synthesis" as shown in Figures 62, 63, and 64. (Experiment II) Then, the previous experiment is repeated by using the proposed facial points detector [22]. This experiment is denoted as "Proposed detector + No Synthesis". (Experiment III), the last experiment deals with replacing the 3D reconstructed shape in the proposed framework with a generic 3D shape to show the effect of reconstruction accuracy on recognition. This experiment is denoted as "Proposed detector + Mean Shape Synthesis". The generic shape is constructed similar to that in [110]. Notice the incremental improvement of the recognition results as more components are added towards the proposed system framework.

Figures 62, 63, and 64 show the results of these four experiments on CMU-PIE, FERET, and Multi-PIE respectively while the 3D facial shape in these results is constructed from single image using statistical shape from shading. Moreover, Table 5 shows the results of these experiments on CVIP-EWA dataset but the 3D facial shape in these results is constructed using geometric stereo instead of statistical shape from shading.

Table 6 shows a comparison among proposed approaches which are rendering from

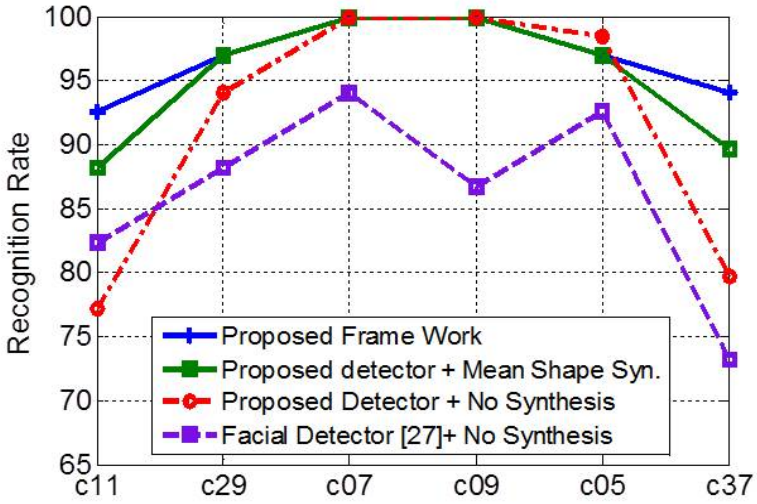


FIGURE 62: Comparison among the proposed framework and its variations to highlight the effect of each component on CMU-PIE database.

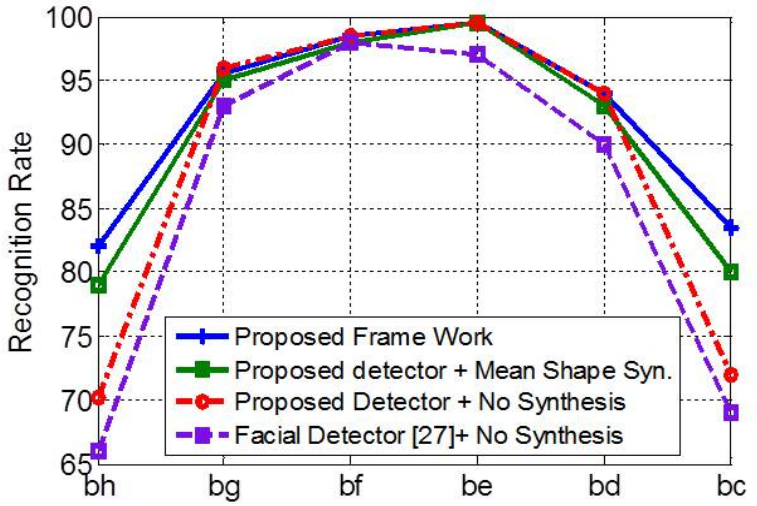


FIGURE 63: Comparison among the proposed framework and its variations to highlight the effect of each component on FERET database.

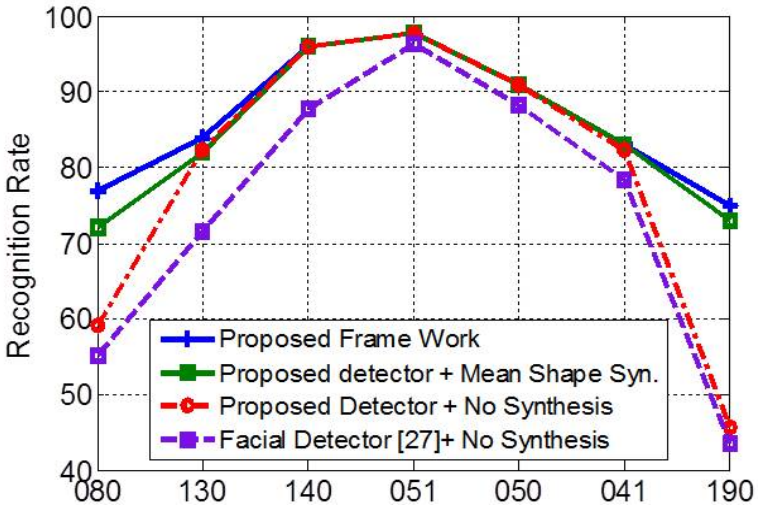


FIGURE 64: Comparison among the proposed framework and its variations to highlight the effect of each component on Multi-PIE database.

TABLE 5: Rank-1 recognition rates (number are percentage) on the UoFL-EWA dataset in three experiments: without including the synthesis images in the gallery (left column in each pose), ”generic+synthesized” approach (middle column in each pose), and ”stereo+synthesized” approach (left column in each pose).

Distance	−45°			−25°			+25°			±15			+45°			**Avg**		
50*m*	63	74	77	71	80	81	—	—	—	86	88	89	74	76	82	74	79	82
80*m*	52	68	75	60	72	73	83	85	85	76	81	83	63	64	68	67	74	77
100*m*	46	53	59	56	64	69	78	78	78	76	79	81	59	62	62	63	67	70
150*m*	31	33	44	41	43	48	49	49	49	33	49	55	26	36	45	36	42	49
Avg	48	57	63	57	65	68	70	71	71	68	74	77	56	60	64			

statistical shape from shading and geometric stereo approach and weighting of facial features approach. The comparison is based on recognition rate, number of images required, computational time, and memory usage on UoFL-EWA dataset since there is no publicly available stereo database.

TABLE 6: Comparison among proposed approaches for pose invariant face recognition [105].

Method	No. Images	Memory usage	offline time	−45°	−25°	+25°	+45°	**Avg**
Synthesis from Mean shape	1	×5	2.5 sec	57	65	74	60	64.0
Hybrid 2D-3D approach	1	×5	2.8 sec	58	64	74	62	64.5
Weighting of facial features	1	×1	0.8 sec	61	66	74	63	66.0
Stereo based from synthesis 2D images	2	×5	80 sec	63	68	77	64	68.0

4.6 Comparison with State-of-Art Methods

Dynamic weighting of facial features shows a better recognition rate than render face pose images. Moreover, the rendered pose face images approach needs five times memory as compared to the memory need by weighting facial feature approach to store the gallery signature. The gallery signature is frontal pose image plus four synthesis images in rendering approach but it is the signature of frontal pose image only in weighting facial feature approach. It also solves the problem in discrete domain that means that there is a sweet spot at pose angle of synthesis images. Therefore, weighting of facial features is compared with the state of art methods.

Tables 7, 8, and 9 show a comparison the weighting of facial features approach with state-of-art approaches on CMU-PIE, FERET, and Multi-PIE database respectively. The comparison is based on recognition rate, face alignment (either automatic or manual), the algorithm is trained on images from the same dataset of probe images, and the algorithm need information about probe image head pose angle.

Many methods (e.g. [88], [87], [86], [83], [80], and [84]) are based on building a model that related images at frontal with images at certain pose angle. Therefore, the information about the pose angle should be known prior to determine which model will be used and these algorithms handle a discrete set of head pose angles. The work in [83] and [84] solved this issue by using statistical model but the recognition drops significantly.

Ho et al. [88] proposed a method based Markov Random Fields and an efficient variant of the Belief propagation algorithm but it has a high complexity. Therefore, using prior information is an important drawback since it hinders to use the method in full automatic system and the performance drops significantly to solve this issue. Manual intervention is another important drawback since it affects the performance dramatically. The proposed method has a recognition rate 88% and 99.5% using manual and automatic detected facial feature points respectively.

It can be seen that the proposed approach in general outperformed the methods proposed in [85], [80], [81], [79], [84], [90] and [75]. The performance of the proposed method is close to the [88], [87], [86]. However, the advantage of proposed approach over [87], and [86] that there is no need for prior information about probe head pose angle. The proposed method solves the pose problem in continues domain (from -45° to $+45^{\circ}$), it is no discrete set of poses. Their results are based on at least four manual detected facial features points. The reported results are reported on whole dataset (68 subjects CMU-PIE, 200 FERET, and 137 Multi-PIE) but they reported in half number of subjects since they use the other half in training. The advantage of the proposed approach over [88] that it takes 2 minutes and the entire proposed approach process in the 2.8 seconds. measured on Intel(R) Xeon CPU 3.2 GHZ (excluding the timing for comparison that varies based on the size of the gallery in both methods).

TABLE 7: Recognition rates of different approaches on the CMU-PIE database [103].

Method	Alignment	Trained on CMU-PIE	NEED POSE	c11 −45°	c29 −22°	c07 ↑ 22°	c09 ↓ 22°	c05 +22°	c37 +45°	**Avg**
Kanade et al. [83]	Manual-3pts	Yes	No	96.8	100	100	100	100	100	99.5
Zhang et al. [85]	Automatic	No	No	71.6	87.9	78.8	93.9	86.4	74.6	82.2
Chai et al. [80]	Manual-3pts	Yes	Yes	89.8	100	98.7	98.7	98.5	82.6	94.7
Castillo et al. [79]	Manual	No	Yes	100	100	90	100	100	99.0	98.2
Sarfraz et al. [84]	Automatic	Yes	Yes	84.0	87.0	–	–	94.0	90.0	88.8
Asthana et al. [75]	Automatic	No	NO	98.5	100	98.5	100	100	97.0	99.0
Li et al. [87]	Manual-5pts	Yes	Yes	100	100	100	100	100	100	100
Ho et al. [88]	Automatic	No	No	97.0	100	100	97.0	98.5	100	98.8
Proposed	**Automatic**	**No**	No	**95.5**	**100**	**100**	**100**	**100**	**100**	**99.3**

TABLE 8: Recognition rates of different approaches on the FERET database [102].

Method	Alignment	Trained on FERET	Need POSE	bh −40°	bg −25°	bf −15°	be +15°	bd +25°	bc +40°	**Avg**
Blanz et al. [77]	Manual-8pts	No	No	95.4	96.4	97.4	99.5	96.9	95.4	96.8
Zhang et al. [85]	Automatic	No	No	62.0	91.0	98.0	96.0	84.0	51.0	80.5
Chai et al. [80]	Manual	No	Yes	55.0	89.5	93.0	89.0	77.0	53.0	76.1
Gao et al. [81]	**Manual**	Yes	No	78.5	91.5	98.	97.0	93.0	81.5	90.0
Sarfraz et al. [84]	Automatic	Yes	Yes	92.4	89.7	100	98.6	97.0	89.0	94.5
Asthana et al. [75]	Automatic	No	No	90.5	98.0	98.5	97.5	97.0	91.9	95.6
Li et al. [87]	Manual-5pts	Yes	Yes	96.0	99.0	98.0	96.0	96.0	91.0	96.0
Sharma et al. [86]	Manual-4pts	Yes	Yes	100	100	100	97.0	100	94.0	98.5
Ho et al. [88]	Automatic	Yes	No	91.0	97.3	98.0	98.5	96.5	91.5	95.5
Proposed	**Automatic**	**No**	No	**91.3**	**98.0**	**100**	**99.5**	**96.0**	**90.7**	**96.0**

TABLE 9: Recognition rates of different approaches on the Multi-PIE database [105].

Method	Alignment	Trained on Multi-PIE	Need POSE	080 −45°	130 −30°	140 −15°	051 0°	050 +15°	041 +30°	190 +45°	**Avg**
Zhang et al. [85]	Automatic	No	No	37.7	62.5	77.0	92.6	83.0	59.2	36.1	64.0
Schwratz et al. [90]	Automatic	No	Yes	65.0	87.0	99.0	–	94.0	84.0	65.0	83.0
Asthana et al. [75]	Automatic	Yes	No	74.1	91.0	95.7	96.9	95.7	89.5	74.8	87.7
Li et al. [87]	Manual-5pts	Yes	Yes	91.0	96.0	99.0	–	100	96.0	85.0	94.5
Sharma et al. [86]	Manual-4pts	Yes	Yes	85.7	93.7	98.7	–	98.7	94.9	87.8	93.3
Ho et al. [88]	Automatic	No	No	86.3	89.7	91.7	92.5	91.0	89.0	85.7	89.4
Proposed	**Automatic**	**No**	No	**81.0**	**88.0**	**96**	**97.7**	**91.0**	**85.0**	**77.0**	**88.0**
Proposed	**Manual-15pts**	**No**	No	**97.0**	**100**	**100**	**100**	**100**	**100.0**	**99.3**	**99.5**

CHAPTER 5
SIMILARITY MEASURE IN FACE RECOGNITION

Computing a similarity measure between a face representation and other representation plays an important role in the success of face recognition. The standard distance measure i.e., Euclidean distance, treats all face representations equally. However, certain image features could be more reliable than others. To overcome this drawback and to enhance the measure performance, prior information to discard bad features selectively in each individual matching circumstance should be used in computing the measure. The similarity measure should satisfy that, in the features space, the distance between same subjects is smaller than the one between different subjects.

Studies on the distance measure use supervised or unsupervised learning techniques to learn a similarity measure. The unsupervised learning is easier than its competitors since there is no need for a labeled training set. However, it is less accurate as compared with supervised learning techniques. Principle Component Analysis (PCA), Multidimensional Scaling (MDS), and Neighborhood Preserving Embedding (NPE) [131] are examples of unsupervised metric learning algorithms. On the other hand, for supervised learning, the need of labeled data is a challenging task, especially, in the case of learning a large scale dataset with a huge amount of data. There are two main settings for supervised labelling: unrestricted and restricted settings. The unrestricted setting where all data points have fully supervised labels is infeasible. The restricted setting, which is the easier one, specifies labels in the form of equivalent constraints. Where, each pair is labeled as similar or dissimilar without any information about the object class.

This chapter focuses on proposing a similarity measure between two pose invariant face representations. This similarity measure is learned using equivalent constraints labeled data (the restricted setting). The proposed similarity measure maps data from its

original features space to a target space such that a simple distance can be adequate for the verification task. The original feature space is invariant to pose but it may be affected by many uncontrolled sources of variations e.g., changes in illumination, expression and camera properties. On the other hand, the target space should be invariant to pose, illumination, and expression.

The organization of the remaining of the chapter is as followed. First related work about supervised distance learning in restricted setting is reviewed. Then, the proposed similarity measure is presented. Finally, experimental results and summary are explained.

5.1 Related Work

Supervised learning approaches that learn distance metrics can be categorized into linear and non-linear techniques. In the linear techniques, a linear mapping is performed to map feature vectors into another space. Then their pairwise Euclidean distances, in the projected space, are computed. Information-Theoretic Metric Learning (ITML) [138], and Distance Metric Learning with Eigenvalue Optimization (DML-eig) [141] are examples of this family. However, the original image space, which is used in many computer vision applications, is highly nonlinear. This non linearity is due to high variability of the image content and style. Thus; nonlinear supervised distance metric learning approaches show a better performance in this case. The non-linear supervised distance metric learning methods are also known as kernelization methods. These methods typically comprise two parts: the first part maps (usually nonlinearly) the input points to a features space often of much higher or even infinite dimensionality. Then the second part applies a relatively simple (usually linear) classifier in the projected features space [133]. An example of these kernelization methods is the localized multikernel metric learning (LMKML) algorithm which was recently presented by Lu et al. [153]. In order to reliably represent a set of images of the same class, authors extracted multiple order statistics as features of this set. Then, these features are mapped to another dimensional features space where a distance between samples is calculated as a dot product.

These supervised approaches can also be classified into global and local methods. In the local methods, the distance metric satisfies some local properties of the dataset. Learning the distance locally make the distance performance better in the retrieval and k-nearest neighbours applications. However, rather than requiring the equivalence constraints, information about the class labels are needed. Thus it is impossible to be used in a restricted setting i.e., training examples are labeled similar or dissimilar. One of these approaches is Locally Linear Embedding (LLE) [132]. In this approach, the authors tried to find the low dimensional embedding such that the local neighborhood structure is preserved. Locally Smooth Manifold Learning (LSML) is another local method, which finds a projection such that the local neighbors of different classes are separated. Also, Large Margin Nearest Neighbor (LMNN) [139] is a local method. This approach encourages target neighbors to be at least one distance unit closer than any imposter using two terms: One term strengthens the correlation to target neighbors while the other weakens it to impostors. Recently, to perform a kinship verification, Lu et al. [148] proposed a local iterative method, which is neighborhood repulsed metric learning (NRML). Their aim is to learn a distance such that samples with a kinship relation are pulled as close as possible and neighbored interclass samples are repulsed as far as possible. Given a set of labeled training images, first, they used Euclidean distance to find k-nearest neighbors for each sample. Then they performed a local optimization using these k-nearest neighbors to learn a metric. After that k-nearest neighbors are updated using the new metric. This is iteratively done until a convergence error is achieved.

On the other hand, the global methods learn the distance such that it satisfies some global properties of the data set. Examples of global methods are Relevant Component Analysis (RCA), Discriminative Component Analysis (DCA), Information-Theoretic Metric Learning (ITML) [138], Logistic Discriminant Metric Learning (LDML) [140], and Distance Metric Learning with eigenvalue optimization (DML-eig) [141]. Davis et al. [138] formulated the problem of learning a linear distance as a minimization of the differential relative entropy between two multivariate Gaussians constraining the distance function. Authors expressed this problem as a particular Bregman optimization problem of minimiz-

ing the LogDet divergence subject to linear constraints. Guillaumin et al. [140] introduced a linear logistic discriminant model to learn a metric. This model estimates the probability of whether the two samples belong to the same class i.e., $p(x_i = x_j)$. The a posteriori probability is modeled by a sigmoid function and model's parameters are estimated by iteratively adapting the Mahalanobis metric to maximize the log-likelihood. Ying et al. [141] introduced a metric learning approach, which minimizes the maximal eigenvalue of the symmetric matrix of Mahalanobis metric. Köstinger et al. [142] introduced a non-iterative algorithm to learn Mahalanobis metric based on statistical inference where the difference between two objects in similar and dissimilar groups are represented as a Gaussian distribution. Cao et al. [154] tried to exploit the good performance of cosine similarity function and the Mahalanobis distance in face verification. They proposed a generalized similarity metric learning approach that combines both distances in a unified formula. However, this method does not have any physical meaning.

5.2 Proposed Approach

A supervised method to learn a nonlinear similarity measure, based on a nonlinear combination of Mahalanobis distances, is proposed. The proposed approach is built on Kostinger et al. [142], where the measure is derived from a log-likelihood ratio of a difference feature vector of two intra-class samples to a difference feature vector of two inter-class samples. Since the training dataset is labeled in form of equivalent constraints (the restricted setting), the proposed method belongs to global approaches.

The identity of a subject and variations in the capturing process are the main components, which influence the appearance of the face image. The identity component of an face is constant regardless the variations in its capturing process. Therefore for the same identity, any feature vector $\mathbf{x}_i$, which represents the face appearance, is drawn from a k mixture of distributions. These distributions reflect the randomness in the capturing process variables e.g., expressions, lightings, backgrounds, hairstyles, etc.

Let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ denotes a sample from the pairwise differences space. If the

two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different identities, a new component, which represents variation in the identities, will appear in $\mathbf{x}_{ij}$. This component that represents the difference in the identities can be assumed as a random variable. Therefore, $\mathbf{x}_{ij}$ will be a sample drawn from a mixture of $k + 1$ distributions: variation in the identities distribution (or identities distribution for simplicity) and k distributions represent the capturing process variables. On the other hand, if the two face images $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same identity, there is no difference in the identities. Thus the identity component will not contribute in the feature vector in the differences space. Then $\mathbf{x}_{ij}$ will be a sample drawn from a mixture of k distributions, which represent the variations in the capturing process variables.

Using the Maximum a Posterior (MAP) rule, the decision about the two face images $\mathbf{x}_i$ and $\mathbf{x}_j$ are belonging to same identity is made by testing the following likelihood ratio:

$$l(\mathbf{x}_i, \mathbf{x}_j) = \frac{p(\mathbf{x}_i, \mathbf{x}_j|H_E)}{p(\mathbf{x}_i, \mathbf{x}_j|H_I)}. \tag{41}$$

Where H_I represents the intra-class variation hypothesis that the two face images $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same identity, and H_E is the inter-class variation hypothesis that the two samples belong to different identities, subjects.

Note that if the hypothesis H_E is valid, the pair $\mathbf{x}_i$ and $\mathbf{x}_j$ will lead to a high likelihood ratio. In the space of pair wise differences $\mathbf{x}_{ij}$, Equation 41 can be rewritten as follows.

$$l(\mathbf{x}_i, \mathbf{x}_j) = \frac{p(\mathbf{x}_{ij}|H_E)}{p(\mathbf{x}_{ij}|H_I)}. \tag{42}$$

According to the assumption, $p(\mathbf{x}_{ij}|H_E)$ is represented by a mixture of $k + 1$ distributions, and $p(\mathbf{x}_{ij}|H_I)$ is represented by a mixture of k distributions. Assuming these distributions are Gaussian, the likelihood ratio can be rewritten as follows.

$$l(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{m=1}^{k+1} \overline{w}_m \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \overline{\Sigma}_m^{-1} \mathbf{x}_{ij})}{\sum_{n=1}^{k} w_n \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \Sigma_n^{-1} \mathbf{x}_{ij})}. \tag{43}$$

The mixture parameters i.e., the covariance matrices $\overline{\Sigma}_m$ and Σ_n and the weights $\overline{w}_m$ and w_n are simultaneously estimated using the Expectation Maximization (EM) algorithm.

1. The relation between proposed approach and Mahalanobis distances

What is the relation between the proposed learned similarity measure and Mahalanobis distances? This is an important question to be answered. Equation 43 can be reformulated as follows.

$$\begin{aligned} l(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{m=1}^{k+1} \overline{w}_m \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \overline{\Sigma}_m^{-1} \mathbf{x}_{ij})(\sum_{n=1}^{k} w_n \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \Sigma_n^{-1} \mathbf{x}_{ij}))^{-1}, \\ &= \sum_{m=1}^{k+1} (\sum_{n=1}^{k} \overline{w}_m^{-1} w_n \exp(\frac{1}{2}\mathbf{x}_{ij}^T \overline{\Sigma}_m^{-1} \mathbf{x}_{ij}) \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \Sigma_n^{-1} \mathbf{x}_{ij}))^{-1}, \\ &= \sum_{m=1}^{k+1} (\sum_{n=1}^{k} \overline{w}_m^{-1} w_n \exp(\frac{-1}{2}\mathbf{x}_{ij}^T (\Sigma_n^{-1} - \overline{\Sigma}_m^{-1}) \mathbf{x}_{ij}))^{-1}, \end{aligned}$$

$$l(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{k+1} (\sum_{n=1}^{k} w_{nm} \exp(\frac{-1}{2}\mathbf{x}_{ij}^T \Sigma_{nm}^{-1} \mathbf{x}_{ij}))^{-1}, \tag{44}$$

where $w_{nm} = \frac{w_n}{\overline{w}_m}$ and $\Sigma_{nm}^{-1} = \Sigma_n^{-1} - \overline{\Sigma}_m^{-1}$.

Let $d_{\Sigma_{nm}}$ denotes the squared Mahalanobis distance $\mathbf{x}_{ij}^T \Sigma_{nm}^{-1} \mathbf{x}_{ij}$, then Eq.44 can be rewritten as follows.

$$l(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{k+1} (\sum_{n=1}^{k} w_{nm} \exp(\frac{-1}{2} d_{\Sigma_{nm}}))^{-1}. \tag{45}$$

As seen in Equation 45, the proposed similarity measure i.e., the likelihood ratio, is expressed as a non-linear combination of $(k(k+1))$ Mahalanobis distances. Where the learning matrix of each Mahalanobis distance is Σ_{nm}. To avoid over fitting, choosing k is critical.

Here, the relation between the proposed similarity measure and the one introduced by Kostinger et al. [142] can be discussed. Actually, Kostinger et al. [142] approach is a special case from the proposed one. Their big assumption is that $\mathbf{x}_{ij}$ is a sample drawn from a single Gaussian distribution weather the two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same identity or different identities. Köstinger et al. [142] did not give any interpretation or evidence of why the distance between two samples can be represented by a single Gaussian distribution. This can be interpreted as follows: Back to Equation 45, let m change from 1 to k instead $k+1$, n changing from 1 to k and $k = 1$, then the likelihood ratio will be a function of a single Mahalanobis distance. This means that $\mathbf{x}_{ij}$ is a sample drawn from a single distribution. This distribution will carry information about the variations in

capturing process, if the two samples belong to a same identity. Where, if the two samples have different identities, $\mathbf{x}_{ij}$ is a sample drawn from a single distribution that captures the identities variation. The limitation of the latter is that it ignores any information about the variations in capturing process.

2. Classification Enhancement by Proposed Approach

To highlight the enhancement of the proposed nonlinear combination of Mahalanobis distances on the classification of the pairs ($\mathbf{x}_i$,$\mathbf{x}_j$) to similar or dissimilar pairs, a simple example is discussed here. It is well known that the Mahalanobis distance $\sqrt{d_{\Sigma_{nm}}(\mathbf{x}_{ij})}$ is the L2-norm of a linearly transformed difference vector $\mathbf{x}_{ij}$. So $d_{\Sigma_{nm}}(\mathbf{x}_{ij})$ can be written as follows.

$$d_{\Sigma_{nm}}(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \Sigma_{nm}^{-1} \mathbf{x}_{ij} = \|\mathbf{L}\mathbf{x}_{ij}\|^2, \tag{46}$$

where $\Sigma_{nm}^{-1} = \mathbf{L}^\top \mathbf{L}$. However, Euclidean distance (L2-norm) does not achieve the classification task in many cases as shown in figure 65. Where, each point in this figure either a triangle or a circle represents a difference feature vector x_{ij}, which is a 2-dimension vector for illustration purpose. It is obvious that using L2-norm directly, i.e., the distance to the origin in X-Y coordinates, will misclassify all the similar samples i.e., triangles. Note that the discrimination between the distances to the origin from the triangles and from the circles is achieved. A better classification can be achieved using one Mahalanobis distance, which is equivalent to compute the L2-norm of the samples in V-U coordinates. However, the circle labeled (a) has a bigger distance than the three encircled triangles. This means that the three encircled triangles will be misclassified as dissimilar samples. Finally, by using the proposed nonlinear combination of just two Mahalanobis distances (i.e., the simplest case $k = 1$), the best classification can be achieved. Assuming the circle labeled (a) has L2-norm 20 in V_1-U_1 coordinates and L2-norm 2 in V_2-U_2 coordinates. The proposed similarity measure is $e^{-1} + e^{-10} = 0.3679$. Note that this is the farthest dissimilar sample, which means that all other circles have bigger similarity measures. On the other hand, assuming the triangle labeled (b) has L2-norm 4 in V_1-U_1 coordinates and L2-norm 4 in

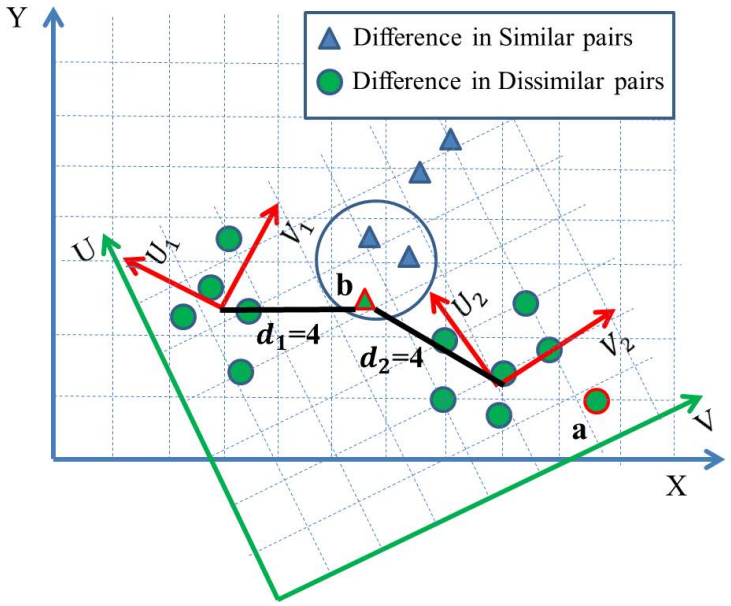


FIGURE 65: A classification example of 2D data illustrates the improvement achieved by the proposed similarity measure. X-Y coordinates is the pairwise difference space, which are linearly transformed to V-U coordinates using a one Mahalanobis distance and are linearly transformed to V_1-U_1 coordinates and V_2-U_2 coordinates using two Mahalanobis distances.

V_2-U_2 coordinates. The proposed similarity measure is $e^{-2} + e^{-2} = 0.27$. Note that this is the closest similar sample, which means that all other triangles have smaller similarity measures. Since the smallest similarity measure of dissimilar samples is greater than the bigger similarity measure of similar samples, all samples are correctly classified using the proposed non-linear combination of two Mahalanobis distances.

5.3 Experimental Results

The experiments are conducted on the two recent face recognition benchmarks: Labeled Faces in the Wild (LFW) [150] and Public Figures Face Database (PubFig) [149]. Since choosing the number of Mahalanobis distances to be learned in the proposed similarity measure is critical for avoiding overfitting, this issue is investigated in the following experiments by using different values of k.

FIGURE 66: Several examples face pairs of the same person from the LFW data set. Left: similar pairs and right: dissimilar pairs.

In the first set of experiments, a very challenging database, the Labeled Faces in the Wild (LFW) dataset [150] is used. LFW's faces are captured in uncontrolled settings i.e., general imaging and environmental conditions. These conditions include different expressions, poses, lightings, backgrounds, hairstyles, etc. This dataset can be considered as the de facto standard dataset for face identification. LFW dataset contains $13,233$ unconstrained face images of $5,749$ individuals. Some illustrative examples are given in figure 66. To be used for cross validation experiments, images of this dataset are divided into ten fully independent folds. Each fold contains 600 pairs of images. For 300 pairs out of 600, each pair belongs to an individual. In the remaining 300 pairs, the images of each pair belong to different individuals. Using this paradigm, the face verification task can be tested by individuals that have not been used in the training stage.

In the first experiment, the SIFT-based descriptor is used, which is proposed by Guillaumin et al. [140], to generate a feature vector. At three different scales, Guillaumin et al. [140] extracted 128 dimensional SIFT descriptors [156] from patches centered on 9 facial features (corners of the mouth, eyes and nose). The descriptor of Guillaumin et al. [140] is a 3456 dimensional feature vector. this descriptor is used, because it provides a fair comparison to the-state-of-the-art in distance learning [138, 140–142] where those authors reported their best results using that descriptor as compared with other descriptors e.g., Local Binary Pattern (LBP) [5], Three-Patch LBP (TPLBP) and Four-Patch LBP (FPLBP) [155].

Face verification results on the LFW dataset are shown in figure 67. Where, Receiver Operator Characteristic (ROC) curves for KISSME [142], DML-eig [141], LDML [140], ITML [138], SVM [55], and the proposed measure as a function of two Mahalanobis distances (i.e., $k = 1$) and as a function of six Mahalanobis distances (i.e., $k = 2$) are generated. The performance of the proposed similarity measure is computed by averaging verification results over the 10 folds. Notice that the proposed method outperforms others and reaches an Equal Error Rate (EER) of 83% at ($k = 1$). While this is a competitive improvement, the proposed method is also computationally efficient compared to ITML and LDML.

It is worth mentioning that other state-of-the-art e.g. [145–147, 151, 152], which focus purely on faces on LFW, provide better results, however, they combine many features and require considerably more domain knowledge. Also, for a fair comparison, the comparison is done only with the global methods e.g., LDML and ITML, which are similar to the proposed method that requires restricted setting to be learned. However, local methods, e.g., LMNN, require unrestricted setting to be learned (i.e., information about the class label is needed).

The influence of using Principle Component Analysis (PCA) for dimension reduction of the used feature vector is investigated. The verification results of the proposed technique, DML-eig [141], KISSME [142], ITML [138], and LDML [140] for the 10-fold cross validation tests are obtained using different principal components. The verification rates of these approaches are plotted versus these different principal components as shown

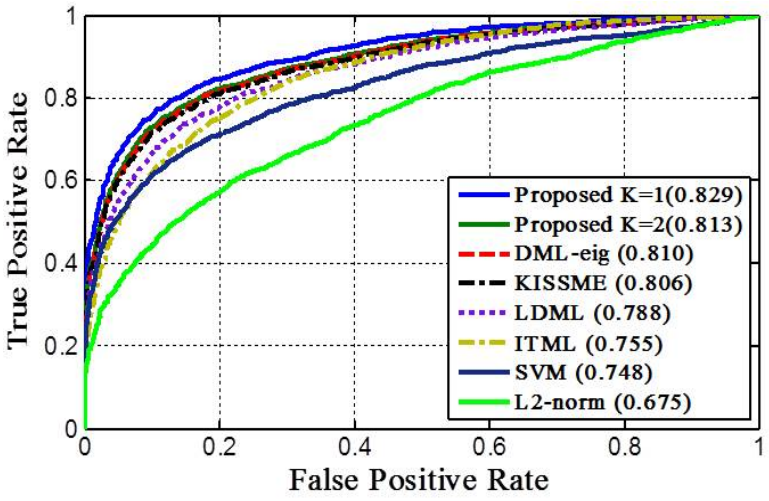


FIGURE 67: Face verification results on the LFW dataset

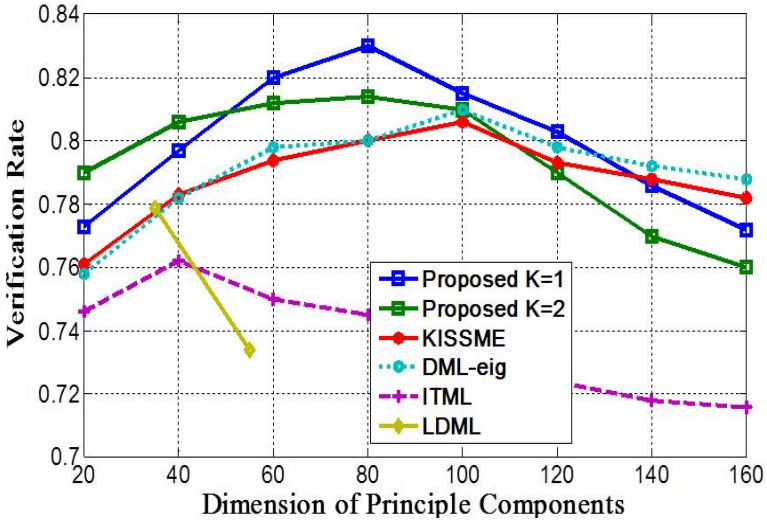


FIGURE 68: The verification rates of the proposed technique compared to other state-of-the-art versus different principal components. The results of LDML is copied from [140]

in Figure 68. When the principal components are 80 at $k = 1$ and 70 at $k = 2$, the proposed method achieves its best performance. These results can be used to draw conclusion that by learning more Mahalanobis distances smaller number of principle components can be used.

Finally, the performance of some "high-level" visual features [149], or attributes (e.g., gender, race, and age), of a face image that are insensitive to imaging conditions

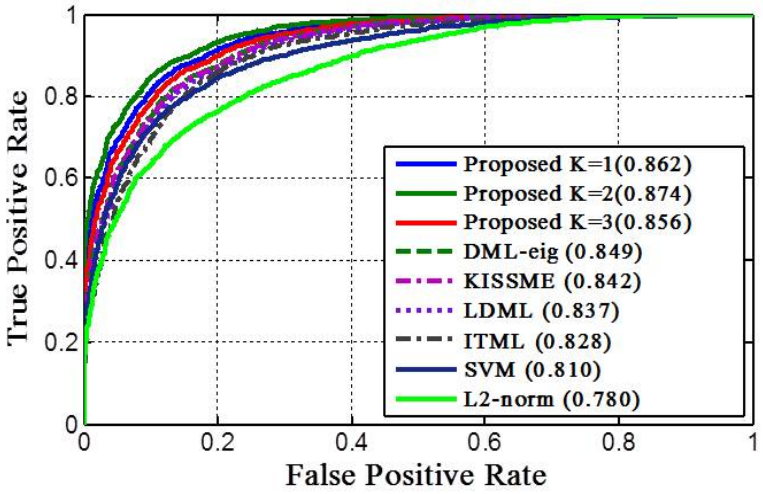


FIGURE 69: ROC curves for the attribute features-based face verification results on the LFW dataset.

(e.g., pose, illumination, and expression), for face representation is investigated. In Kumar et al. [149] approach, a face is warped using detected fiducial points. Then, the warped face is manually divided into parts e.g., the eyes, nose, and mouth. After that low-level features e.g., image intensities in different color spaces, edge magnitudes, and gradient directions, are constructed from these regions. An adaboost method is used to choose from the set of low-level features the one that drops error rates the most. For each attribute, this set of features is used to learn an attribute classifier (i.e., SVM with an RBF kernel) using data labeled by Amazon Mechanical Turk. Then, at the testing stage, these binary trained classifiers can be used to recognize the presence or absence of the attributes.

Figure 69 shows the ROC curves of the proposed approach compared to the-state-of-the-art approaches curves. The results reported in figures 69 and 67 can be used to conclude that there is no optimum number of Mahalanobis distances that always achieves best results. The optimum k in the case of low dimension vector length and large training dataset is usually around $k = 2$. In the high dimension vector length, the gain of using a nonlinear combination of Mahalanobis distances over one Mahalanobis distance is small.

In the following experiment, the PubFig dataset [149] in the face verification test is used. Similar to the LFW dataset, the PubFig dataset [149] is a challenging large-scale

database, consisting of 58,797 images collected from the internet. In total 200 people appear in the images. Also, like the LFW dataset, images of this dataset are divided into ten fully independent folds, which can be used for cross validation experiments. Each fold contains 1000 pairs of images from the same identity and another 1000 pairs of images from different identities.

Unlike the common faces dataset e.g., LFW, due to copyright issues, the authors of PubFig dataset do not distribute images in any format. Instead, the PubFig dataset consists of a list of image URLs that can be used to download the images by the user. However, this makes it impossible to exactly compare numbers, as image links disappear over time. To overcome this problem they extracted attribute features of these images and made them available for comparison. Therefore, in this experiment, we use these features, which are high-level descriptors and are robust against image variations compared to low-level features. Moreover, these high-level features help on evaluating the performance of the similarity measure learning algorithms.

Face verification results on the PubFig dataset are shown in figure 70. Where, Receiver Operator Characteristic (ROC) curves for KISSME [142], DML-eig [141], LDML [140], ITML [138], SVM [149], and the proposed method are generated. The proposed method reaches with an Equal Error Rate (EER) of 80.9% at $k = 2$, which is better than all the others.

As expected, the experiments illustrate that there is no optimum number of Mahalanobis distances that always achieves best results. Actually, this optimum k is based on both the length of the feature vector and the size of the training dataset. This is because for a feature vector of a length d, there are d^2 unknowns for each Mahalanobis distance. Since the proposed similarity measure has $k(k+1)$ Mahalanobis distances (Eq. 45), $k(k+1)d^2$ unknowns need to be learned. This means that one needs more than $k(k+1)d^2$ samples to learn the Mahalanobis distances otherwise overfitting occurs. Back to Figure 68, for a fixed number of samples, increasing d from 20 to 80 enhances the performance, since the number of samples is still greater than $k(k+1)d^2$. On the other hands, as d becomes more than 80, the number of samples becomes less than $k(k+1)d^2$ and the performance is degraded

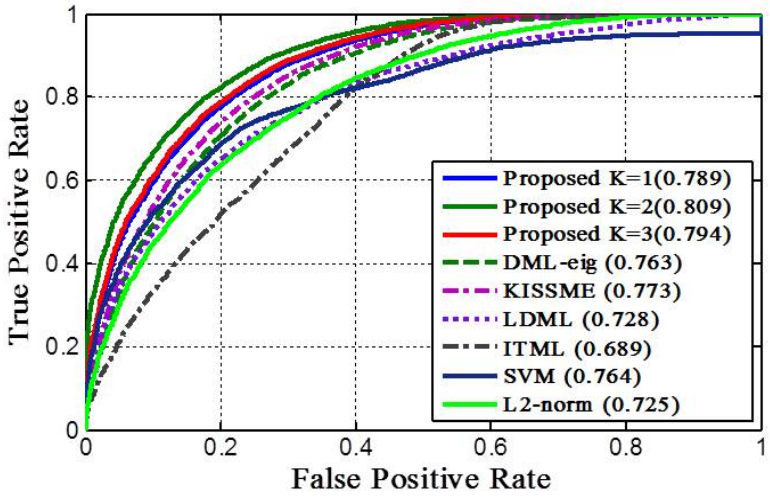


FIGURE 70: ROC curves for the attribute features-based face verification results on the PubFig dataset.

due to overfitting. Also, this explains why the performance in the case of $k = 2$ enhances faster than the case of $k = 1$ when $d = 20$ or 40, and it degrades faster when $d > 80$. This conclusion is confirmed in Figures 69 and 70. Since the dimensions of the attribute feature is small ($d = 40$), the performance in the case of $k = 2$ is better than the performance in the case of $k = 1$. However, in the case of $k = 3$ overfitting occurs.

CHAPTER 6
STEPS TOWARDS EMOTION RECOGNITION FOR AUTISTIC CHILDREN

The Centers for Disease Control and Prevention (CDC) conducted a survey study in 2006. In this study, they stated that one child in every 110 aged 8 years old in the United States is currently diagnosed with an Autism Spectrum Disorders (ASD). According to the national research council, children with ASD have major difficulties in expressions and emotions recognition. They can only show a little verbal and nonverbal communication. In other words, they have a problem in revealing their expressions and emotions, and in understanding others' emotions and expressions. Therefore, it is very hard to understand their emotions based on gesture and facial expressions.

6.1 Background

1. Emotion

There are differing theories and models regarding the relationship between bodily changes, cognitive processes and emotions. The most famous model is the representation of emotions in 2D valence/arousal space, see Figure 71, where valence represents the way one judges a situation, from unpleasant to pleasant and arousal expresses the degree of excitement felt by people, from calm to excited. Emotions drastically influence many aspects of human activity such as:perception, intention, communication, organization of memory, learning, attention, performance, goal generation, evaluation, social interaction and decision making. Emotion recognition may be studied using contact or contactless sensors. Contactless sensors include video, to capture facial expression and gesture, and microphones, for analysis of the vocal intonations [159]. Unfortunately, individuals with ASD

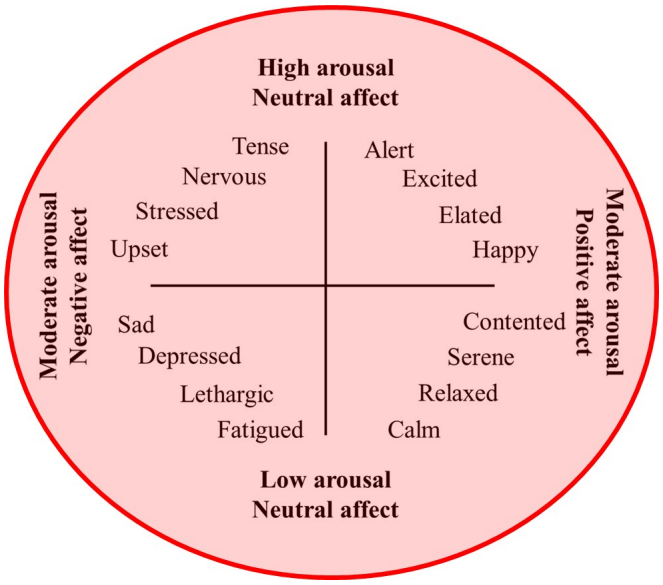


FIGURE 71: Representation of emotions in 2D valence/arousal space

often do not readily expose emotions; hence, these contactless sensors may not be a directly applicable. The other way to recognize emotion is using contact sensors that measure the physiological activities related to emotional states. The usage of physiological parameters in such context is beneficial because they are mostly under control of Autonomic Nervous System (ANS), which means that they are less affected by the conscious manipulations.

2. Physiological metrics for Emotion Recognition in Autism

Liu et al [177, 178] established an affective model to relate the physiological measures to the emotion state of children with ASD. The physiological measures were measured using a BioPac system. The BioPac system measured three different groups of physiological signals which are electromyogram (EMG), electrodermal activity (EDA), and cardiovascular measure. This physiological data was collected using contact sensors placed on chest, face, and one of the subject's hands, see Figure 72. Electromyogram signals have been used as strong indicators of emotion state for typical individuals [160]. While in this experiment, it is observed that it is less discriminatory than the cardiovascular and electrodermal activities. The reason behind that, is children with ASD often have nonverbal

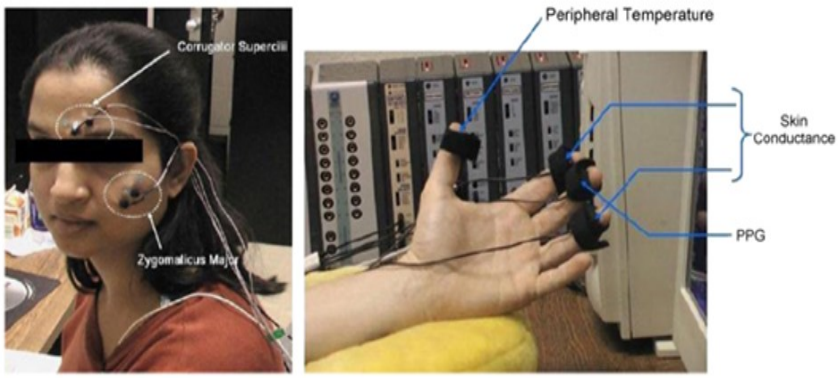


FIGURE 72: Electrode arrangements for collecting physiological data during an exercise, which invokes emotional activities of autistic individual (adopted from [177])

communicative impairments regarding expression of affective states (e.g., abnormal body postures and gestures and absence of facial expression), which reduce the discriminatory capability of EMG signals (e.g., muscle activities from both the corrugator supercilii and the zygomaticus major) to reveal affective cues of the participants. Cardiovascular and electrodermal activities are the main indicator of the emotion state for children with ASD. However, collecting the measurements for cardiovascular and electrodermal activities by contact sensors limit the usage of these measurements in real life situations.

3. Physiological Measurements by Thermal IR sensors

The heart consists of four chambers, namely the left and right atria and ventricles, which work in unison acting as a two-stroke pump to circulate the blood. During systole, the heart contracts and the blood heated in the core of the body is circulated through the various tissue layers via the arterial network and eventually reaches the skin via the capillaries. As blood passes through the capillary bed, the temperatures between the skin and blood equilibrate. During diastole, the heart expands and the blood exits the capillary bed via the veins to the venous return channels. As a result, the blood temperature in a vein represents the average temperature of the tissues drained by that vein. The blood is reheated as the process repeats. When the body attempts to maintain homeostasis, heat variations

resulting from the underlying vasculature are conducted through soft tissues so it can be measured measurable by an Infrared camera [174–176].

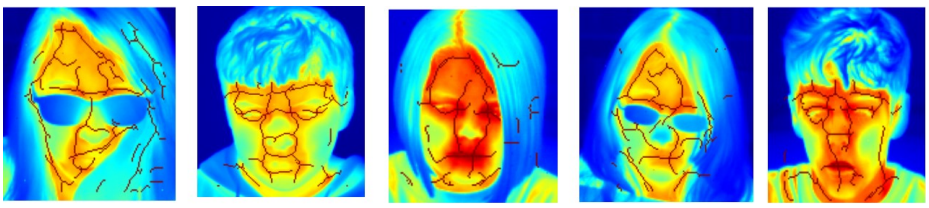


FIGURE 73: Samples of extracted underlying vasculature map in face image

While the carotid complex, STA/SO, and forearm are the most accessible superficial arteries for thermal IR measurements, the carotid artery is prone to being covered by fatty deposits while also located in a deformable region on the neck and the forearm is not nearly as convenient or practical to image as the face. The superficial temporal arterial (STA) branches in the forehead region are identified as the regions of interest (ROI) [176]. The STA is a continuation of the carotid artery on the neck, which splits into the parietal, frontal, and supra orbital (SO) branches. The skin in the forehead over the STA/SO is less susceptible to movements introduced by facial expression, neck movement, and breathing. These structures are sufficiently large and close to the skin's surface, creating a distinct heat signature. Figure 74 shows the regions of interest which is the center of forehead area around supraorbital branch for measuring heart activities. Moreover, the nasal tissue area is another region of Interest for measuring the breathing function. In [174–176], the dominant heart rate frequency is estimated by averaging the power spectra of each pixel in the vascular map in the regions of interest. The power spectra is calculated by adaptive Fourier-based signal filtering method on the pixel appearance over N frames. Figure 73 shows the automatic computer extract of underlying vasculature in thermal images.

6.2 Related Work

It is one of the important missing components for making the contactless emotion recognition for children with Autism Spectrum Disorders is tracking and detecting a Region

of Interest (ROI). However, robust tracking of a region Of Interest (ROI), center of forehead area around supraorbital branch or nasal tissue area, is a challenging task, therefore it is the focus of this work. Recently, region-based tracking approaches have seen very popular in visual tracking. Tracking a region of interest (ROI) is accomplished by searching for the best match for the ROI. Some methods limit the search domain in an area where the ROI is expected [180], while others use the state predication [181] (e.g. particle filter [179]). The ROI can be represented as a template or a generative model. The template can be static (i.e., the template does not change over time) or adaptive (i.e., a new template is always extracted from the previous frame).

Jepson et al. [163] proposed a statistical appearance template, which gives pixels with stable behavior heavier weights than pixels with less stable behavior. This algorithm shows a good performance in visible images, however, it fails in thermal images. This is due to that it does not support abrupt changes in the appearance and saturates after a long tracking periods (i.e., it needs re-initialization). While tracking the ROI for vital signs should support fast changes in the appearance since the appearance changes with the blood flow in the body.

In the thermal imaging domain, Dowdall et al. [164] proposed a network of particle filters trackers driven by deterministic template function. Each pixel in the template is updated if a respective difference exceeds a predetermined threshold. This approach also could not handle the abrupt changes in the appearance and is vulnerable to the drifting problem. Zhou et al. [161] proposed a particle filter tracker driven by a probabilistic template mechanism, which is based on the Matte algorithm [162]. This algorithm handles changes in appearance better than the algorithm proposed in [164]. Certainly these methods face several challenges in practice such as abrupt motion of the ROI, frame-cut, and leaving of the ROI the field of view.

6.3 Proposed Framework

The focus of this work is the detection and tracking of ROI, center of forehead area

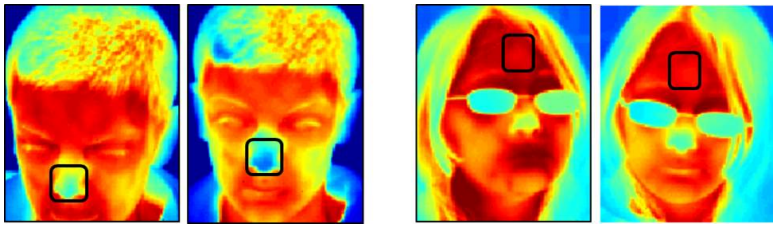


FIGURE 74: The variation in the appearance in the two different regions of interest.

around supraorbital branch or nasal tissue area, in thermal video. This framework handles the two unique problems for detecting ROI for vital signs. The first one is the small size of ROI (e.g., around 20-30 pixels height and width)which makes confusion in the detection since many other areas in the face may have same appearance. The second challenge is that the appearance of ROI changes with time in response to change in cardiovascular activities, as shown in figure 74.

The proposed framework, as shown in figure 75, consists of three main modules: an adaptive particle filter tracker of the ROI, a detector of the ROI, and a unit of integration and learning decision. A thermal video is processed by the particle filter and the detector independently. The outputs (i.e., bounding boxes of ROI) from the detector and from the particle filter tracker are combined in the integrator and learning decision module into one ROI. If both outputs can not be combined, ROI is assumed to be invisible. Moreover, the integrating and learning module decides if the frame work needs to be learned or does not. In the learning stage, the template in the particle filter is updated and positive and negative samples are generated from current frame to update the random ferns classifier.

The tracker is based on adaptive particle filter to overcome the problem of appearance change in ROI. While the detector is based on a novel concept which uses supporters in detection of ROI to handle the challenge of small size of ROI. The supporters are five facial feature points that have more discriminative appearance than ROI. The detection is done by detecting ROI that follows the geometric constrain with the facial feature points.

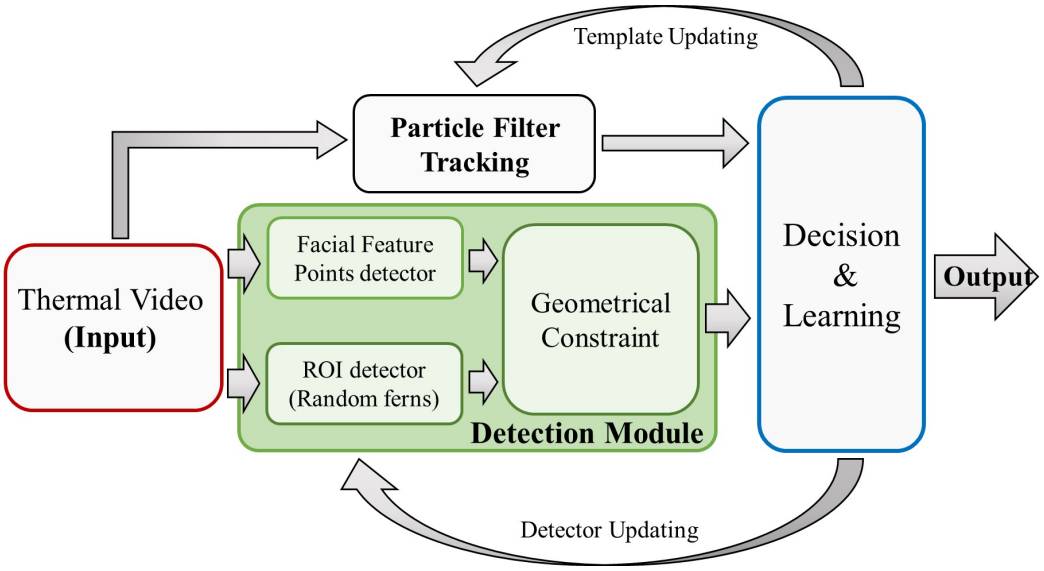


FIGURE 75: Proposed long term tracking framework in thermal imaging for vital signs

1. Particle filter Tracking of ROI

Particle filtering to track the ROI's position in the current frame, based on template matching is used. The particle filter tracker uses $r = 60$ particles (candidate ROIs) in a single iteration per frame. The tracker output ROI C_t^* is selected from the candidate ROIs (particles) by choosing the closest one to the template. The similarity measure between the template T at time $t-1$ and one of the candidate ROIs C^r is estimated as follows.

$$S(T, C^r) = \frac{\Sigma_i (T_{t-1}[i] - \mu(T_{t-1}))(C_t^r[i] - \mu(C_t^r))}{\sigma(T_{t-1})\sigma(C_t^r)}, \tag{47}$$

where $\mu(T_{i-1})$ and $\mu(C_t^r)$ denote the means of the template at time $t-1$ and the r candidate at time t, respectively. $\sigma(T_{t-1})$ and $\sigma(C_t^r)$ are the standard deviations. The index i represents the i^{th} pixel.

Since the appearance of ROI exhibits a large variation over time, the template should be updated. Updating the template every frame makes the tracker being prone to drifting. Unlike Zho et al. [161], the template is updated using the output of the integrator module whenever the learning decision module is permitted for learning the framework. The updated template, at time t, will be a weighted sum of the previous template and the integrator

output P_t^* from the current frame:

$$T_t[i] = w_t[i]P_t^*[i] + (1 - w_t[i])T_{t-1}[i], \tag{48}$$

where $w_t[i]$ is called Matte value [162], which indicates the necessary degree of updating for each pixel. To compute the matte of the current ROI, the intensity of each pixel in the image $I[i]$ is assumed to be a convex combination of a stable $S[i]$ and an unstable $U[i]$ component.

$$I[i] = w[i]S[i] + (1 - w[i])U[i]. \tag{49}$$

Various methods have been proposed to solve for $w_t[i]$ [162]. The cost function proposed by [168] is used in this work since it considered the spatiotemporal smoothing term, which is very important in the tracking. The estimation of the Matt value is formulated as a quadratic optimization problem with respect to w_t as follows.

$$w_t^* = \arg\min_{w_t}(w_t^T L_t w_t + (w_t - w_{t-1})^T(w_t - w_{t-1}) + \epsilon(w_t - d_t)^T D_t(w_t - d_t)), \tag{50}$$

where ϵ is chosen to be a large number 10^3 to avoid updating seeds in the optimization. L_t is the the similarity measure between a pair of pixels in the integrator output module P_t^*. D_t is a diagonal matrix whose diagonal elements are 1 for seeds (stable and unstable) and 0 for all other pixels. d_t is a vector containing 1 for unstable seeds and zero for all other pixels. The iterative generalized minimal residual method linear equation solver is used to solve the optimization formula.

The stable and unstable seeds from P_t^* are extracted based on the following inequalities:

$$|P_t^*[i] - T_{t-1}[i]| < \lambda_s, \quad and \quad |P_t^*[i] - T_{t-1}[i]| > \lambda_u. \tag{51}$$

where λ_s and λ_u are predetermined thresholds. In [168], authors showed that choosing these values is flexible and the performance is not very sensitive to this choice. In this paper, these thresholds are chosen to be 5 and 15, respectively.

2. Detection Region Of Interest (ROI)

Region of Interest (ROI), center of forehead area around supraorbital branch or nasal tissue area, detection is a task of localizing the ROI in an input image. Sliding window-based approaches show great success in the object detection, where the input image is scanned by different sizes of windows and for each window a decision is made if the window contains the object or does not. This work focuses on building online ROI detector where an offline model for ROI does not exist. Building an ROI detector, which is based on an offline model, is a challenge due to the following reasons. The ROIs are areas, which have a signification appearance variance over time. This is due to heart activity and breathing cycle. Also, The ROI's appearance differs from subject to subject and does not have a shape structure. Furthermore, the online ROI detector is flexible enough to be used for any ROI. In vital signs measuring, the ROIs differs in their positions based on the interested physiological measure (e.g., the nose area in measuring breathing cycle, part of forehead region in the blood pressure waveform). In this work, only five facial feature points are detected as supporters for detecting ROIs to construct a geometrical constraints for the position of the ROIs. Figure 77 shows facial feature points that collaborate in detecting ROI.

2..1 Classifier and ROI representation Due to the efficiency of the randomized ferns classifier, which is widely used in recognition [170] and tracking [167], it has been deployed in this work to detect the ROI in every frame. Randomized ferns were originally proposed by Ozuysal et al. [171] to increase the speed of randomized forest approach [182]. Unlike the tree-structure in randomized forest, ferns have non-hierarchical structures and consists of a number of binary testing functions. In this work, each fern classifier is learned by a set of Binary Pattern features. The Binary Pattern features are generated by performing pixel comparisons in the ROI similar to Binary Robust Independent Elementary Features (BRIEF) [184] and Oriented FAST and Rotated BRIEF (ORB) [183] descriptors. Each leaf in a fern records the number of added positive and negative samples during training. For a test sample, its evaluation by calculating the binary pattern features leads it to a leaf in the

fern. After that, the posterior probability for that test sample (i.e., feature vector f_i) to be labeled as an ROI (i.e., $y = 1$) by fern j is computed as $Pr_j(y = 1|f_i) = p/(p + n)$ where p and n are the numbers of positive and negative samples recorded by the leaf. The posterior probability is set to zero if there is no record in the leaf, respectively. The final probability is calculated by averaging the posterior probabilities given by all the ferns (N):

$$Pr(y = 1|f_i) = \sum_{j=1}^{N} Pr_j(y = 1|f_i) \tag{52}$$

The number of ferns is chosen to be 10. Each fern learn on an 8-bit comparison in the ROI. Thus the ROI is represented by 80 binary bits each bit represents a pixel comparison in the ROI.

2..2 Geometrical Constraint The five collaborated facial feature points and the center of ROI construct a shape vector Z. This shape vector Z is given by $[x_{roi} + iy_{roi} \;\; x_l + iy_l \;\; x_r + iy_r]$. This vector represents the spatial coordinates of the centroid of the ROI, and the position of the five facial feature points. To be in-plane rotation invariant, this shape vector is modeled as multivariate complex gaussian distribution. The mean vector μ_z is initially estimated as the center position of manually detected ROI and the positions of detected facial feature points in the first frame. The covariance matrix Σ_z is modeled as a random noise with zero mean and unit standard deviation. Finally, the confidence score of the output of randomized ferns classifier based on geometric constraints is given by:

$$Pr(y = 1|Z) = \frac{1}{\sqrt{(2\pi)^3|\sum_z|}} \exp(\frac{-1}{2}(Z - \mu_z)^T \Sigma_z^{-1}(Z - \mu_z)). \tag{53}$$

This shape constraint works as a filter for the output of random ferns classifier. The overall confidence level of the detector is given by

$$Pr(y = 1) = Pr(y = 1|Z)Pr(y = 1|f_i) \tag{54}$$

3. Integrating and learning Module

The integrator combines the bounding box of the tracker and the bounding boxes of the detector into a single bounding box. The ROI is declared as not visible if there is no

detector output and the tracking confidence level is less than 0.5, or the confidence levels of all bounding boxes are less than 0.2. If the detected boxes overlap with the bounding box of the tracker, the integrator outputs the maximally confident bounding box which is given by:

$$Total\ Confidenec = Pr(y = 1)\ Overlap(C^*, fi), \tag{55}$$

where $Overlap(C^*, fi)$ is the percentage of area of overlap between the detector and tracker bounding boxes. In the case of no overlapping boxes with the bounding box of the tracker, the integrator outputs the maximally confident bounding box that correspondence to $\max(Pr(y = 1), S(T, C^*))$.

The learning step is performed when the tracker is valid with a confidence more than 0.8 and no detector output, or there is a detection box that overlaps with tracker output and both of them have confidence more than 0.2. In the learning procedure, positive and negative samples are generated from the current frame to update the random ferns. The positive samples are warped versions from output of the integrator. The negative samples consist of random boxes that are generated far from the integrator's output box and the boxes that feed into the integrator and the integrator decided that they are false positives. Also, the template of the tracker is updated as shown in subsection 6.3.1.

6.4 Experimental Results

The experiment is conducted on collected thermal faces database from Bluegrass autism center. The motivation behind collecting this dataset is that there is no public thermal imaging datasets that capture the face movement for tracking to measure the vital signs.

The collected dataset consists of 40 subjects (children aged 6-8 years), 20 control subjects and 20 autistic children. Figure 76 shows a subset of subjects who participate in this study. The subject is captured while playing a face recognition game, with two different levels of difficulty to excite different emotional states of the subject. In the easy level, the subject tries to match visible images with visible images that are captured with

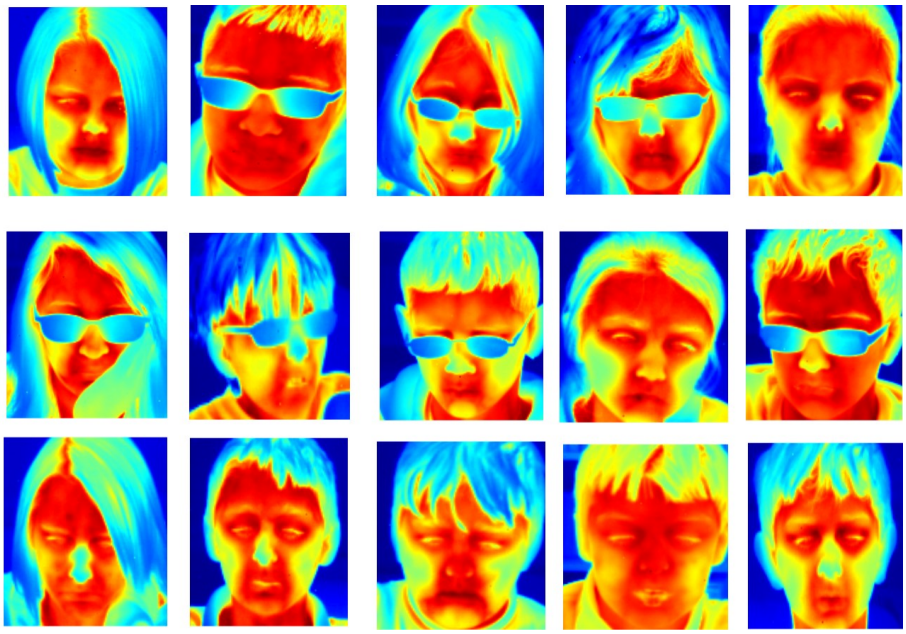


FIGURE 76: Thermal images for subset of children with ASD who participate in this study.

time lapse. In the difficult level, the subject tries to match a vascular map for a face to other vascular map. The dataset is collected on these children since the ultimate goal of the project is emotion understanding for children with ASD. Each subject is captured for at least 8 minutes with N2-cooled Indigo Phoenix QWIP LWIR Camera System with Real Time Imaging Electronics (Product 420-0011-007, Rev. 120) and Talon Ultra 5.2 image acquisition software (Ver. 4.5.1.27) to acquire 1024 video frames at 20 fps, spatial resolution of 320x256 and thermal resolution (MRTD) of 35mK. The regions of interest in this experiment are the nasal tissue area and the center of forehead area around the supraorbital artery.

To quantify how well each tracker performs, one needs to have the ground truth location of the ROI and compare the detected ROI with the ground truth through time lines. However, with hundreds of thousands of frames in the dataset (40 subjects, 8 min/subject, 1200 frames/min), manual ground truth is not practical. Instead, the data is downsampled to 100 frames/min and two videos are clipped per subject: every video is three minutes to reduce the number of frames. The detected ROI using the proposed approach and other approaches are compared with the manually annotated ground truth.

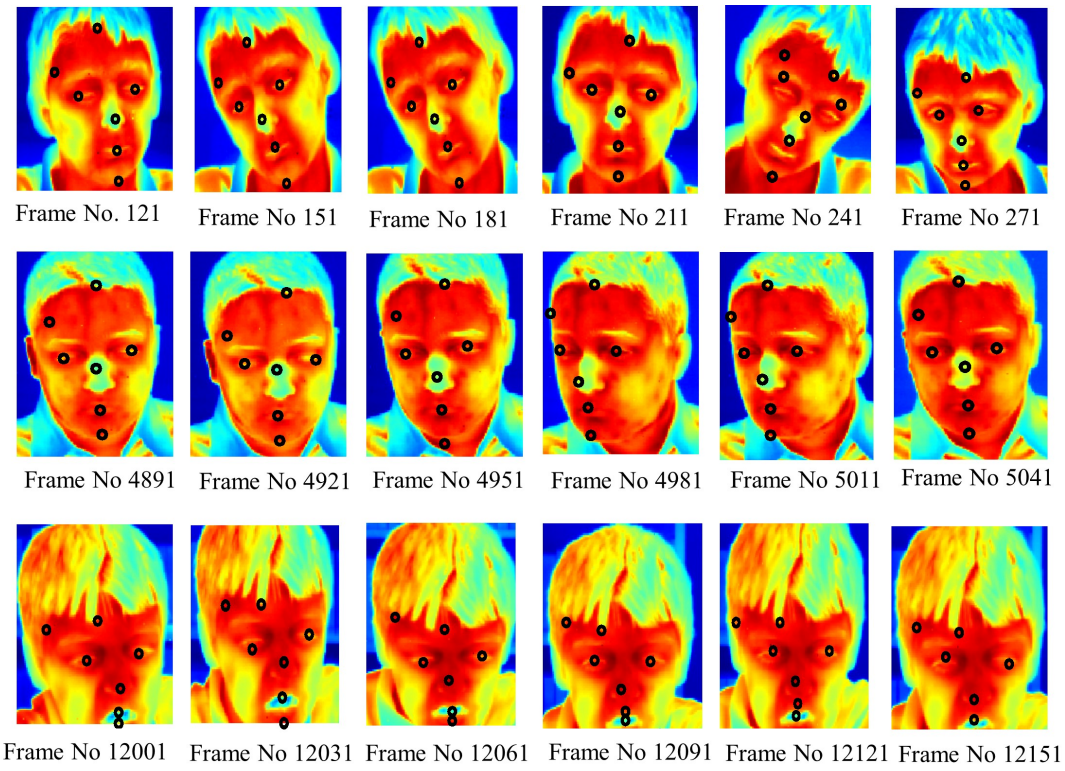


FIGURE 77: Samples of the detecting and tracking facial feature points.

Figure 77 shows sample of the proposed approach results for detecting the region of interest and detecting other facial feature points.

Tables 10 and 11 show the performance of the proposed tracker comparing with other recently proposed trackers on four autistic subjects and four control subjects on the nasal tissue area and the center of forehead, respectively. The performance evaluation is measured by Precision, Recall, and F-measure which are calculated from true positive, false positive, false negative, and true negative quantity. The performance of the tracker for each subject is evaluated on two videos. The first video is captured in the beginning of recording while playing the easy game so the movement is not big and the appearance of the ROI does not change very fast, since the subject is in a normal situation, the results are shown in the upper row for each subject in the table. The second video is captured at the end of session while a subject playing the difficult game so the movement is large since he/she is stressed, the results are shown in the lower row for each subject in the table. The processing

TABLE 10: A comparison of the nasal tracking results using the proposed algorithm and other alternatives.

Group	Proposed	TLD [167]	STM [168]	OB [172]	BS [173]
Autistic	1.00/1.00/1.00	0.89/0.91/0.90	0.95/1.00/0.98	0.86/0.89/0.87	0.90/0.81/0.86
	0.90/0.89/0.90	0.86/0.87/0.86	0.69/0.61/0.65	0.69/0.06/0.11	0.98/0.03/0.07
Autistic	1.00/1.00/1.00	1.00/1.00/1.00	1.00/1.00/1.00	1.00/0.92/0.96	1.00/0.83/0.91
	1.00/1.00/1.00	1.00/0.96/0.98	0.98/0.73/0.84	0.98/0.46/0.63	0.89/0.49/0.64
Autistic	0.95/0.93/0.94	0.80/0.83/0.82	0.88/0.85/0.87	0.76/0.82/0.79	0.83/0.84/0.83
	0.75/0.53/0.62	0.50/0.47/0.48	0.50/0.53/0.52	0.36/0.40/0.38	0.95/0.09/0.16
Autistic	0.97/0.93/0.95	0.87/0.88/0.88	0.72/0.83/0.77	0.94/0.84/0.89	0.98/0.62/0.76
	0.96/0.96/0.96	0.86/0.98/0.92	0.70/0.76/0.73	0.90/0.14/0.25	0.93/0.32/0.47
Control	1.00/0.95/0.97	1.00/0.92/0.96	0.98/0.97/0.98	0.98/0.91/0.94	0.99/0.89/0.94
	0.96/0.92/0.94	0.92/0.76/0.83	0.85/0.64/0.73	0.94/0.35/0.50	0.93/0.66/0.77
Control	0.98/0.92/0.95	0.96/0.82/0.89	0.97/0.99/0.98	0.98/0.87/0.92	0.99/0.67/0.80
	0.95/1.00/0.97	0.83/0.91/0.87	0.82/1.00/0.90	0.97/0.22/0.36	0.86/0.69/0.77
Control	1.00/1.00/1.00	1.00/1.00/1.00	1.00/0.00/0.97	1.00/0.89/0.94	1.00/0.93/0.96
	0.98/0.86/0.91	0.94/0.96/0.95	0.90/0.57/0.70	0.75/0.62/0.68	0.91/0.67/0.77
Control	0.99/0.87/0.93	0.98/0.89/0.93	0.98/0.89/0.93	0.92/0.46/0.61	0.97/0.58/0.73
	1.00/0.43/0.60	1.00/0.21/0.35	0.89/0.12/0.20	0.83/0.05/0.10	0.82/0.29/0.43

speed of all algorithms are greater than 25 fps (e.g. real time processing). The proposed tracker has the best performance, since it enhances the performance of TLD [167] by adding geometrical constraint to remove false positives and avoid drifting. Also, proposed tracker combines the advantages of STM tracker [168] in the framework. Combing STM tracker in the framework avoids drifting and solves the problem of when the ROI re-enters the field of view.

Figures 78 shows samples of the nasal tracking results using the proposed algorithm and other alternatives.

TABLE 11: A comparison of the forehead tracking results using the proposed algorithm and other alternatives.

Group	Proposed	TLD [167]	STM [168]	OB [172]	BS [173]
Autistic	0.99/0.90/0.94	0.76/0.77/0.77	0.85/0.90/0.87	0.81/0.98/0.8869	0.90/0.79/0.84
	0.84/0.77/0.81	0.57/0.71/0.63	0.54/0.53/0.53	0.98/0.20/0.34	0.86/0.82/0.84
Autistic	0.99/0.90/0.94	0.67/0.84/0.75	0.86/0.82/0.84	0.56/0.51/0.53	0.77/0.81/0.79
	0.92/0.84/0.88	0.46/0.58/0.51	0.91/0.63/0.75	00.84/0.69/0.76	0.93/0.40/0.56
Autistic	0.82/0.81/0.82	0.88/0.71/0.79	0.85/0.78/0.81	0.70/0.76/0.73	0.59/0.78/0.67
	0.57/0.43/0.49	0.38/0.49/0.43	0.36/0.42/0.39	0.63/0.91/0.75	0.26/0.32/0.28
Autistic	0.90/0.85/0.87	0.66/0.98/0.79	0.63/0.86/0.73	0.760.54/0.63	0.87/0.81/0.84
	0.92/0.87/0.90	0.78/0.92/0.89	0.55/0.58/0.57	0.80/0.13/0.22	0.8670/0.08/0.85
Control	0.98/0.86/0.91	0.98/0.92/0.95	0.97/0.99/0.98	0.99/0.67/0.80	0.97/0.83/0.86
	0.95/1.00/0.97	0.92/0.46/0.61	0.82/1.00/0.90	0.97/0.22/0.36	0.86/0.67/0.75
Control	0.98/0.92/0.95	0.96/0.82/0.89	0.97/0.99/0.98	0.98/0.87/0.92	0.99/0.65/0.75
	0.95/1.00/0.97	0.83/0.91/0.87	0.82/1.00/0.90	0.97/0.22/0.36	0.86/0.68/0.77
Control	0.97/0.99/0.98	0.98/0.92/0.95	0.97/0.99/0.98	0.98/0.92/0.95	1.00/0.93/0.96
	0.99/0.87/0.93	0.98/0.89/0.93	0.98/0.89/0.93	0.92/0.46/0.61	0.97/0.58/0.73
Control	0.98/0.86/0.91	0.94/0.96/0.95	0.90/0.57/0.70	0.75/0.62/0.68	0.91/0.67/0.77
	0.96/0.92/0.94	0.92/0.76/0.83	0.85/0.64/0.73	0.94/0.35/0.50	0.86/0.69/0.77

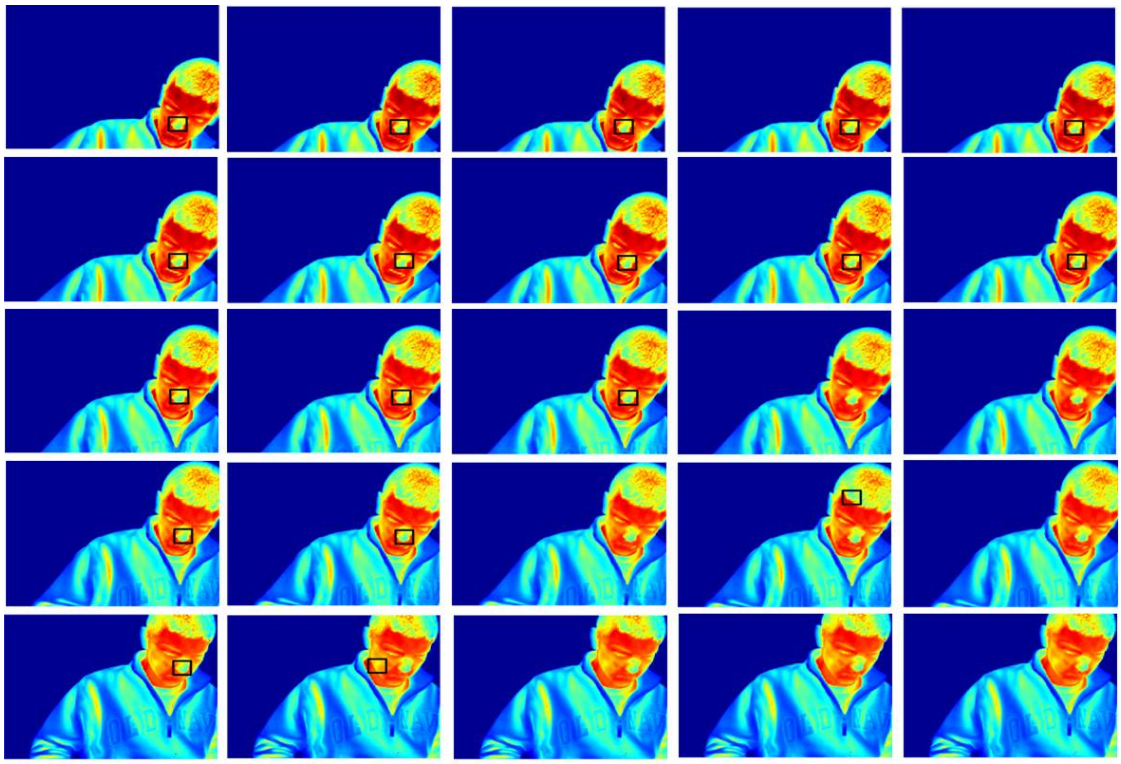


FIGURE 78: Samples of the nasal tracking results using the proposed algorithm and other alternatives.

CHAPTER 7
CONCLUSIONS AND FUTURE WORK

Face understanding is considered one of the most important topics in computer vision field since the face is a rich source of information in social interaction. Not only does the face provide information about the identity of people, but also their membership in broad demographic categories of human (including sex, race, and age), and about their current emotional state. Facial landmarks extraction is the corner stone in the successful development of different face analysis and understanding applications. In this dissertation, a novel face modeling is designed for facial landmarks detection in unconstrained real life environment from different image modality, infra-red and visible images.

The model for landmark detection has been studied in the two face understanding applications which are face recognition from visible images and physiological measurements for autistic individual from thermal images. However maturity of face recognition systems, recognize face identity with different poses, expressions and lighting conditions from a complex background is still unsolved problem even with accurate detection of landmark. To handle illumination and expression variation, learning similarity measure between two face image representation is proposed that it only responds to the difference in identity and filter the illumination and expression variation. The pose challenges are tackled by two new approaches where the three dimensional face representations from images has been involved.

The other framework for the face understanding application is a physiological measures for autistic individual from infra-red images. Accurate detecting and tracking of superficial temporal arteria while the subject is moving, playing, and interacting in social communication is challenging while it is must. The challenges come from two sources: the appearance of the STA region changes over time. The appearance is not discrimina-

tive enough from other area in face region since it is few pixel width and height. A novel concept,supporter collaboration, in detection is introduced.

7.1 Summary of Contributions

This dissertation proposed a novel human face model for facial landmarks detection. The facial landmarks detection has been apart for two face understanding applications/frameworks face recognition from visible images and emotion recognition for children with Autism spectrum disorder from thermal images. The proposed model incorporates the part based model with holistic face information. The part based model is based on soft combining a texture classifier with complex Bingham distribution as a shape representation. The texture classifier is built by a support vector machine classifier that used a novel feature representation, which is called the pixel difference feature. The complex Bingham distribution is adapted from statistical community into computer vision for face shape representation since it is invariant to the in-plane rotation which gives this model superiority. An energy minimization function is formulated to incorporate information from both the texture classifier and the shape model simultaneously. In the final stage, global information is used to improve the results of the part based model by using regression model that does not penalize the outliers of the human face shape due to extreme expression, occlusion, and different ethnicity.

In the presented face recognition framework, the face image is represented by concatenating and transforming the appearance around the facial landmarks. This representation is not robust for illumination, expression, and pose. To handle illumination and expression, learning a similarity measure between two face representations is proposed that it only responds to the difference in identity and remains almost constant with illumination and expression variation. The learning similarity measure aims to discard bad features selectively in each individual matching circumstance which should not be used on computing the measure. The proposed similarity measure is derived from a statistical inference as a non-linear combination of mahalanobis distance. This similarity measure is evaluated on

two different datasets in the wild which are Labeled Face in the Wild and Public Figure using different face representations. The proposed similarity measure outperforms the accuracy of state-of-art similarity measures, which are based on learning using the equivalent constraint.

The pose invariant face representation is achieved by proposing two approaches. The first approach is based on dynamic weighting of the contribution of each facial landmark in the similarity measure between two face representations. This method has been called ”dynamic weighting for pose invariant face recognition”. Dynamic weights are assigned for each facial feature at each pose based on the overlapping scores. This score is based on the number of pixels in the patch in the frontal gallery image and the captured pose image. These pixels correspond to the same vertices in the 3D person's head. The second approach for the pose invariant face representation is based on the existence of multiple images for a subject at different poses in the gallery, therefore each identity is represented by multiple face representations. These images are virtual images that are synthesized from a reconstructed 3D face. The 3D face is reconstructed using two different techniques. The first technique is based on a single image where the problem of 3D construction is formulated as regression relation between face landmarks in two dimensional space and the dense three dimensional face shape. This approach is called ”Hybrid 2D-3D”. The second technique for 3D face reconstruction is based on geometric stereo. In geometric stereo, two different cameras simultaneously captured two images and the distance between two cameras is known.

The approaches for pose invariant face recognition has been tested on three public datasets which are FERET, CMU-PIE, and Multi-PIE, and in-house collected dataset (UofL-EWA). UofL-EWA dataset had been collected since there is no public available dataset for stereo. The experimental results show that dynamic weighting for pose invariant face recognition achieves state-of-art results in indoor environments with limited variation in expression and illumination since it can not be combined with learned distance metric. While ”Hybird 2D-3D” and stereo face recognition outperform the dynamic weighting approach in uncontrolled environment since the distance metric learning can be combined

with these approaches for pose handling while they show slightly decrease in performance in indoor environment as compared to the dynamic weighting approach.

The dissertation moves the research in non-intrusive physiological measurements for emotion recognition for children with autism spectrum disorder forward by proposing framework for continuously tacking and detection superficial temporal arterial (STA) branches from thermal camera while the subject is moving, playing, and interacting in social communication. A long term tracking and detection framework is proposed. The proposed framework consists of three main modules: (1) an adaptive particle filter tracker for (STA) branches area which is used to overcome continuous change in the appearance with changing the blood flow, (2) online detector that used a new concept which is called supporters to avoid the confusion that results from the small size of STA branches area, and (3) a unit of integration and learning decision. Moreover, a dataset consists of thermal and visible videos for children with autism spectrum disorder and controlled at age 6-8 years old has been collected while they engaged in playing game that has different difficulty levels.

7.2 Limitations and Suggested Future Directions

This work tackled the following face recognition problems: illumination, expression, and pose variations. Pose is handled using information about the 3D structure of the face while illumination and expression variations in recognition are handled by learning similarity measure. The main assumption behind dividing the illumination and expression into one category and the pose in another category is that the pose is the rotation of the subject's head in the 3D dimensional space while illumination and expression can be understood from the two dimensional space, image plane. In extreme illumination and expression, this assumption is not valid and a 3D model is needed. The main challenge is combining different 3D models that are used to solve different face challenges. It is still an open question in face recognition research with an unconstrained environment.

Face is represented by features captured around facial landmarks using hand crafted

features. Recently, many researchers have moved toward combining face representation with learning a similarity measure using a convolution deep belief network known as deep learning. These approaches show similar performances to the approaches that separate face representation and similarity measure when the data available for learning similarity measure in order of thousands of the face images. However, deep learning may significantly improve the performance when the training data in order of millions/billions face images which is not publicly available yet since the learning algorithm for the similarly measures may not scalable to big dataset.

The dissertation has additionally attempted to move forward research in the emotion recognition for children with autism spectrum disorder with the useof thermal camera by detection and tracking superficial temporal arterial. Thus is very beneficial for real life situations while the subject is moving, playing, and interacting in social communication. However, this work does not go explore ideas of movement on the model of extracting physiological measurement from thermal images.

REFERENCES

[1] P. Belhumeur ,J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 19, no. 7, pp. 711-720, Jul. 1997.

[2] M. Turk ,and A. Pentland, "Face recognition using eigenfaces", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586-591.

[3] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognit.*, vol. 42, no. 11, pp. 2876-2896, Nov. 2009.

[4] X. Tan, S. Chen, Z. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey", *Pattern Recogn.*, vol. 39, no. 9, pp. 1725-1745, Sep. 2006.

[5] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[6] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local Binary Patterns and Its Application to Facial Image Analysis: A Survey," *IEEE Trans. on Sys., Man, and Cyber., Part C: Applications and Reviews,*, vol. 41, no. 6, pp. 765-781, Dec. 2011.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, "Cascade object detection with deformable part models," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2010, pp. 2241-2248.

[8] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proceedings of International Conference in Computer Vision*, 1998, pp. 555-562.

[9] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Optical Society of America, Journal, A: Optics and ImageScience*, vol. 2, pp. 1160-1169, 1985

[10] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91-110, Nov. 2004.

[11] I. Dryden and K. Mardia, *Statistical Shape Analysis,*, John Wiley and Sons, 1998.

[12] X. Gao, Y. Su, X. Li, D. Tao,"A review of active appearance models," in *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 40, no. 2, pp. 145-158, 2010.

[13] S. Zhou, D. Comaniciu, "Shape regression machine," in *Proceedings of the 20th International Conference on Information Processing in Medical Imaging*, 2007, pp. 13-25.

[14] Y. Freund, R. Schapire, "decision-theoretic generalization of on-line learning and an application to boosting," in *Journal of Computer and System Science*, vol.55, pp. 119-139, 1997.

[15] J. Friedman, T. Hastie, R. Tibshiani, "Additive logistic regression: a statistical view of boosting," in *The Annals of Statistics*, vol. 38, no. 2, pp. 337-374, 2000.

[16] X. Cao, Y. Wei, F. Wen,J. Sun, "Face alignment by explicit shape regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2887-2894.

[17] X. Burgos, P. Perona, P. Dollr, " Robust face landmark estimation under occlusion", in *Proceedings of International Conference in Computer Vision*, 2013.

[18] N. Duffy, "Boosting methods for regression," in *Machine Learning*, vol 47, pp. 153-200, 2002.

[19] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... Buffy - Automatic Naming of Characters in TV Video," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 92.1-92.10.

[20] D. Cristinacce, T. Cootes, and I. Scott, "A Multi-Stage Approach to Facial Feature Detection," in *Proceedings of the British Machine Vision Conference*, 2004, pp. 231-240.

[21] M. Valstar and B Martinez and X Binefa and M Pantic, "Facial Point Detection using Boosted Regression and Graph Models," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2010, pp. 2729-2736.

[22] E. Mostafa, and A. Farag, "Complex Bingham Distribution for Facial Feature Detection," in *Proceedings of European Conference on Computer Vision Workshops*, 2012, pp. 330-339.

[23] O. Hamsici, and A. Martinez, "Rotation Invariant Kernels and Their Application to Shape Analysis," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 31, no. 11, pp. 1985-1999, Nov. 2009.

[24] N. Wang, X. Gao, D. Tao, X. Li, "Facial Feature Point Detection: A Comprehensive Survey", in emphInternational Journal of Computer Vision, accepted to appear, 2015.

[25] O. eliktutan, S. Ulukaya, and B. Sankur,"Abstract A comparative study of face landmarking techniques," in emphJournal on Image and Video Processing, 2013.

[26] T. Cootes, G. Edwards, C. Taylor, "Active appearance models," in *Proceedings of European Conference on Computer Vision*, 1998, pp. 484-498.

[27] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 23, no. 6, pp. 681-685, Jun. 2001.

[28] I. Matthews, S. Baker, D. Cooper, and J. Graham, "Active Appearance Models Revisited," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 135-164, Nov. 2003.

[29] T. Cootes, C. Taylor,"An algorithm for tuning an active appearance model to new data", in *Proceedings of British Machine Vision Conference*, 2006, pp. 919-928.

[30] J. Saragih, S. Lucey, J. Cohn, "Deformable face fitting with soft correspondence constraints," *Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1-8.

[31] B. Amberg, A. Blake, and T. Vetter, "On compositional Image Alignment, with an application to Active Appearance Models," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2009, pp. 1714-1721.

[32] A. Ashraf, S. Lucey, T. Chen, "Fast image alignment in the Fourier domain," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2480-2487.

[33] A. Asthana, S. Lucey, R. Goecke, "Regression based automatic face annotation for deformable model building". *Pattern Recognition*, vol. 44, pp. 2598-2613, 2011.

[34] M. Hansen, J. Fagertun, R. Larsen, "Elastic appearance models," in *Proceedings of British Machine Vision Conference*, 2011, pp 91.1-91.12.

[35] M. Kostinger, P. Wohlhart, P. Roth, H. Bischof, " Annotated facial landmarks in the wild: a largescale, real-world database for facial landmark localization," in *International Conference on Computer Vision WorkShops*, 2011, pp. 2144-2151.

[36] G. Tzimiropoulos, J. Alabort, S. Zafeiriou, M. Pantic, "Generic active appearance models revisited," in *Proceedings of Asian Conference on Computer Vision*, 2012, pp. 650-663.

[37] G. Tzimiropoulos, M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 593-600.

[38] P. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models: their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38-59, Jan. 1995.

[39] D. Cristinacce, T. Cootes, "Facial feature detection using AdaBoost with shape constraints", in *Proceedings of British Machine Vision Conference*, 2003, pp 24.1-24.10.

[40] D. Cristinacce, T. Cootes,"A comparison of shape constrained facial feature detectors," *Proceedings of Inteernational Conference on Automatic Face and Gesture Recognition*, 2004, pp 375-380.

[41] D. Cristinacce, T. Cootes, "Boosted regression active shape models," in *Proceedings of British Machine Vision Conference*, 2007, pp. 1-10.

[42] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," in *Proceedings of European Conference on Computer Vision*, 2008, pp. 504-513.

[43] J. Saragih, S. Lucey, J. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp 1034-1041.

[44] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2011, pp. 545-552.

[45] J. Saragih, S. Lucey, and J. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift," *Int. J. Comput. Vision*, vol. 91, no. 2, pp. 200-215, Jan. 2011.

[46] X. Zhu, and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879-2886.

[47] P. Martins, R. Caseiro, J. Henriques, J. Batista, "Let the shape speak-discriminative face alignment using conjugate priors", in *Proceedings of British Machine Vision Conference*, 2012, pp 118.1-118.11.

[48] I. Dryden and K. Mardia, *Statistical Shape Analysis,*, John Wiley and Sons, 1998.

[49] L. Trujillo, G. Olague, R. Hammoud, and B. Hern, "Automatic feature localization in thermal images for facial expression recognition", in *CVPR'05 Workshops. Joint IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2005, pp. 7-14.

[50] B. Martinez, X. Binefa, and M. Pantic, "Facial Component Detection in Thermal Imagery", in *CVPR'10 Workshops. Joint IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2010, pp. 48-54.

[51] P. Viola, and M. Jones, "Robust Real-time Object Detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137-154, May 2001.

[52] E. Mostafa, R. Hammoud, A. Ali, and A. Farag "Face Recognition in Low Resolution Thermal Images ," in *Computer Vision and Image Understanding Journal*, Volume 117, Issue 12, December 2013, pp. 1689 -1694.

[53] F. Felzenszwalb, and P. Huttenlocher, "Pictorial Structures for Object Recognition," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55-79, Jan 2005.

[54] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

[55] C. Chang, and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[56] K. Mikolajczyk, and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 27, no. 10, pp. 1615-1630, Oct. 2005.

[57] C. Erdem, S. Ulukaya, A. Karaali, and A.Erdem, "Combining Haar Feature and skin color based classifiers for face detection", in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 1497-1500.

[58] H. Rara, A. Ali, S. Elhabian, T. Starr, A. Farag,"Face Recognition at-a-Distance Using Texture, Dense- and Sparse-Stereo Reconstruction", *International Conference on Pattern Recognition*, pp. 121-1224, 2010.

[59] J. Parris, M. Wilber, B. Heflin, H. Rara, A. El-barkouky, A. Farag," Face and eye detection on hard datasets",*International Joint Conference on Biometrics*, pp. 1-10, 2011.

[60] C. Zhang , and Z. Zhang, "A Survey of Recent Advances in Face detection", Microsoft Research, Redmond, WA, Tech. Rep. MSR-TR-2010-66, Jun. 2010

[61] M. Yang ,D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 24, no. 1, pp. 34-58, Jan. 2002.

[62] P. Kakumanu ,S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods", *Pattern Recogn.*, vol. 40, no. 3, pp. 1106-1122, Mar. 2007.

[63] B. Jun , and D. Kim, "Robust Real-Time Face Detection Using Face Certainty Map", in *Proceedings of the international conference on Advances in Biometrics*, 2007, pp. 29-38.

[64] C. Conaire ,N. O'Connor, and A. Smeaton, "Detector adaptation by maximising agreement between independent data sources", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-6.

[65] V. Jain , and E. Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings", University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, Jan. 2010

[66] V. Le, J. Brandt, Z. Lin, V. Jain , and E. Miller, "Interactive Facial Feature Localization", in *Proceedings of European Conference on Computer Vision*, 2012.

[67] V. Jain , and E. Miller, "Online domain adaptation of a pre-trained cascade of classifiers", University of Massachusetts, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 577 -584.

[68] K. Mikolajczyk, F. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proceedings of European Conference on Computer Vision*, 2004, pp. 69-81.

[69] B. Venkatesh, and S. Marcel, "Fast Bounding Box Estimation based Face Detection," in *Proceedings of Workshop on Face Detection of the European Conference on Computer Vision*, 2010, pp. 69-81.

[70] V. Subburaman. S. Marcel,"Fast Bounding Box Estimation based Face Detection", *Workshop on Face Detection: Where we are, and what next*, 2010.

[71] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf, "Face Detection: Efficient and Rank Deficient," in *Proceedings of Advances in Neural Information Processing Systems*, 2005, pp. 673-680.

[72] E. Mostafa ,A. El-Barkouky, H. Rara, and A. Farag, "Rejecting pseudo-faces using the likelihood of facial features and skin," in *Proceedings of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 365-370.

[73] E. Mostafa, and A. Farag, "Dynamic Weighting of Facial Features for Automatic Pose-Invariant Face Recognition," in *Proceedings of IEEE Ninth Conference on Computer and Robot Vision*, 2012, pp. 411-416.

[74] A. Asthana, T. Gedeon,R. Goecke, and C. Sanderson, "Learning-based face synthesis for pose-robust recognition from single image," in *Proceedings of the British Machine Vision Conference*, Dublin, UK , 2009, pp. 31.1-31.10.

[75] A. Asthana, M. Jones, T. Marks, K. Tieu, and R. Goecke, "Pose Normalization via Learned 2D Warping for Fully Automatic Face Recognition," in *Proceedings of the British Machine Vision Conference*, 2011, pp. 127.1-127.11.

[76] J. Atick, P. Griffin, and A. Redlich, "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Comput.*, vol. 8, no. 6, pp. 1321-1340, Jun. 2006.

[77] V. Blanz, and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 25, no. 9, pp. 1063-1074, Sep. 2003.

[78] M. Castelan, and W. Smith, "A Coupled Statistical Model for Face Shape Recovery From Brightness Images," *IEEE Trans. on Image Process.*, vol. 16, no. 4, pp. 1139-1151, Apr. 2007.

[79] C. Castillo, and D. Jacobs, "Using Stereo Matching with General Epipolar Geometry for 2D Face Recognition across Pose," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 31, no. 12, pp. 2298 -2304, Dec. 2009.

[80] X. Chai, and S. Shan, X. Chen, and W. Gao, "A Coupled Statistical Model for Face Shape Recovery From Brightness Images," *IEEE Trans. on Image Process.*, vol. 16, no. 7, pp. 1716 -1725, Jul. 2007.

[81] H. Gao, H. Ekenel, and R. Stiefelhagen, "Pose Normalization for Local Appearance-Based Face Recognition," in *Proceedings of International Conference on Biometrics*, 2009, pp. 32-41.

[82] J. Guillemaut, J. Kittler, and M. Sadeghi, and W. Christmas, "General pose face recognition using frontal face model," in *Proceedings of the Iberoamerican conference on Progress in Pattern Recognition, Image Analysis and Applications*, 2006, pp. 79-88.

[83] T. Kanade, and A. Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *Proceedings of Computational Intelligence in Robotics and Automation*, 2003, pp. 954 - 959.

[84] M. Sarfraz, and O. Hellwich, "Probabilistic learning for fully automatic face recognition across pose," *Image Vision Comput.*, vol. 28, no. 5, pp. 744-753, May 2010.

[85] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model For face representation and recognition," in *Proceedings of International Conference in Computer Vision*, 2005, pp. 786791.

[86] A. Sharma, M. Haj, J. Choi, L. Davis, and D. Jacobs, "Robust pose invariant face recognition using coupled latent space discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 11, pp. 1095-1110, Nov. 2012.

[87] A. Li, S. Shan, and W. Gao, "Coupled bias-variance trade off for cross-pose face recognition", *IEEE Trans. on Image Process.*, vol. 21, no. 1, pp. 305-315, Jan. 2012.

[88] H. Ho, and R. Chellappa, "Pose-invariant face recognition using markov random fields," *IEEE Trans. on Image Process.*, vol. 22, no. 4, pp. 1573-1584, Apr. 2013.

[89] L. Ding, X. Ding, and C. Fang, "Continuous pose normalization for pose-robust face recognition," *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 721-724, Nov. 2012.

[90] W. Schwartz, H. Guo, and L. Davis, "A robust and scalable approach to face identification" , in *Proceedings of European Conference on Computer Vision*, 2010, pp. 476-513.

[91] A. Serrano, I. Diego, C. Conde, and E. Cabello,, "Recent advances in face biometrics with gabor wavelets: A review," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 372-381, May 2010.

[92] S. Xie, S. Shan, X. Chen, and J. Chen,"Fusing local patterns of gabor magnitude and phase for face recognition," *IEEE Trans. on Image Process.*, vol. 19, no. 5, pp. 1349-1361, May 2010.

[93] H. Nguyen, L. Bai, and L. Shen, "Local gabor binary pattern whitened pca: A novel approach for face recognition from single image per person," in *Proceedings of International Conference on Biometrics*, 2009, pp. 269-278.

[94] B. Zhang, S, Shan, X. Chen, and W. Gao, "Histogram of gabor phase patterns (HGPP): a novel object representation approach for face recognition, *IEEE Trans. on Image Process.*, vol. 16, no. 1, pp. 57-68, Jan. 2007.

[95] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," in *Proceedings of the International conference on Image Processing*, 1997, pp. 129-132.

[96] V. Kolmogorov. Graph Based Algorithms for Scene Reconstruction from Two or More Views. PhD thesis, Cornell University, Ithaca, NY, 2004.

[97] S. Birchfield and C. Tomasi. "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 20, no. 4, pp. 401406, Apr. 1998.

[98] G. Finlayson and R. Xu, "Illuminant and gamma comprehensive nor-malisation in log rgb space," *Pattern Recognition Letters*, vol. 24, no. 11, pp. 16791690, Nov. 2003.

[99] Y. Heo, K. Lee, and S. Lee, "Illumination and camera invariant stereo matching", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[100] S. Gupta, J. Aggarwal, M. Markey, and A. Bovik., "3D face recognition founded on the structural diversity of human faces", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[101] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[102] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[103] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 25, no. 12, pp. 1615-1618, Oct. 2003.

[104] S. Sarkar, "USF DARPA Human-ID 3D FaceDatabase," *http://www.cse.usf.edu/ sarkar/*, University of South Florida, Tampa, FL.

[105] R. Gross ,I. Matthews, J. Cohn, T. Kanade, and S. Baker "Multi-PIE," in Proceedings *IEEE Conference on Automatic Face Gesture Recognition*, pp. 1-8, 2008.

[106] H. Rara, A. Farag, and T. Davis, "Model-based 3D shape recovery from single images of unknown pose and illumination using a small number of feature points," in *Proceedings of International Joint Conference on Biometrics*, 2011, pp. 1-7.

[107] Shirren Elhabian, Eslam Mostafa, Ham Rara, and Aly Farag, "Non-Lambertian Model-based Facial Shape Recovery from Single Image Under Unknown General Illumination," in *Proceedings of Ninth Conference on Computer and Robot Vision*, 2012, pp. 252-259.

[108] O. Aldrian, and W. Smith, "A Linear Approach of 3D Face Shape and Texture Recovery using a 3D Morphable Model," in *Proceedings of the British Machine Vision Conference*, 2010, pp. 75.1-75.10.

[109] W. Smith, and E. Hancock, "Recovering Facial Shape Using a Statistical Model of Surface Normal Direction," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 28, no. 12, pp. 1914-1930, Oct. 2006.

[110] X. Zhang, Y. Gao and M. Leung, "Automatic Texture Synthesis for Face Recognition from Single Views," in *Proceedings of International Conference on Pattern Recognition*, 2006, pp. 1151-1154.

[111] X. Liu, "Pose-Robust Face Recognition Using Geometry Assisted Probabilistic Modeling," in *Proceedings of Computer Vision and Pattern Recognition Conference*, 2005, pp. 502-509.

[112] U. Prabhu, H. Jingu, and M. Savvides, "Unconstrained Pose-Invariant Face Recognition Using 3D Generic Elastic Models," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 33, no. 10, pp. 1952-1961, Oct. 2011.

[113] S. Prince ,P. Li, Y. Fu, U. Mohammed, and J. Elder, "Probabilistic models for inference about identity," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 34, no. 1, pp. 144-157, Jan. 2012.

[114] S. Prince ,J. Elder, J. Warrell, and F. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 30, no. 6, pp. 970-984, Jun. 2008.

[115] Eslam Mostafa, Asem Ali, Naif Alajlan, and Aly Farag, "Pose invariant approach for face recognition at distance," in *Proceedings of European Conference on Computer Vision - Volume Part VI,* 2012, pp. 15-28.

[116] Eslam Mostafa, Moumen Elmelegy, and Aly Farag, "Passive Single Image-based Approach for Camera Steering in Face Recognition at-a-Distance Applications," in *Proceedings of IEEE Conference on Biometrics: Theory, Applications and Systems,* 2012, pp. 371-376.

[117] M. Greiffenhagen, M. Enhagen, V. Ramesh, D. Comaniciu, and H. Niemann, "A Statistical Modeling and Performance Characterization of a Real-Time Dual Camera Surveillance System," 2000.

[118] S. Stillman, R. Tanawongsuwan, and I. Essa, "A System for Tracking and Recognizing Multiple People with Multiple Cameras," in *Proceedings of IEEE Conference on Audio-Vision based Person Authentication,* 1998, pp. 96-101.

[119] F. Wheeler, R. Weiss, and P. Tu, "Face recognition at a distance system for surveillance applications," in *Proceedings of IEEE Conference on Biometrics: Theory, Applications and Systems,* 2010.

[120] S. Prince, and J. Warrell, J. Elder, and F. Felisberti, "Tied Factor Analysis for Face Recognition across Large Pose Differences," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 30, no. 6, pp. 970 -984, Jun. 2008.

[121] D. Rother, K. Patwardhan, I. Aganj, and G. Sapiro, "3D Priors For Scene Learning From A Single View," in *Proceedings of Computer Vision and Pattern Recognition Conference Workshops*, 2008.

[122] L. Marchesotti, L. Marcenaro, and C. Regazzoni, "Dual camera system for face detection in unconstrained environments," in *Proceedings of IEEE International Conference on Image Processing,* 2003, pp. 681-684.

[123] J. Elder, S. Prince, and Y. Hou, and M. Sizintsev, "Pre-attentive and attentive detection of humans in wide-field scenes," *International Journal of Computer Vision*, vol. 72, no. 1, pp. 47-66, Sep. 2007.

[124] J. Elder, S. Prince, and Y. Hou, and M. Sizintsev, "Pre-attentive and attentive detection of humans in wide-field scenes," *International Journal of Computer Vision*, vol. 72, no. 1, pp. 47-66, Sep. 2007.

[125] N. Dodgson, "Variation and extrema of human interpupillary distance," *Proceedings of SPIE: Stereoscopic Displays and Virtual Reality Systems XI,*, vol. 5291, pp. 3646, 2004.

[126] N. Dodgson, "Variation and extrema of human interpupillary distance," *Proceedings of SPIE: Stereoscopic Displays and Virtual Reality Systems XI,*, vol. 5291, pp. 3646, 2004.

[127] X. Zhou, R. Collins, T. Kanade, and P. Metes, "A Master-Slave System to Acquire Biometric Imagery of Humans at Distance," *ACM International Workshop on Video Surveillance,*, Nov. 2003.

[128] A. Hampapur, S. Pankanti, and A. Senior, Y. Tian, L. Brown, and R. Bolle "Face cataloger: multi-scale imaging for relating identity to location," in *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance,* 2003, pp. 13-20.

[129] E. Mostafa, A. Ali, and A. Farag, "Learning A NonLinear Combination of Mahalanobis Distances Using Statistical Inference For Similarity Measure," in *Institution of Engineering and Technology Journal on Computer Vision*, Accepted to appear 2015.

[130] L. Wolf, T. Hassner, Y. Taigman, "Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10,pp. 1978-1990, 2011.

[131] X. He, D. Cai, S. Yan, H. Zhang, "Neighborhood preserving embedding", *IEEE International Conference on Computer Vision*, pp. 1208-1213, 2005.

[132] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *SCIENCE*, vol. 290, pp. 2323-2326, 2000.

[133] H. Chang, D. Yeung,"Kernel-based distance metric learning for content-based image retrieval",emphImage Vision Computing Journal", vol. 25, no. 5, pp. 695-703, 2007.

[134] M. Dikmen, E. Akbas, T. Huang, N. Ahuja,"Pedestrian recognition with a learned metric", *Asian Conference on Computer Vision (ACCV)*, 2011.

[135] M. Guillaumin, J. Verbeek, C. Schmid,"Multiple instance metric learning from automatically labeled bags of faces", *European Conference on Computer Vision (ECCV)*, 2010.

[136] J. Ye, Z. Zhao, H. Liu, "Adaptive Distance Metric Learning for Clustering",*IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007

[137] K. Weinberger, L. Saul,"Fast solvers and efficient implementations for distance metric learning",*International Conference on Machine Learning (ICML)*, 2008.

[138] J. Davis, B. Kulis, P. Jain, S. Sra, I. Dhillon, "Information-theoretic metric learning", *International Conference on Machine Learning (ICML)*, 2007.

[139] K. Weinberger, L. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification", *Journal Machine Learning Resarch*, 2009.

[140] M. Guillaumin, J. Verbeek, C. Schmid, "Is that you? Metric learning approaches for face identification",*IEEE International Conference on Computer Vision*, 2009.

[141] Y. Ying, P. Li,"Distance Metric Learning with Eigenvalue Optimization",*Journal of Machine Learning Resarch*, vol. 13, pp. 1-26, 2012.

[142] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, H. Bischof, "Large scale metric learning from equivalence constraints",*IEEE conference on Computer Vision and Pattern Recognition*, 2012.

[143] M. Hirzer, P. Roth , M. Kstinger , H. Bischof,"Relaxed Pairwise Learned Metric for Person Re-identification",*European Conference on Computer Vision*, 2012.

[144] W. Zheng, S. Gong, T. Xiang,"Person re-identification by probabilistic relative distance comparison",*IEEE conference on Computer Vision and Pattern Recognition*, 2011.

[145] Y. Sun, X. Wang, X. Tang, "Deep learning face representation from predicting 10,000 classes", *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[146] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification", *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[147] T. Berg and P. Belhumeur, "POOF: Part-Based One-vs-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation", *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[148] J. Lu, X. Zhou, Y. Tan, Y. Shang, J. Zhou,"Neighborhood Repulsed Metric Learning for Kinship Verification",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 331-345, 2014.

[149] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, "Attribute and Simile Classifiers for Face Verification",*IEEE International Conference on Computer Vision (ICCV)* , 2009.

[150] G. Huang, M. Ramesh, T. Berg, E. Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments", *University of Massachusetts, Amherst*, 2007.

[151] H. Nguyen, L. Bai, "Cosine Similarity Metric Learning for Face Verification",*10th Asian Conference on Computer Vision (ACCV)*, 2010.

[152] D. Cox, N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition", *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, 2011.

[153] J. Lu, G. Wang, P. Moulin, "Image Set Classification Using Holistic Multiple Order Statistics Features and Localized Multi-kernel Metric Learning", *IEEE International Conference on Computer Vision (ICCV)*, pp.329-336, 2013.

[154] Q. Cao, Y. Ying, P. Li,"Similarity Metric Learning for Face Recognition",*IEEE International Conference on Computer Vision (ICCV)*, 2013.

[155] L. Wolf, T. Hassner, Y. Taigman, "Descriptor based methods in the wild",*In: Faces in Real-Life Images Workshop in ECCV*, 2008.

[156] D. Lowe,"Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, November 2004

[157] P. Velsor, "School counselors as social-emotional learning consultants: Where do we begin?" *Professional School Counseling*, vol. 13, no. 1, pp. 5058, 2009.

[158] B. Kort, R. Reilly, R. Picard "An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion," In Proc. *IEEE International Conference on Advanced Learning Technologies*, pp. 43-46, 2001.

[159] P. Ekman, "Facial Expressions of Emotion: An Old Controversy and New Findings," *Philosophical Transactions of the Royal Society of London*, vol. B, no. 335, pp. 63-69, 1992.

[160] P. Rani, C. Liu, N. Sarkar,E. Vanman, "An empirical study of machine learning techniques for affect recognition in humanrobot interaction", *Pattern Analysis and Applications*, vol. 9, pp. 5869, 2006.

[161] Y. Zhou, P. Tsiamyrtzis, I. Pavlidis, "Tissue Tracking in Thermo-physiological Imagery through Spatio-temporal Smoothing", in Proc. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2009.

[162] A. Levin, D. Lischinski, Y. Weiss, "A Closed-Form Solution to Natural Image Matting", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228-242, 2008.

[163] A. Jepson, D. Fleet, T. El-Maraghi, "Robust online appearance models for visual tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1296-1311, 2003.

[164] J. Dowdall, I. Pavlidis,P. Tsiamyrtzis, "Coalitional tracking", *Computer Vision Image Understing Journal*, vol. 106, no.2 pp. 205-219, 2007.

[165] D. Thang, N. Vo, G. Medioni,"Context tracker: Exploring supporters and distracters in unconstrained environments", in proc *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1177-1184, 2011.

[166] H. Grabner, J. Matas, L. Van Gool, P. Cattin, "Tracking the invisible: Learning where the object might be", in proc *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1285-1292, 2010.

[167] Z. Kalal, K. Mikolajczyk, J. Matas,"Tracking-Learning-Detection",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409-1422, 2012.

[168] Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis,"Spatio-Temporal Smoothing as a Basis for Facial Tissue Tracking in Thermal Imaging", *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1280-1289, 2013.

[169] Z. Kalal, J. Matas, K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 49-56, 2010.

[170] A. Bosch, A. Zisserman, X. Muoz, "Image Classification using Random Forests and Ferns",*International Conference on Computer Vision*, pp. 1-8, 2007.

[171] M. Ozuysal, M. Calonder, V. Lepetit, P. Fua,"Fast Keypoint Recognition Using Random Ferns",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448-461, 2010.

[172] H. Grabner, H. Bischof,"On-line Boosting and Vision",*IEEE Conference on Computer Vision and Pattern Recognition*, pp. 260-267, 2006.

[173] S. Stalder, H. Grabner, L. Van Gool,"Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition",*International Conference on Computer Vision Workshops*, pp. 1409-1416, 2009

[174] N. Sun,M. Garbey, A. Merla, I. Pavlidis,"Imaging the cardiovascular pulse", *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 416-421, 2005.

[175] R. Murthy, I. Pavlidis,"Noncontact measurement of breathing function",*IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 3, pp. 57-67, 2006.

[176] S. Chekmenev,A. Farag, E. Essock,"Thermal Imaging of the Superficial Temporal Artery: An Arterial Pulse Recovery Model",*IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-6, 2007.

[177] C. Liu, K. Conn, N. Sarkar, W. Stone,"Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder", *International Journal Human-Computer Studies*, vol. 66, no. 9, pp. 662-677, 2008.

[178] K. Welch, U. Lahiri, C. Liu, R. Weller, N. Sarkar,Z. Warren,"An Affect-Sensitive Social Interaction Paradigm Utilizing Virtual Reality Environments for Autism Intervention", in proc*International Conference on Human-Computer Interaction. Part III: Ubiquitous and Intelligent Interaction*, pp. 703-712, 2009.

[179] I. Michael, A. Blake,"CONDENSATION - conditional density propagation for visual tracking",*International Journal of Computer Vision*, vol. 29, pp. 5-28, 1998.

[180] B. Babenko, Y. Ming-Hsuan, S. Belongie, "Visual tracking with online Multiple Instance Learning",*IEEE Conference on Computer Vision and Pattern Recognition*, pp. 983-990, 2009.

[181] X. Mei, H. Ling, "Robust visual tracking using 1 minimization",*International Conference on Computer Vision*, pp. 1436-1443, 2009

[182] L. Breiman,"Random Forests",*Machine Learning*, pp. 5-32, 2001.

[183] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ”ORB: An efficient alternative to SIFT or SURF”,*IEEE International Conference on Computer Vision*, pp. 2564-2571, 2011.

[184] M. Calonder, V. Lepetit, C. Strecha, P. Fua, ”BRIEF: binary robust independent elementary features”, in Proc. *European conference on Computer vision: Part IV*, pp. 778-792, 2010.

CURRICULUM VITAE

Name: Eslam Mostafa

Contact Information:

Computer Vision and Image Processing Laboratory

2211 South Brook

Lutz Hall, Room #6, Louisville, KY 40292, USA

E-mail: eslam.a.mostafa@gmail.com

Mobile: +1-954-8306428

Objective Pursuit of advanced research and development in computer vision, image processing, and medical imaging as a postdoctoral fellow.

Research Interests Highly competent researcher in computer vision, machine learning, biometrics, and image processing, areas with 21 peer reviewed published papers. Experienced in C/C++, Matlab, Python, Ruby, and SQL on Linux/Mac OS/Windows environments Strong interpersonal skills in collaboration with academic and industry professionals.

Career Profile

- Have the will and the desire to learn.
- Like to be responsible.
- Capable to work for long continuous period under pressure, in time sensitive and fast paced environment.
- Performed successfully in a team setting as well as individually.

Education

University of Louisville,Louisville, Kentucky, USA

- Thesis Topic: *FACE MODELING FOR FACE RECOGNITION IN THE WILD*
- GPA : 4.0 out of 4.0
- Advisor: Professor Aly A. Farag
- Thesis Committee:
  - Professor Aly A. Farag, ECE Department and Director of CVIP-Lab@UofL
  - Professor John F. Nabe, ECE Department, University of Louisville
  - Professor Sahoo, Prasanna K.,A&S Mathematics, University of Louisville
  - Professor Robert W. Cohn, ECE Department
  - Asstistant Professor Roman V. Yampolskiy, CECS Department, University of Louisville
- Area of Study: Computer Vision and Image Processing

Alkexandria University,Alexandria, Egypt

M.Sc., Faculty of Engineering, 2009

- Thesis Topic: *ITraffic Engineering in Wireless OFDM Cognitive Netwrok*
- Advisors: Professor Said Elnoubi
- Area of Study: Wireless Communications

B.Sc., Faculty of Engineering, 2006

- GPA: 4.0
- Implementation of GSM-UMTS using Software Defined Radio on FPGA Completed in the top 1

Academic Appointments Graduate Research Assistant August 2009 to present

Department of Electrical and Computer Engineering,

University of Louisville

Research and Development:

- Key member of Biometric Optical Surveillance System funded by the Department of Homeland Security with budget 5.3 million for establishing a system for face recognition at 100 meter range, appeared in New York Times news, Aug.13. - Develop a new similarity measure for handling illumination, expression, and aging challenges in face recognition systems, published IET 15. - Developed a face detection algorithm based on likelihood of boosted classifier and facial features, published in BTAS 12. - Devised an algorithm for multi camera network management, published BTAS 11. - Achieved a pose invariant approach for face recognition based on stereo Imaging-3D reconstruction /Rendering, published ECCV 12.

Graduate Teaching Assistant June 2002 to July 2007

Faculty of Engineering,

Alexandria University

Activities:

- Lecturing classes and labs.
- Helping students throughout their study (office hours)
- Helping students with their thesis projects
- Grading students' assignments
- Grading midterm exams

Refereed Journal Publications

- Eslam Mostafa, Asem Ali, and Aly Farag, Learning A NonLinear Combination of Mahalanobis Distances Using Statistical Inference For Similarity Measure, Institution of Engineering and Technology Journal on Computer Vision, Accepted to appear 2015.

- Eslam Mostafa, Riad Hammoud, Asem Ali, and Aly Farag, Face Recognition in Low Resolution Thermal Images, Computer Vision and Image Understanding Jo urnal, Volume 117, Issue 12, December 2013, pp. 1689 -1694.

- Eslam Mostafa, and Aly Farag, Weighting of facial features for pose Invariant face recognition, Under Review in IEEE Transaction of Image Processing.

Conference Publications

- Eslam Mostafa, Aly Farag, Shireen Elhabian, Aly Abdelrahim, and Salwa Elshazly, Statistical Morphable Model for Human Teeth Restoration, in Proc. of IEEE International Conference on Image Processing, October, 2014.

- Eslam Mostafa, Travis Gault, Asem Ali, and Aly Farag, Long term Facial Parts Tracking in thermal Imaging for Uncooperative Emotion Recognition, in Proc. of IEEE Conference on Biometrics: Theory, A pplications and Systems, 2013.

- Ahmed Shalaby, Eslam Mostafa, Aly A. Farag, and Todd Hockenbury, 2D-3D Registration: A Step towards Image-Guided Ankle Fusion, , MWBIVPCS 2013 - MICCAI workshop on Bio- Imaging and Visualization for Patient-Customized Simulations.

- Eslam Mostafa, Asem Ali, Naif Alajlan and Aly Farag, Pose invariant approach for face recognition at distance, in Proc. of European Conference on Computer Vision-Volume Part VI, 2012, pp. 15-28.

- Eslam Mostafa, and Aly Farag, Complex Bingham Distribution for Facial Feature Detection, in Proc. of European Conference on Computer Vision Workshops, 2012, pp. 330-339.

- Eslam Mostafa, Moumen Elmelegy, and Aly Farag, Passive Single Image-based Approach for Camera Steering in Face Recognition at-a-Distance Applications, in Proc. IEEE Conference on Biometrics: Theory and Systems, 2012, pp. 371-376.

- Eslam Mostafa , Ahmed El-Barkouky, Ham Rara, and A. Farag, Rejecting pseudo-faces using the likelihood of facial features and skin, in Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems, 2012, pp. 365-370.

- Aly Adelrahim, Aly Farag, Shireen Elhabian, Eslam Mostafa and Wael Aboelmaaty, Occlusal Surface Reconstruction of Human Teeth From A Single Image Based On Object and Sensor Physical Characteristics, 19th IEEE International Conference on Image Processing (ICIP), 2012.

- Eslam Mostafa, and Aly Farag, Dynamic Weighting of Facial Features for Automatic Pose-Invariant Face Recognition, in Proc. of Ninth Conference on Computer and Robot Vision, 2012, pp. 411-416.

- Shirren Elhabian, Eslam Mostafa, Ham Rara, and Aly Farag, Non-Lambertian Model-based Facial Shape Recovery from Single Image Under Un-known General Illumination, in Proc. of Ninth Conference on Computer and Robot Vision, 2012.

- Melih Aslan, Eslam Mostafa, Hossam Abdelmunim, Ahmed Shalaby, Aly Farag, and Ben Arnold, A novel probabilistic simultaneous segmentation and registration using level set, Proc. of IEEE International Conference on Image Processing (ICIP), pp. 2161 - 2164, Sept 2011.

- Travis Gault, Eslam Mostafa, Ahmed Farag and Aly Farag, Less is more: Cropping to Improve Facial Recognition with Thermal Images, Proc. of International Conference on Multimedia Technology (ICMT), 2011.

Honors and Awards

- Doctoral Dissertation Compilation Award, SIGS

- Outstanding Graduate Student University of Louisville, Dean of Students, 2013
- Outstanding Graduate Student Computer Engineering Department, 2013
- Theobald Award, Computer Engineering Department, 2013
- Honorable Mention in Graduate Reserach Symposium, Oral presentation, 2013
- Best Student Organizer in Computer Vision and Pattern Recognition Conference, 2012
- Travel grant Award from European Conference on Computer Vision, 2012
- Faculty of Engineering Certificate of Honor, Alexandria University, 2002, 2003, 2004

Skills Programming Languages and Toolkits

- C and C++ Fair and In Progress
- Matlab Highly Competent
- C# Fair and In Progress
- Java Basic
- Visulization Toolkit (VTK) Fair and In Progress

Software

Microsoft Office	Highly Competent
Latex	Highly Competent
Microsoft Visual Studio	good
Mathematica	Basic

Areas of Research Focus Mathematics:

- Real and complex analysis, differential geometry, linear algebra, multilinear algebra, and spherical harmonics

Signal Processing:

- Signal and image processing, probability, random variables, stochastic processes, information theory, subspace learning, pattern recognition and machine learning

Computer Vision:

- Shape-from-shading, statistical shape-from-shading, stereo reconstruction and image formation

Medical Imaging:

- Statistical shape analysis, image segmentation and image processing

Relevant Coursework:

- Electronic circuits,
- Control System Principles,
- Autonomous Robots,
- Communications and modulation,
- Measurements,
- Digital Signal Processing,
- Computer Vision,
- Image Processing,
- Database,
- Data structure,
- Computer graphics,

- Networks,
- Digital Design,
- Computer Architecture
- Microprocessor.

Languages

- Arabic — Mother Tongue
- English — Very Good
- French — Basic

Hobbies: sports, fishing, collecting stamps, electricity, mechanics and carpentry