FACE RECOGNITION FOR VEHICLE PERSONALIZATION

A Thesis
Presented to
The Academic Faculty

by

Jinwoo Kang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2016

FACE RECOGNITION FOR VEHICLE PERSONALIZATION

Approved by:

Professor David V. Anderson, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Biing-Hwang Juang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Monson H. Hayes III
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Edward J. Coyle
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mark T. Smith
School of Information and Communication Technology
Swedish Royal Institute of Technology (KTH)

Professor Aaron D. Lanterman
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: August 2016

To my mother and father,

for their

boundless love and patience.

ACKNOWLEDGEMENTS

First and foremost, I thank my parents for their unconditional love and support, without which none of the worthwhile achievements in my life would have been possible. I also give my deepest thanks to my advisors, Prof. Monson Hayes and Prof. David Anderson, for their excellent research guidance and training and for teaching me the meaning of hard work and dedication by show of example. They are true researchers and advisors. I thank the many other faculty members who interacted with me in varying degrees and helped me in the process of developing as a graduate student, including Prof. Mark Smith and Prof. Fred Juang. The friendship and support of peers was also an indispensable part of my tenure as a graduate student. At the risk of inadvertently omitting some names, I thank Michael Farrell, Yeongseon Lee, Woojay Jeon, Joon Hyun Sung, Smita Vemulapalli, Emily Xu, Muhammad Rizwan, Amol Borkar, Angelique Yeung, Nancy Nong, Jaegul Choo, Byungki Byun, Michael Lee, Nathan Parrish, Bradley Whitaker, Kaitlin Fair, Femi Odelowo, Brandon Carroll, Salman Aslam, Dongshin Kim, Sangmin Oh, Kihwan Kim and all the people participated in the vehicular face recognition database acquisition. I also express my gratitude to the Ministry of Information and Communication of Korea for its generous financial support that funded part of my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The objective of this dissertation is to develop a system of practical technologies to implement an illumination robust, consumer grade biometric system based on face recognition to be used in the automotive market. Most current face recognition systems are compromised in accuracy by ambient illumination changes. Especially outdoor applications including vehicle personalization pose the most challenging environment for face recognition. The point of this research is to investigate practical face recognition used for identity management in order to minimize algorithmic complexity while making the system robust to ambient illumination changes. We start this dissertation by proposing an end-to-end face recognition system using near infrared (NIR) spectrum. The advantage of NIR over visible light is that it is invisible to the human eyes while most CCD and CMOS imaging devices show reasonable response to NIR. Therefore, we can build an unobtrusive night-time vision system with active NIR illumination. In day time the active NIR illumination provides more controlled illumination condition. Next, we propose an end-to-end system with active NIR image differencing which takes the difference between successive image frames, one illuminated and one not illuminated, to make the system more robust on illumination changes. Furthermore, we addresses several aspects of the problem in active NIR image differencing which are motion artifact and noise in the difference frame, namely how to efficiently and more accurately align the illuminated frame and ambient frame, and how to combine information in the difference frame and the illuminated frame. Finally, we conclude the dissertation by citing the contributions of the research and discussing the avenues for future work.

CHAPTER I

INTRODUCTION

1.1 Face Recognition as a Consumer Grade Biometric

Biometrics as an identity management discipline seeks to either identify a person or verify a person's claimed identity. In most such systems a design goal is to minimize the probability of false admission even at the expense of increasing the probability of false rejection [82]. The economics behind this are driven by applications that must manage a significant economic, safety, or security threat. Any such system that incorrectly admits even a vanishingly small number of subjects will fail in the marketplace. However, the market for ID management extends well beyond high-security applications. Proponents of pervasive and context-aware computing have argued that a knowledge of the user's identity can greatly enhance the perceived value of an application by personalizing it [122]. The value increase comes from a combination of usability enhancements such as automatic configuration or customization, and through the perception that an otherwise shared resource is virtually one's own and reflects favorable attributes tied to one's identity, such as opinion, emotion, or fashion. There are countless such applications that are enabled by knowing the user's identity, but carry no significant risk of personal, physical, or economic harm should the user's identity be incorrectly determined. Unlike high-security applications where it is preferable to reject the identity of a person under any reasonable doubt, in no- or low-security consumer applications, it is preferable to always attempt to converge on an identity of the subject, even if it is wrong. Consumer-grade biometrics are therefore characterized by minimizing the probability of a false rejection at the expense of increasing the probability of false admission. This reflects the difference in priorities

between consumer-grade and high-security applications. Usability is an issue in all identity management methods and in turn will contribute to the overall ease of use of an application in which it is used. Ease of use in consumer applications is very desirable, so consumer grade biometrics ideally will involve a simple training process. The actual use of the system should be invisible. Operation should be automatic and require no direct user cooperation.

In this dissertation, face recognition is used because it is a well-studied biometric [129] using both static- and video-based image collection, and also has the attributes of being noninvasive and potentially requiring no explicit cooperation from the user for the biometric to work. With respect to performance, an interesting difference in this work compared with previous research on face recognition is the idea of consumer-grade biometrics and how it can alter the goals of biometric device design.

1.2 Vehicle Personalization

The application area selected as a target for this dissertation is biometrics for the automotive market. This is an attractive application space due to a growing number of opportunities and needs that can be addressed. Traditional application ideas using personalization involve automatically adjusting physical properties of the car, such as the positions of seats and mirrors. New regulatory requirements also play a role, such as the need to configure and manage hands-free telecommunications for car drivers [3]. The amount and diversity of information technology being introduced into new cars in the form of entertainment, email, navigation, telemetry, and driver assistance services are increasing, all of which can be personalized in some way. Beyond personalization, many expected features of new cars will exploit imaging or video for some other purpose. Examples include parental supervision, driver distraction monitoring, and autonomous operation of vehicles. This makes using an image-based biometric in cars more attractive, as the imaging hardware used can be shared across

many applications. Currently, image-based biometrics are not commonly found in commercially available cars, but the market may grow if practical solutions can be found. Precedent for this is being set by several automotive manufacturers who are developing sensor-based systems for car security, advanced driver support, and other tasks.

Several enabling technologies already exist for driver identification, one of which is embedding special identification devices such as RFID tags in a car key or a key fob. Many car manufacturers are developing RFID-based smart car keys, which enable automatic keyless entry based on proximity sensing or alerting the presence of an intruder in the vehicle cabin [17, 55]. Vehicle personalization is another application of the RFID smart car key. These systems are not looked at in this dissertation as they imply an ownership and involve the inconvenience of having something that must be carried and potentially lost. Password-based identity management methods are also not used in this study as they require user cooperation for the system to work.

1.3 Face Recognition using Near-Infrared Illumination

As face recognition is the biometric chosen for the vehicular personalization application, the proposed algorithm needs to be robust under operation during both day and night. In this regard, the choice and tailoring of the algorithm depend on the mode of video acquisition. With non-intrusiveness being one of the main features of this application, the selection of sensors and illuminators is crucial. The initial choice was to use a color camera as the mode of acquisition. As this camera would work near flawlessly during the day, it turns out to be almost useless at night without a reliable illumination source. With sunlight not available at night, an artificial light illuminator would be needed to aid the camera. Depending on the specifications of the algorithm, this illuminator would be continuously on or intermittently pulse, which

can be troublesome and in turn annoy the driver in both cases. As a result, this option was eliminated and suggestions to explore the use of Infrared (IR) illumination were taken into consideration. The IR spectrum is further subcategorized depending on its wavelength; however, for the purpose of this work, only near-infrared (NIR, 0.7 - 1.4μm) is being considered [59]. The key advantage of NIR over visible light is that it is invisible to the human eye. Compared to the previously described scenario, NIR is easily available from the sun during the day since it is an abundant source. At night however, an NIR illuminator can be used to provide the controlled artificial illumination without bothering the driver. This feature lays the foundation for the end product to be non-intrusive.

Although NIR is invisible to the human eye, most CCD and CMOS imaging devices show reasonable responses to these wavelengths, making them applicable to this work. Color cameras are typically equipped with an optical "IR-cut" filter in front of the CCD or CMOS sensors to control the red color photons contributed by the IR spectrum [53]. As a result, monochromatic cameras are preferred as they do not include this filter. An NIR camera is easily implemented using a monochromatic camera along with a NIR wavelength-passing optical filter. The choice of CMOS- or CCD-based cameras varies depending on application requirements. The artificial illumination is also easily implemented using a network of NIR light emitting diodes (LEDs) which are available from most electronic component vendors.

1.4 Organization of the Dissertation

This dissertation focuses on the development of a novel illumination-robust face recognition system for vehicle personalization with the active NIR illumination and camera. The dissertation is organized as follows:

Chapter 2 introduces the background literature review on previous illumination-robust approaches for face recognition including passive and active methods.

Chapter 3 presents a consumer grade biometric system based on face recognition using infrared imaging with successful recognition result in a small group of subjects. The system consists of three stages; face detection, eye detection and face recognition.

Chapter 4 focuses on the development of face recognition system with NIR active image differencing. The NIR active image differencing produces images independent of the ambient illumination. End-to-end face recognition system is presented including foreground/background segmentation, motion detection, face detection, pose clustering and face recognition modules. It is shown that the image differencing method makes the modules more robust to the ambient illumination variation. Additionally, we present face video acquisition hardware which implements the NIR active image differencing in real vehicular environment and large face video dataset taken with the hardware. Finally, extensive test results on the dataset are provided to evaluate the end-to-end system.

Chapter 5 addresses several aspects of the problem in active NIR image differencing which are motion artifact and noise in the difference frame, namely how to efficiently and more accurately align the illuminated frame and ambient frame, and how to combine information in the difference frame and the illuminated frame. Extensive experimental results on video dataset introduced in Chapter 4 show performance increase using the proposed methods.

Chapter 6 presents the summary and conclusion reported in this dissertation. In addition, avenues to continue and extend the presented research are also discussed.

CHAPTER II

BACKGROUND

For many applications, the performance of face recognition systems in controlled environments has now reached a satisfactory level; however, there are still many challenges posed by uncontrolled environments. Some of these challenges are posed by the problems caused by variations in illumination, face pose, expression, and etc. The effect of variation in the illumination conditions in particular, which causes dramatic changes in the face appearance, is one of those challenging problems [129] that a practical face recognition system needs to face. To be more specific, the varying direction and energy distribution of the ambient illumination, together with the 3D structure of the human face, can lead to major differences in the shading and shadows on the face. Such variations in the face appearance can be much larger than the variation caused by personal identity [4]. The variations of both global face appearance and local facial features also cause problems for automatic face detection/localisation, which is the prerequisite for the subsequent face recognition stage. Therefore, the situation is even worse for a fully automatic face recognition system. Moreover, in a practical application environment, the illumination variation is always coupled with other problems such as pose variation and expression variation, which increase the complexity of the automatic face recognition problem.

A number of illumination invariant face recognition approaches have been proposed in the past years. Existing approaches addressing the illumination variation problem fall into two main categories. The approaches in the first category is called "passive" approaches, since they attempt to overcome this problem by studying the visible spectrum images in which face appearance has been altered by illumination

variations. The other category contains "active" approaches, in which the illumination variation problem is overcome by employing active imaging techniques to obtain face images captured in consistent illumination condition, or images of illumination invariant modalities. Existing reviews related to illumination invariant face recognition can be found in [69, 73, 129].

2.1 Passive Approaches

Passive approaches can be divided into four groups: illumination variation modelling, illumination invariant features, photometric normalisation, and 3D morphable model.

2.1.1 Illumination Variation Modelling

The modelling of face images under varying illumination can be based on a statistical model or physical model. For statistical modelling, no assumption concerning the surface property is needed. Statistical analysis techniques, such as PCA(*Eigenface*) and LDA(*Fisherface*), are applied to the training set which contains faces under different illuminations to achieve a subspace which covers the variation of possible illumination. In physical modelling, the model of the process of image formation is based on the assumption of certain object surface reflectance properties, such as Lambertian reflectance.

2.1.1.1 Linear Subspaces

Hallinan [46] showed that five eigenfaces were sufficient to represent the face images under a wide range of lighting condition. Shashua proposed *photometric alignment* approach to find the algebraic connection between all images of an object taken under varying illumination conditions [96]. An order k linear reflectance model for any surface point p is defined as the scalar product $x \cdot a$, where x is a vector in the k-dimensional Euclidean space of invariant surface properties (such as surface normal, albedo, and so forth), and a is an arbitrary vector. The image intensity $I(p)$ of an

object with an order k reflection model can be represented by a linear combination of a set of k images of the object. For Lambertian surface under distant point sources and in the absence of shadows, all the images lie in a 3D linear subspace of the high dimensional image space, which means that they can be represented by a set of 3 images, each from a linearly independent source. Given three images of this surface under three known and linearly independent light sources, the surface normal and the albedo can be recovered. This is known as *photometric stereo*. Shashua claimed the *attached shadows*, which are caused by points where the angle between surface normal and the direction of light source is obtuse($n_p \cdot s < 0$, therefore $I(p) = 0$), do not have a significant adverse effect on the photometric alignment scheme. However, the *cast shadows* caused by occlusion cannot be modeled using the above framework.

Belhumeur *et al.* [13] presented the so-called *3D linear subspace* method for illumination invariant face recognition, which is a variant of the photometric alignment method. In this linear subspace method, three or more images of the same face taken under different lighting are used to construct a 3D basis for the linear subspace. The recognition proceeds by comparing the distance between the test image and each linear subspace of the faces belonging to each identity. The Fisher Linear Discriminant(also called *FisherFace*) method is also proposed in [13] in order to maximise the ratio of the between-class scatter and the within-class scatter of the face image set to achieve better recognition performance.

Batur and Hayes [10] proposed a segmented linear subspace model to generalize the 3D linear subspace model so that it is robust to shadows. Each image in the training set is segmented into regions that have similar surface normals by k-mean clustering, then for each region a linear subspace is estimated. Each estimation only relies on a specific region, so it is not influenced by the regions in shadow.

2.1.1.2 Illumination Cone

Belhumeur and Kriegman [14] proved that all images of a convex object with Lambertian surface from the same viewpoint but illuminated by an arbitrary number of distant point sources form a convex Illumination Cone. The dimension of this illumination cone is the same as the number of distinct surface normals. This illumination cone can be constructed from as few as three images of the surface, each under illumination from an unknown point source. The illumination cone is a convex combination of extreme rays given by $x_{i,j} = \max(Bs_{ij}, 0)$, where $s_{i,j} = b_i \times b_j$, and b_i, b_j are two different rows of a matrix B where each row is the product of albedo with surface normal vector. Kriegman and Belhumeur showed in [65] that for any finite set of point sources illuminating an object viewed under either orthographic or perspective projection, there is an equivalence class of object shapes having the same set of shadows. These observations are exploited by Georghiades *et al.* [42] for face recognition under variable lighting.

2.1.1.3 Spherical Harmonics

Spherical harmonics method is proposed by Basri and Jacobs [9], and contemporarily by Ramamoorthi and Hanrahan [86]. Assuming arbitrary light sources (point sources or diffuse sources) distant from an object of Lambertian reflectance property, Basri and Jacobs [8] show that ignoring cast shadow the intensity of object surface can be approximated by a 9-dimensional linear subspace based on a *spherical harmonic* representation.

Zhang and Samaras [125] proposed two methods for face recognition under arbitrary unknown lighting by using the spherical harmonics representation, which requires only one training image per subject and no 3D shape information. In the first method [124] the statistical model of harmonic basis images are built based on a collection of 2D basis images. For a given training face image, the basis images for

this face can be estimated based on maximum a posterior estimation. In the second method a 3D morphable model and the harmonic representation are combined to perform face recognition with both illumination and pose variation.

2.1.1.4 Nine point lights

Lee *et al.* [67] showed that there exists a configuration of nine point source directions such that a subspace resulting from nine images of each individual under these nine lighting sources is effective at recognition under a wide range of illumination conditions. The advantage of this method is that there is no need to obtain a 3D model of surface as in the spherical harmonics approach [8], or to collect a large number of training images as in the statistical modelling approaches.

2.1.1.5 Generalized Photometric Stereo

Recently, Zhou *et al.* [132] analyzed images of the face class with both the Lambertian reflectance model and the linear subspace approach. The human face is claimed to be an example of a so-called *linear Lambertian object*, which is not only an object with Lambertian surface, but also a linear combination of basis objects with Lambertian surfaces. The albedo and surface normal vectors of each basis object for the face class form a matrix called class-specific albedo/shape matrix, which can be recovered by a *generalized photometric Stereo* process from the bootstrap set. The model is trained using Vetters 3D face database [16]. Excellent performance was reported. The work was further extended for multiple light sources.

2.1.2 Illumination Invariant Features

Adini *et al.* [4] presented an empirical study that evaluates the sensitivity of several illumination insensitive image representations to changes in illumination. These representations include edge map, image intensity derivatives, and image convolved with a 2D Gabor-like filter. All of the above representations were also followed by a log

function to generate additional representations. However, the recognition experiment on a face database with lighting variation indicated that none of these representations is sufficient by itself to overcome the image variation due to the change of illumination direction.

2.1.2.1 Features Derived from Image Derivatives

Line edge map [40] is proposed for face recognition by Gao and Leung. The edge pixels are grouped into line segments, and a revised Hausdorff Distance is designed to measure the similarity between two line segments. Chen *et al.* [23] showed that for any image, there are no discriminative functions that are invariant to illumination, even for objects with Lambertian surface. However, they showed that the probability distribution of the image gradient is a function of the surface geometry and reflectance, which are the intrinsic properties of the face. The *direction of image gradient* is revealed to be insensitive to illumination change. The recognition performance using gradient direction is close to the illumination cone approach. *Relative Image Gradient* feature is applied by Wei and Lai [116] and Yang *et al.* [105] for robust face recognition under lighting variation. The relative image gradient $\overline{G}(x, y)$ is defined as $\overline{G}(x, y) = \frac{|\Delta I(x,y)|}{\max_{(u,v)\in W(x,y)}|\Delta I(x,y)|+c}$, where $I(x, y)$ is the image intensity, Δ is the gradient operator, $W(x, y)$ is a local window centered at (x, y), and c is a constant value to avoid dividing by zero.

Zhao and Chellappa [130] presented a method based on Symmetric Shape from Shading for illumination insensitive face recognition. The symmetry of every face and the shape similarity among all faces are utilized. A prototype image with normalized illumination can be obtained from a single training image under unknown illumination. Their experiments showed that using the prototype image significantly improved the face recognition based on PCA and LDA.

Sim and Kanade [100] developed a statistical shape from shading model to recover

face shape from a single image and to synthesize the same face under new illumination. The surface radiance $i(x)$ for location x is modeled as $i(x) = n(x)^T \times s + e$, where $n(x)$ is the surface normal with albedo, s is the light source vector, e is an error term which models shadows and specular reflections. A bootstrap set of faces with labeled varying illuminations is needed to train the statistical model for $n(x)$ and e. The illumination for an input image can be estimated using kernel regression based on the bootstrap set, then $n(x)$ can be obtained by maximum a posterior estimation and the input face under a new illumination can be synthesized.

2.1.2.2 Quotient Image

Shashua and Riklin-Raviv [95] treat face as an *ideal class of object*, i.e. the objects that have the same shape but differ in the surface albedo. The *quotient image* $Q_y(u, v)$ of object y against object a is defined by $Q_y(u, v) = \frac{\rho_y(u,v)}{\rho_a(u,v)}$, where $\rho_y(u, v)$, $\rho_a(u, v)$ are albedo of the two objects. The image Q_y depends only on the relative surface texture information, and is independent of illumination. A bootstrap set containing N faces under three unknown independent illumination directions is employed. Q_y of a probe image $Y(u, v)$ can be calculated as $Q_y(u, v) = \frac{Y(u,v)}{\sum_j \overline{A}_j(u,v)x_j}$, where $\overline{A}_j(u, v)$ is the average of images under illumination j in the bootstrap set, and x_j can be determined from all the images in bootstrap set and $Y(u, v)$. Then the recognition is performed based on the quotient image.

Based on the assumption that faces are an ideal class of objects, Shan *et al.* [93] proposed *Quotient Illumination Relighting*. When the illumination in the probe image and the target illumination condition are both known and exist in the bootstrap set, the rendering can be performed by a transformation learnt from the bootstrap set.

Chen and Chen [22] proposed a *generic intrinsic illumination subspace* approach. Given the ideal class assumption, all objects of the same ideal class share the same

generic intrinsic illumination subspace. Considering attached shadows, the appearance image of object i in this class under a combination of k illumination sources $\{l_i\}_{i=1}^{k}$ is represented by $I_i(x,y) = \rho_i(x,y)\sum_{j=1}^{k}\max(n(x,y)i_j, 0)$, where $\rho_i(x,y)$ is the albedo, and $n(x,y)$ is the surface normal vector of all objects in the class. The illumination image is defined as $L(x,y) = \sum_{j=1}^{k}\max(n(x,y)l_j, 0)$. The illumination images of a specific ideal class form a subspace called generic intrinsic illumination subspace, which can be obtained from a bootstrap set. For a given image the illumination image can be estimated by $L = Bl$, where $l = argmin\|Bl - L^*\|$. Here B is the basis matrix of the intrinsic illumination subspace, and L^* is an initial estimation of illumination image based on smoothed input image. Finally $\rho(x,y)$ can be obtained by $\rho(x,y) = \frac{I(x,y)}{L(x,y)}$. The method was evaluated on CMU-PIE and Yale B face databases and showed significantly better results than the quotient image method. It is also shown that enforcing nonnegative light constraint will further improve the results.

2.1.2.3 Retinex Approach

In *retinex* approaches the luminance is estimated by the smoothed image. The image can then be divided by the luminance to obtain the reflectance, which is an invariant feature to illumination. A single Gaussian function is applied to smooth the image in the single scale retinex approach [57], and the sum of several Gaussian functions with different scales is applied in the multi-scale retinex approach [56]. Logarithm transform is employed to compress the dynamic range in [57] and [56].

Wang *et al.* [115] defined Self-Quotient Image, which is essentially a multi-scale retinex approach. However, instead of using isotropic smoothing as in [56], anisotropic smoothing functions with different scales are applied. Each anisotropic smoothing

function is a Gaussian weighted by a thresholding function. Zhang *et al.* [127] proposed a morphological quotient image (MQI) method in which mathematical morphology operation is employed to smooth the original image to obtain a luminance estimate.

Gross and Brajovic [45] solve luminance L for the retinex approach by minimizing an anisotropic function over the image region $\Omega : J(L) = \iint_\Omega \rho(x,y)\,(L-I)^2\,dxdy + \lambda \int\int_\Omega \left(L_x^2 + L_y^2\right) dxdy$, where $\rho(x,y)$ is space varying permeability weight which controls the anisotropic nature of the smoothing. L_x and L_y are the spacial derivatives of L, and I is the intensity image. The isotropic version of function $J(L)$ can be obtained by discarding $\rho(x,y)$.

In the total-variation based quotient image(TVQI) approach [24], the luminance $u(x)$ is obtained by minimizing $\iint_\Omega |\Delta u(x)| + \lambda\,|I(x) - u(x)|\,dx$ over all points x in image $I(x)$.

2.1.2.4 Transformation domain features

Recently methods based on the frequency domain representation have received attention. Savvides *et al.* [88] performed PCA in the phase domain and achieved impressive results on the CMU-PIE database [99]. This so-called *Eigenphase* approach improved the performance dramatically compared to Eigenface, Fisherface and 3D linear subspace approach. Meanwhile, they further showed that even with partial face images the performance of the Eigenphase approach remains excellent and the advantages over other approaches are even more significant. Heo *et al.* [50] showed that applying Support Vector Machines directly on phase can lead to even better performance than the Eigenphase approach mentioned above.

In [118] a quaternion correlation method in a wavelet domain is proposed and good performance is achieved on the CMU-PIE database with only one training sample per subject. The subband images after discrete wavelet decomposition are encoded

into a 2-D quaternion image. Quaternion Fourier Transform is then performed to transfer the quaternion image to quaternion frequency domain, where a quaternion correlation filter is applied. Qing *et al.* [84] showed that the Gabor phase is tolerant to illumination change and has more discriminative information than phase in the Fourier spectrum.

Savvides *et al.* proposed a series of work based on advance correlation filters [87, 89]. A pre-whitening spectrum stage is usually adopted to emphasize higher frequency components followed by phase matching. Llano *et al.* [41] examined the sensitivity of several frequency domain representations of face image to illumination change. Those representations are the magnitude, phase, real part and imaginary part of the Fourier spectrum of original face image, and those of gradient image. The gradient image is defined as an image where each pixel has a complex value with the horizontal gradient of the original image as the real part, and the vertical gradient as imaginary part. The experimental results on the normal illumination set and the darken set of the XM2VTS face database showed that the real part of the Fourier spectrum of the gradient image is less sensitive to illumination change than other representations.

2.1.2.5 Local Binary Pattern

Local binary pattern (LBP) is a local feature which characterizes the intensity relationship between a pixel and its neighbors. LBP is unaffected by any monotonic grayscale transformation in that the pixel intensity order is not changed after such a transformation. Furthermore, for a region with a number of pixels, a histogram of the LBP patterns associated with respective pixels within this region tends to be a good feature for face recognition. LBP has been used in [51, 69] as an illumination invariant feature.

2.1.3 Photometric Normalization

Histogram Equalisation [44] is the most commonly used approach. By performing histogram equalisation, the histogram of the pixel intensities in the resulting image is flat. It is interesting that even for images with controlled illumination (such as face images in the XM2VTS database), applying histogram equalisation still offers performance gain in face recognition. Shan *et al.* [93] proposed Gamma Intensity Correction for illumination normalisation. The corrected image $G(x, y)$ can be obtained by performing an intensity mapping: $G(x, y) = cI(x, y)^{\frac{1}{\gamma}}$, where c is a gray stretch parameter, and γ is the Gamma coefficient.

In Homomorphic filtering approach [44] the logarithm of the equation of the reflectance model is taken to separate the reflectance and luminance. The reflectance model often adopted is described by $I(x, y) = R(x, y) \times L(x, y)$, where $I(x, y)$ is the intensity of the image, $R(x, y)$ is the reflectance function, which is the intrinsic property of the face, and $L(x, y)$ is the luminance function. Based on the assumption that the illumination varies slowly across different locations of the image and the local reflectance changes quickly across different locations, a high-pass filtering can be performed on the logarithm of the image $I(x, y)$ to reduce the luminance part, which is the low frequency component of the image, and amplify the reflectance part, which corresponds to the high frequency component.

Du and Ward [32] performed illumination normalization in the wavelet domain. Histogram equalisation is applied to low-low subband image of the wavelet decomposition, and simple amplification is performed for each element in the other 3 subband images to accentuate high frequency components. Uneven illumination is removed in the reconstructed image obtained by employing inverse wavelet transform on the modified 4 subband images.

Xie and Lam [119] proposed an illumination normalization method which is called Local Normalization. They split the face region into a set of triangular facets, the

area of which is small enough to be considered as planar patch. The main idea of this approach is to normalize the intensity values within each facet to be of zero mean and unit variance.

Short *et al.* [97] compared five photometric normalization methods, namely illumination insensitive eigenspaces, multiscale Retinex method, homomorphic filtering, a method using isotropic smoothing to estimate luminance, and one using anisotropic smoothing [45]. Each method is tested with/without histogram equalisation performed in advance. Interestingly it was found that histogram equalisation helped in every case. It is shown that using anisotropic smoothing method as photometric normalisation led to the most consistent verification performance for experiments across the Yale B, BANCA [6] and XM2VTS [77] databases.

Chen *et al.* [26] employed DCT to compensate for illumination variation in the logarithm domain. The uneven illumination is removed in the image reconstructed by inverse DCT after a number of DCT coefficients corresponding to low frequency are discarded.

2.1.4 3D Morphable Model

Blanz and Vetter [16] proposed face recognition based on fitting a 3D morphable model. The 3D morphable model describes the shape and texture of face separately based on the PCA analysis of the shape and texture obtained from a database of 3D scans. To fit a face image under unknown pose and illumination to the model, an optimisation process is needed to optimize shape coefficients, texture coefficients along with 22 rendering parameters to minimise the difference of the input image and the rendered image based on those coefficients. The rendering parameters include pose angles, 3D translation, ambient light intensities, directed light intensities and angles, and other parameters of the camera and color channels. The illumination model of Phong is adopted in the rendering process to describe the diffuse and specular

reflection of the surface. After fitting both the gallery images and the probe images to the model, the recognition can be performed based on the model coefficients for shape and texture. Good recognition performance across pose and illumination is achieved in experiments on CMUPIE and FERET face database.

2.2 Active Approaches

In active approaches additional devices (optical filters, active illumination sources or specific sensors) usually need to be involved to actively obtain different modalities of face images that are insensitive to or independent of illumination change. Those modalities include 3D face information [19] and face images in those spectra other than visible spectra, such as thermal infrared image [43] and near-infrared hyperspectral image [83].

2.2.1 3D information

3D information is one of the intrinsic properties of a face, which is invariant to illumination change. The surface normal information is also used in some passive approaches described in the previous section, however, they are recovered from the intensity images captured by the visible light camera. This section discusses the 3D information acquired by active sensing devices like 3D laser scanners or stereo vision systems.

3D information can be represented in different ways. The most commonly used representations are range image, profile, surface curvature, Extended Gaussian Image(EGI), Point Signature, and etc. Surveys on 3D face recognition approaches can be found in [18,19,90]. The 3D modality can be fused with 2D modality, i.e. texture, to achieve better performance [19, 21]. Nevertheless, it should be noted that the 2D face images which are combined with 3D face info as reported in [19,21] are captured in a controlled environment. It is still not clear how much the fusion will help in the case of uncontrolled environment due to the impact of uncontrolled illumination on

the 2D face intensity images.

Kittler *et al.* [61] reviewed the full spectrum of 3D face processing, from sensing to recognition. The review covers the currently available 3D face sensing technologies, various 3D face representation models and the different ways to use 3D model for face recognition. In addition to the discussion on separate 2D and 3D based recognition and the fusion of different modalities, the approach involving 3D assisted 2D recognition is also addressed.

2.2.2 Infrared

Visible light spectrum ranges from $0.4\mu m - 0.7\mu m$ in the electromagnetic spectrum. The infrared spectrum ranges from $0.7\mu m - 10mm$. It can be divided into 5 bands, namely: Near-Infrared (NIR) $(0.7 - 1.4\mu m)$, the Short-Wave Infrared (SWIR) $(1.4 - 3.0\mu m)$, the Mid-Wave Infrared(MWIR) $(3.0 - 8.0\mu m)$, the Long-Wave Infrared(LWIR) $(8.0 - 15.0\mu m)$, and Far-Infrared(FIR) $(15.0\mu m - 10mm)$. NIR and SWIR belong to reflected infrared $(0.7 - 3.0\mu m)$, while MWIR and LWIR belong to thermal infrared $(3.0\mu m - 15.0\mu m)$. Similar to the visible spectrum, the reflected infrared contains the information about the reflected energy from the object surface, which is related to the illumination power and the surface reflectance property. Thermal Infrared directly relates to the thermal radiation from object, which depends on the temperature of the object and emissivity of the material [62]. Figure 1 shows face images in different infrared spectrum range.

2.2.2.1 Thermal Infrared

A survey on visual and infrared face recognition is presented in [62]. Wilder at al. [117] showed that with minor illumination changes and for subjects without eyeglasses, applying thermal image for face recognition does not lead to significant difference compared to visible images. However, for scenarios with huge illumination changes and facial expressions, superior performance was achieved based on radiometrically

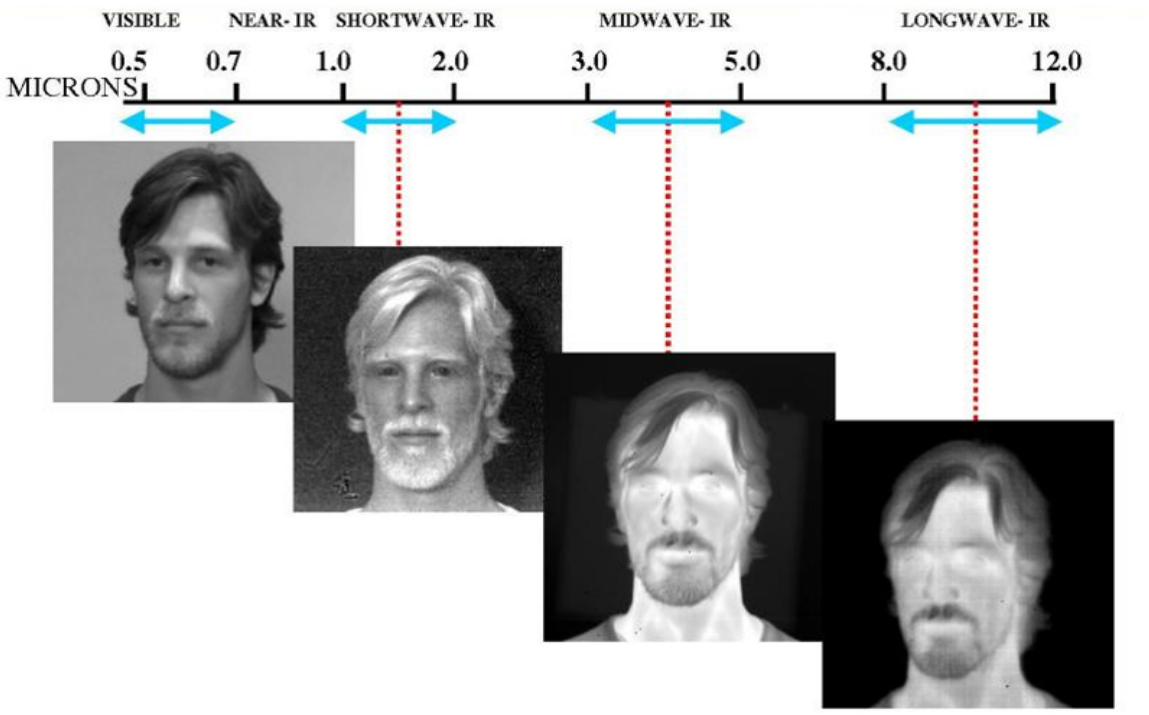


Figure 1: Infrared spectrum [1].

calibrated thermal face images than that based on visible image [101, 103]. The experiments in [27] show that the face recognition based on thermal images degrades more significantly than visible images when there is a substantial passage of time between the acquisition of gallery images and probe images. This result was proved to be reproducible by [102]. However, it is shown that with a more sophisticated recognition algorithm the difference of recognition performance across time based on thermal face and visible face is small.

Despite the independence of visible light, the thermal imagery has its own disadvantages. The temperature of the environment, physical conditions and psychological conditions will affect the heat pattern of the face [11]. Meanwhile, the infrared is opaque to eyeglasses. All the above motivate the fusion of thermal infrared image with visible images for face recognition. Various fusion schemes have been proposed [11, 27, 62, 102] and shown to lead to better performance than the recognition based on either modality alone. The thermal face recognition experiments are

usually conducted on the face database from the University of Notre Dame [27] or the Equinox face database [1]. The former contains the visible spectrum images and LWIR images of 240 subjects without glasses, but with different lighting and facial expressions. The latter was collected by Equinox Corp. and contains the visible images and LWIR images of a total of 115 subjects. In [60], thermal face recognition is performed in an operational scenario, where both indoor and outdoor face data of 385 subjects is captured. When the system is trained on indoor sessions and tested on outdoor sessions, the performance degrades no matter whether one is using thermal imagery or visible imagery. However, the thermal imagery substantially outperformed visible imagery. With the fusion of both modalities, the outdoor performance can be close to indoor face recognition.

2.2.2.2 Active Near-IR Illumination

The Near-IR band falls into the reflective portion of the infrared spectrum, between the visible light band and the thermal infrared band. It has advantages over both visible light and thermal infrared. Firstly, since it can be reflected by objects, it can serve as an active illumination source, in contrast to thermal infrared. Secondly, it is invisible, making active Near-IR illumination unobtrusive. Thirdly, unlike thermal infrared, Near-IR can easily penetrate glasses.

Pan *et al.* [80] performed face recognition in hyperspectral images. A CCD camera with a liquid crystal tunable filter was used to collect images with 31 bands over near-infrared range. It was shown the hyperspectral signatures of the skin from different persons are significantly different, while those belonging to the same person are stable. Above 91% rank one correct identification rate is obtained in the recognition experiments on frontal hyperspectral face images.

Most recently, Li *et al.* [69] proposed a face recognition system based on active Near-IR lighting provided by Near-IR Light-Emitting Diodes(LEDs). The Near-IR

face image captured by this device is subject to a monotonic transform in the gray tone, then LBP feature is extracted to compensate for this monotonic transform to obtain an illumination invariant face representation. Zhao and Grigat [128] performed face recognition in Near-IR images based on Discrete Cosine Transform(DCT) feature and SVM classifier. Although infrared image is invariant to visible illumination, it is not independent of the environmental illumination. This is because environmental illumination contains energy in a wide range of spectrum, including infrared. The variation of the infrared component in the environmental illumination will impose variation in the captured image.

One solution to maximize the ratio between the active source and the environmental source is to apply synchronized flashing imaging by Hizem *et al.* [52]. A powerful active illumination source is desirable. Illuminants such as LEDs can provide very powerful flash but only for very short time to avoid the internal thermal effects which might destroy the LEDs. The idea in Hizem *et al.* [52] is to synchronise the sensor exposure time with the powerful flash. The sensor is exposed to the environmental illumination only for the same short exposure time as the flash. Since the power of the flash is usually much stronger than the environmental illumination, the contribution of the environmental illumination to the captured image will be minimised.

Nevertheless, the illumination variation problem can only be alleviated but not completely solved by the above mentioned approach. For indoor environment, the infrared energy in environmental illumination is low and will not cause much problem, while in outdoor environment, the infrared energy can be very strong.

CHAPTER III

NEAR INFRARED FACE RECOGNITION

3.1 Overview

The goal of this chapter is to provide a consumer grade biometric with NIR face recognition. Due to restrictions on the mathematical capabilities of the consumer grade hardware, proven algorithms of low complexity such as intensity based segmentation, cross correlation, principal component analysis (PCA) and linear discriminant analysis (LDA) are used since they will work well in conjunction with the provided NIR band limited CMOS imager. As a result, the volume of silicon used in manufacturing will be far less than of digital hardware with very low production costs, e.g. US $5 or less.

Research has already been done in the field of face detection and recognition; for detection, some of the computationally less intensive approaches use variations of skin color matching or shape detection based on pre-defined databases. Other more intensive approaches use Support Vector Machine, multi layer Probabilistic Neural Networks and Wavelets. Since these are application based approaches, some of the assumptions made are ad-hoc and cannot be generalized towards a particular approach [74], [92], [60]. In our system, near infrared (NIR) of a wavelength of 940 nm was chosen as an illumination source. Prior research has been done on detection using mid and long wave IR (thermal imaging) [63]. Solid state NIR emitters are inexpensive, and bright enough to provide a fill in for shadows during daytime operation, or to provide total face illumination at night. Although the illumination is invisible to the subject being imaged, most CMOS based imagers have a high degree of sensitivity at these wavelengths. A low cost system using NIR illuminators and a CMOS based

imager with integrated electronics to perform the recognition algorithms should be practical.

To achieve the goal, the most indispensable tasks are included in the system which is face detection, eye detection and recognition. Figure 2 illustrates how the subsystems are interconnected in the overall system. Cluster analysis in face detection part decides whether the input video frame has a face. If a face is present, the face region is passed to eye detection otherwise the frame is disregarded. Eye detection performs boundary analysis on the face image and extracts eye coordinates. Accurate positioning of eyes is indispensable to make the face recognition work. The face image is preprocessed based on the eye locations. And finally identification is made based on the result of face recognition. A detailed functional description of the algorithms for the subsystems is given in the following sections relating them to previous work in each area.

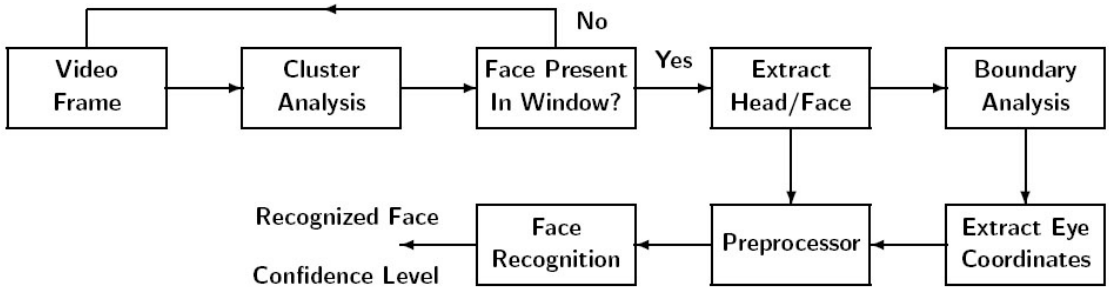


Figure 2: System flowchart.

In order to train and test each subsystem and the end to end system, data was acquired in the form of video captures that simulated a person sitting in the driver's seat of a vehicle, fastening the seat belt and starting to drive with mild body movements. Two videos were acquired of each of the four subjects tested, with each video containing approximately 1000 frames. The videos are in grey scale and the frame size is 320 by 240. A simple off the shelf imager coupled with an optical IR filter was used to obtain the video stream.

3.2 Face Detection

Visible and NIR light obey the *Inverse Square Law* which states that brightness of the light source is inversely proportional to the square of its distance [2]. Using this property, one can observe that objects closer to the light source will be more illuminated than those farther away. Consider a person located near the lighting source, this will cause the person to "glow" or be saturated as compared to the background. Our face detector is based on this phenomenon. We assume that after converting a gray level image with a face to a binary image by thresholding, the largest cluster of white pixels (which are corresponding to the intensity values larger than a threshold value in the gray level image) is a face or an upper body including a face.

The face detection subsystem has three steps: thresholding, cluster analysis, and the final decision step. In the cluster analysis step, a simple erosion and dilation is performed on the binary image to rid any stray pixels that could be present because of noise (Fig. 3). Next, the various clusters are collected with centers computed for each cluster using median information on each axis. The cluster that is closest to the center of the frame is of interest to us. We choose this cluster because the driver of a vehicle is going to be located near the center of the camera's view and sitting in an upright vertical position. The median is preferred over the mean because the median will be directed towards the center of a concentration of pixels and is not easily steered away by stray pixels.

The selected cluster belongs to one of three categories under the assumption mentioned above: face, upper body (including the face), or neither of them. The category that the cluster belongs to is determined in the final decision step with statistical measures of face dimensions: height and width. In the first pass, the cluster height is compared to it's statistical value to decide the category of the cluster. The details are shown below.

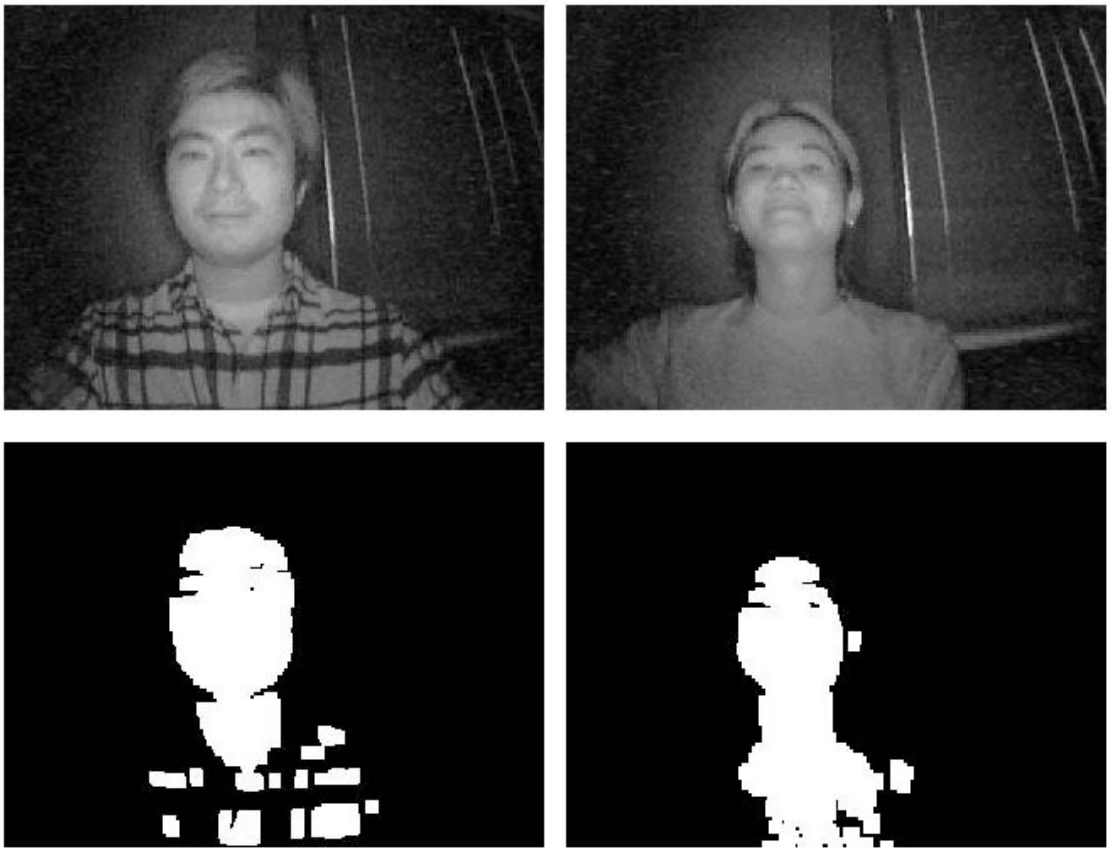


Figure 3: The original images on the top row and the thresholded binary images on the bottom row.

```
select cluster height
    Case ≥ mean height + 2 × standard deviation
        Possible upper body (including face);
    Case ≤ mean height ± 2× standard deviation
        Possible face;
    Case default
        Neither;
end select
```

Given that the classification results in a possible face, we use the height of the cluster combined with the statistical measure of the aspect ratio to determine the

width. The face region with the height and width is extracted from the image. Given that the classification results in a possible upper body, the proper face region in the upper body cluster is extracted as the pseudo code below. If no face is found in either of two cases, we decide that there is no face in the image.

```
if cluster is classified as body
    face width = distance between points at location (mean
            height ÷ 2 + top of cluster)
    if face width < mean width + 3 × standard deviation
        Face region found;
endif
```

Approximately 200 face images were manually extracted from the 7 videos. The binary image threshold was set to one standard deviation less than the mean of the pixel intensity values of the face images, which allows us to segment images with low illumination. The face images also serve as training data to get statistics of face dimensions.

The 7 videos were tested to verify the accuracy of the face detection. To perform this task successfully, each frame was visually inspected by one of the authors. Table 1 shows the result for each video. The detection rate varies from 87.28% to 99.81%. Figure 4 shows sample results of the face detection.

Table 1: Face detection success rates.

Video	Total Frames	Frames with faces present	Frames with faces incorrectly detected	Frames with face correctly detect	Detection Rate (%)
1	1088	931	45	813	87.28
2	1128	1046	2	1044	99.81
3	750	690	11	678	98.26
4	1342	1126	2	1116	99.11
5	563	462	5	410	88.74
6	1324	1168	0	1146	98.12
7	747	676	21	618	91.42

Figure 4: Sample results of the face detection using thresholds.

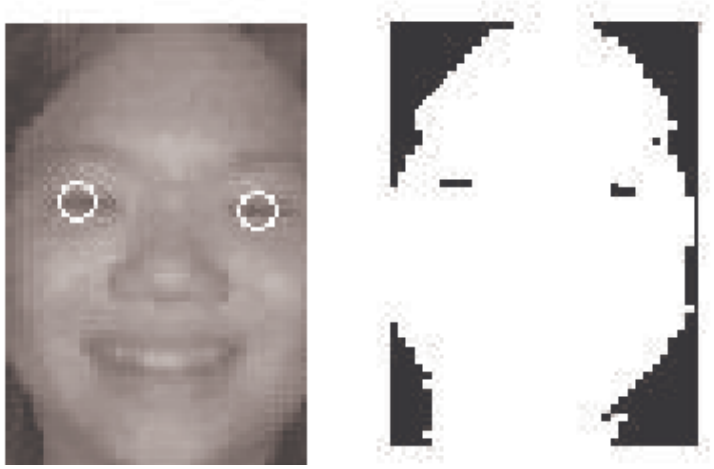
3.3 Eye Detection

A large amount of work has been done on finding facial features like eyes. One approach is to locate the eyes in images without locating the face [64]. Another approach is to use a combination of physiological properties of the eyes, Kalman trackers to model eye/head dynamics, and a probabilistic appearance models to locate the eye appearances [49].

The eye location algorithm is implemented on images from which the background is cropped out leaving only the face. The eye co-ordinates are used to perform the required transformation on the face image. Thus, all the face images are centered based on the eye coordinates in the preprocessing step before they are sent into the recognition system.

Binary representations of the face images were utilized to outline the face and eyes. The exterior boundary points of the binary face were determined, as well as the boundaries of holes inside the face by using 8-connected neighborhood connectivity. It is assumed that the eyes were the largest objects in the upper face region. Therefore, only the set of boundary points is chosen which contains the largest number of points. Once the face is outlined the upper half of the face is analyzed to locate the eyes. Either the left eye or right eye is detected first. This can be determined from the x

location of the first eye compared to the vertical midline of the face. Based on which eye is detected first, the left upper region or the right upper region of the face is utilized to search for the second eye (See Fig. 5). The accuracy of the eye detection was verified by the visual inspection of 4 test videos by one of the authors (Table 2). The detection rates of eyes are lower than the detection rates of faces because the pixel intensity difference between eye regions and face regions is not as large as the pixel intensity difference between face regions and background regions.



(a) Original (b) Binary

Figure 5: Eye detection example.

Table 2: Results of the eye detection.

Video	Face frames	Correctly found eyes	Incorrectly found eyes	Incorrectly rejected eyes	Detection rate (%)
1	662	393	249	20	59.37
2	374	268	98	8	71.66
3	564	441	99	24	78.19
4	375	242	121	12	64.53

3.4 Face Recognition

Among many approaches to the problem of face recognition, appearance based subspace analyses is one of the oldest approaches delivering the most promising results. Two of the more popular appearance based subspace analysis are Eigenface methods

which are equivalent to Principal Component Analysis (PCA) and Fisherface methods which are the combination of PCA and Linear Discriminant Analysis (LDA). PCA finds a set of the most representative projection vectors such that the projected samples retain most information about the original samples [107]. On the other hand, LDA uses the class information to find a set of vectors that maximize the between-class scatter while minimizing the within-class scatter [12].

Before further explanation we will clarify the terms for the data sets used in the recognizer. The recognizer finds an adequate subspace using a set of face images labeled with the subject's identity, which we will call *the training set*. After the subspace is trained, unlabeled face images (*test set*) are to be identified as one of the subjects whose face images are given (*gallery set*). Figure 6 shows the face recognition flowchart including the usage of image sets.

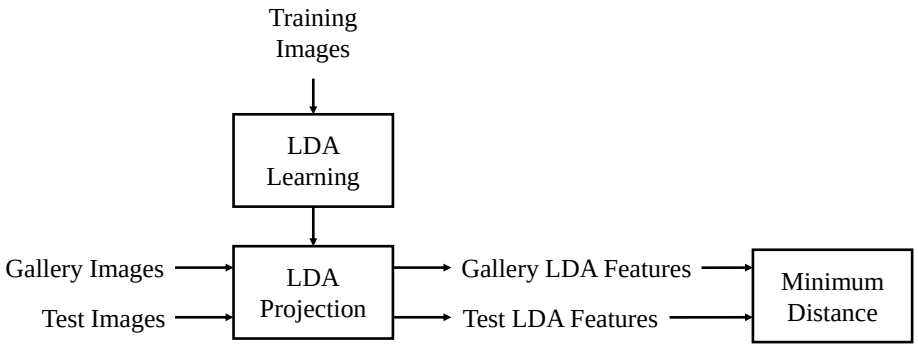


Figure 6: Face recognition flowchart.

An experiment was performed to determine if a training set built from subject images not in the gallery set could result in recognition performance close to that of a recognizer using a training set built from actual gallery images. Frontal face images of CMU PIE database [98] are used as the training set that is different from the gallery set. The training set has approximately 170 images for each of the 68 subjects. 4 videos of 4 subjects in our data set are used for gallery set and the remaining 4 videos are used for test set. Eigenface methods and Fisherface methods are applied to both

cases. Figure 7 shows the recognition rates on various numbers of coefficients. It shows that the recognition rate for the case when CMU PIE database is used as a training set is almost as good as the case when the gallery set is used as a training set for both of subspace methods. Since the difference is less than 10%, we can say that the CMU PIE images are good enough as the training set of our application. The result also confirms that Fisherface methods perform better than Eigenface methods.

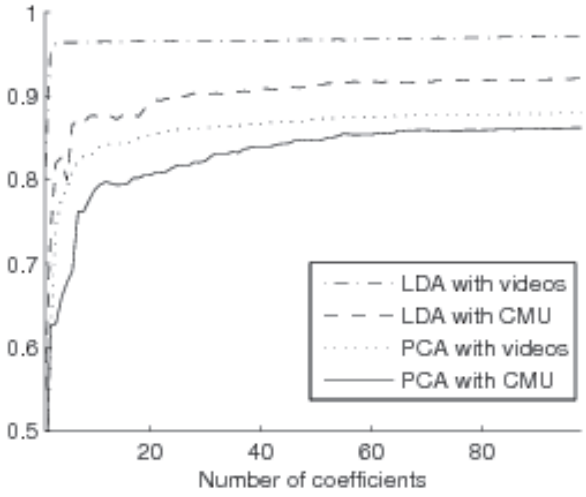


Figure 7: Recognition hit rates comparison.

Before subspace methods are applied to the detected face images in the video frames, the images go through a preprocessing step. It includes geometrical transformation, masking and pixel value normalization. Geometrical transformation processes the face images so that eyes are located in the predefined positions. Masking takes pixels inside the face boundary. And the pixel values are normalized so that they have 0 mean and 1 standard deviation.

Preprocessed images are projected on the Fisherface subspace and minimum distance calculated between the projected test data and the mean values of the projected gallery data of each subject is applied to identify the test subjects of the images. The minimum distance method is equivalent to the maximum likelihood decision assuming that the distributions of all the subjects are equal and priors are equal. The decisions

on the frames from the beginning of the video to the current frame are used to vote the overall decision.

The projected test data are also used to decide whether the preprocessed image comes from a frontal pose or not. If the image is not a frontal face, the projected test data will have a distance from origin of the subspace larger than that of a frontal face. A simple threshold can be used to reject the image. This approach is similar to the method mentioned in [107] but different. The authors suggest using the distance from a data point in the image space to the face space as a measure for face detection. But what we use is the distance from a point to the origin in the Fisherface subspace.

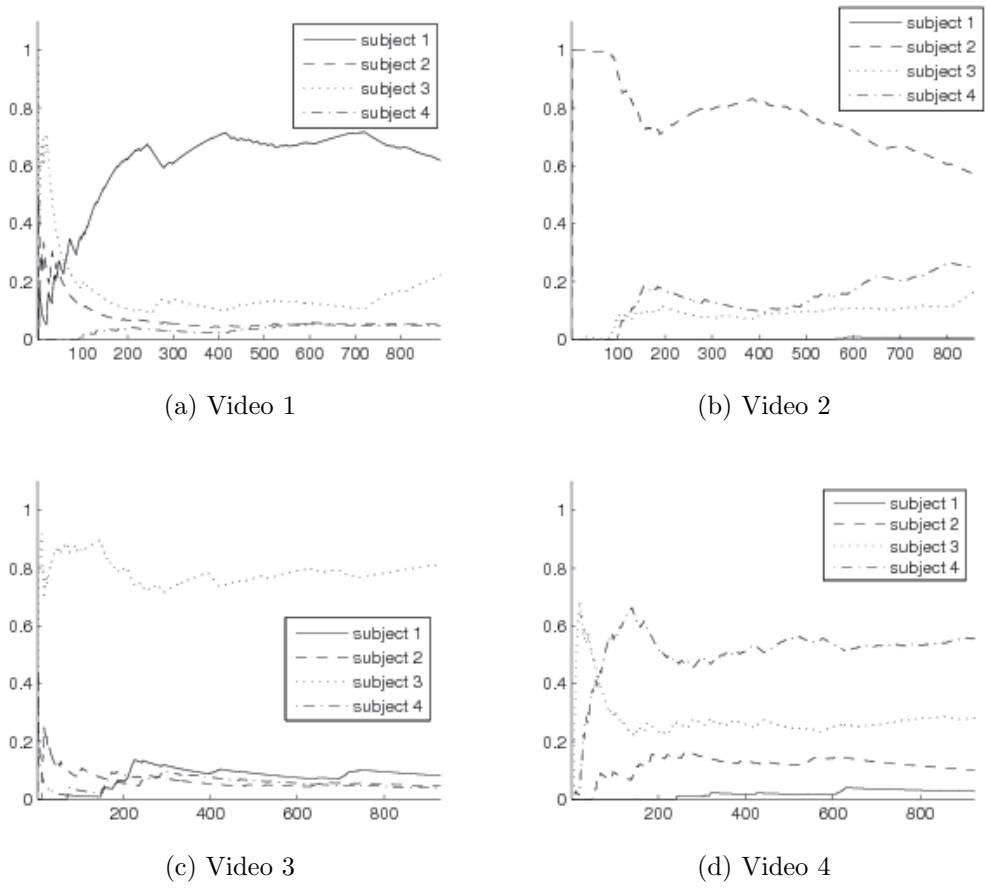
3.5 System Results and Discussion

We performed an end to end evaluation of our system using the 8 videos of 4 subjects, which was introduced in Section 3.1. Considering each subject as a member of a family of four, the goal is to recognize which one of the 4 is sitting in the driver's seat of a car. The overall voting ratios on all the subjects at the final frames of videos are shown in Table 3. Since all videos have the maximum voting ratios on the corresponding subjects,it is concluded that all the subjects are successfully identified at the end of videos. Specifically, video 1 and 3 shows relatively high detection ratios compared to the others. The Graphs shown in Fig. 8 illustrate the change of voting results over all the video sequences. Video 1, 3 and 4 indicate that the voting ratios are stabilized after the 400th frame approximately. Detection ratio of video 2 is high at the beginning of the sequence but then begins to decrease. To figure out the difference between the results, we examined the videos and came to find that the subject in video 4 is in frontal position at the beginning of the video but it moves actively as time progresses. That explains why the detection ratio is decreased over time.

Additional tests were conducted to evaluate the performance of subsystems (Table

Table 3: Voting ratios of 4 videos.

Video	Subject in Video	Subject 1	Subject 2	Subject 3	Subject 4
1	1	0.618	0.050	0.222	0.044
2	2	0.005	0.570	0.163	0.247
3	3	0.078	0.033	0.843	0.028
4	4	0.027	0.099	0.281	0.556



(a) Video 1

(b) Video 2

(c) Video 3

(d) Video 4

Figure 8: Change of voting ratios over the sequence of frames.

4). In the first test, the recognition part of the system is only evaluated with manually specified eye locations on manually determined frontal face frames. In the second test, performance of the overall system is measured on manually detected frontal face frames. And in the third test, performance of the overall system is assessed on all frames of videos. The result of the first test is equivalent to the performance of the face recognition subsystem. If the face and eye detectors find the locations accurately

Table 4: Detection ratio comparison for additional tests.

Video	test 1	test 2	test 3
1	0.956	0.854	0.618
2	0.662	0.591	0.570
3	0.948	0.947	0.843
4	0.969	0.709	0.556

in the presence of face, the difference between the results of test 1 and 2 would be unnoticeable. If the overall system has low false alarm rate, the difference between the results of test 2 and 3 would be insignificant. From the results, we finally conclude that the face recognition subsystem shows a high detection rate. And the difference between the results of test 1 and test 2, and the difference between the results of test 2 and test 3 cannot be decided to be significantly indifferent. Therefore, the face and eye detection subsystems find face with enough accuracy to get a correct identification result in a small group of subjects, but they need to be improved for larger dataset.

This chapter presented a consumer grade biometric system based on face recognition using infrared imaging, with a successful detection result in a small group of subjects. This low cost approach is intended for practical, high volume applications where the distance is minimum and controlled, such as automotive applications and hand held devices. The results can be further improved by increasing the accuracy of the eye detection and the frontal pose detection. Future work will address how to deal with other sources of variations such as illumination, facial expression, and occlusion.

CHAPTER IV

FACE RECOGNITION WITH ACTIVE NEAR INFRARED IMAGE DIFFERENCING

4.1 Overview

Despite many years of active research on the topic, illumination invariant face recognition remains a very difficult and challenging problem. In outdoor environments, for example, lighting conditions vary dramatically throughout the day and from one day to the next. In addition, there may be unknown and highly variable shadows that are cast by static or moving objects. Despite the abilities of humans to recognize faces under these conditions, current face recognition algorithms perform very poorly. Face recognition becomes even more difficult when face recognition must be robust to pose and expression variations.

One approach that has been studied to solve the illumination problem in computer vision systems is to extract features that are invariant to lighting changes. Unfortunately, however, Chen *et al.* showed that there are no illumination invariant discriminative functions for an object with Lambertian reflectance [23]. Shashua and Riklin-Raviv [95] reported that the quotient image, *i.e.*, the image ratio between a test image and linear combination of three non-coplanar illuminated images, depends only on the albedo information, which is illumination free. Jobson, *et al.*, [58] proposed the Multiscale Retinex (MSR) method, which reduces the effect of illumination by using the ratio of the original image to its smoothed version. The self-quotient image (SQI) model [114] is similar but uses a weighted Gaussian filter to obtain the smoothed version. In [25], the authors proposed a logarithmic total variation (LTV) model to

improve SQI using the edge-preserving capability of the total variation model. Authors in [126] proposed illumination insensitive measure called Gradientfaces using the ratio of the vertical derivative to the horizontal derivative. Wang *et al.* [113] proposed Weberfaces method which uses the ratio between local intensity variation and the original pixel value. However, the underlying assumption of a Lambertian model is too simple to describe the real face surface under various illuminations [120] and these methods are unstable in complicated situations when other uncontrolled variations are mixed in [66]. Furthermore, since these methods are sensitive to noise, images are smoothed first with Gaussian kernel filter beforehand.

A completely different approach to illumination-robust face recognition is based on building appearance-based face models that try to model varying illumination explicitly instead of trying to be invariant to it. The main idea behind these model-based approaches is to treat the human face as a Lambertian surface and to find a mathematical description of all images it can produce under all possible illuminations. One such basic model is the illumination cone model proposed by Belhumeur and Kriegman, which states that, when cast shadows are ignored, Lambertian surfaces produce images that lie in a convex polyhedral cone in the image space [14]. Unfortunately, this cone model is very complex to build, providing motivation to search for simpler models that approximate it in the best possible way with minimum complexity. Previously proposed techniques include a spherical harmonics representation independently proposed by Ramamoorthi [85] and Basri and Jacobs [8] and a segmented linear subspace model proposed by Batur and Hayes [10]. Although these methods have achieved a degree of success in some face database such as Yale and CMU PIE, they need a large volume of gallery face images under various illumination conditions which cannot be achieved in many practical applications [120].

Other directions have also been explored to overcome problems caused by illumination changes. One direction is to use 3D (in many case, 2.5D) data obtained from a

laser range scanner or 3D vision method [131]. Because such data captures geometric shapes of face, such systems are less affected by environmental lighting. Moreover, it can cope with rotated faces because of the availability of 3D (2.5D) information for visible surfaces. The disadvantages are the increased cost and slowed speed as well as specular reflections [69]. More importantly, it is shown that the 3D method may not necessarily produce better recognition results: recognition performances achieved by using a single 2D image and by a single 3D image are similar [21]. A commercial development is A4Vision [123]. It is basically a 3D (or 2.5D) face recognition system, but it creates 3D mesh of the face by means of triangulation based on an NIR light pattern projected onto the face. While not affected by lighting conditions, background colors, facial hair, or make-up, it has problems in working under conditions when the user is wearing glasses or opening the mouth, due to limitations of its 3D reconstruction algorithm.

Rather than trying to solve the problem of dealing with illumination variations in the visible band, another approach is using modalities other than the visible spectrum where illumination variations may be less or where variations may be more easily controlled or compensated such as infrared spectrum [20, 43]. The infrared spectrum is further categorized into four sub-bands: near infrared (NIR, 0.75 1.4μm), short wave infrared (SWIR, 1.4 3μm), medium wave infrared (MWIR, 3 8μm), and long wave infrared (LWIR, wavelength 8 15μm). The LWIR spectrum is also refered to as thermal infrared since the spectrum corresponds to thermal radiation from objects near room temperature, including the human body which is slightly higher than room temperature. Face recognition methods using the thermal infrared shows robustness on ambient illumination changes in visible band [43]. However, thermal infrared images are not robust on environmental temperatures and the physical, emotional and health condition of the subject. And another problem of thermal infrared is that it is opaque to eyeglasses. On the other hand, NIR has advantages over both thermal infrared

and visible light. First, it is invisible to human eyes but falls into the reflective range of the infrared spectrum. Therefore, it is possible to build an unobtrusive night-time vision system with active NIR illumination such as NIR light-emitting diodes (LEDs). In day time when there is ambient NIR illumination from the sun, the active NIR illumination sources that are typically mounted near the camera lens illuminate the face from the frontal direction and provide more controlled illumination condition. Second, most CCD and CMOS imaging devices show reasonable responses to NIR, enabling low cost implementation of the system. Third, unlike thermal infrared, NIR is independent of the body temperature and can easily penetrate glasses. Authors in [128] present an NIR-based face recognition system using discrete cosine transform coefficients as features and a support vector machine as the classifier. A linear discriminant analysis-based face recognition system with NIR illumination was proposed in [59]. Li, *et al.*, [69] present an indoor illumination-invariant face recognition system using local binary features. In [39], authors propose NIR face recognition by combining Zernike moments and undecimated discrete wavelet transform. However, as the authors reported in [69], these approaches are not suitable for outdoor face recognition due to the strong NIR component in sunlight.

NIR has also been used for eye tracking and gaze analysis. Such systems are based on the specular and retroreflective response of human eyes to active NIR illumination [28, 36, 37, 54]. Recently, robust eye tracking methods have been developed that are based on the bright/dark pupil images with a differential lighting scheme [47, 48, 133]. Here, several LEDs are placed next to the camera lens, considered to be on the camera axis, and several additional LEDs are placed away from the camera lens and are considered to be off-axis. Common approaches using this setup involve the subtraction of two consecutive video frames with different illuminations and then applying a threshold to the difference image to obtain a binary image. Then,

a connected component analysis is applied to get initial eye pupil candidates. Unfortunately this technique relies on the experimental setup to be placed in a controlled environment such as a laboratory or dark room where the background or lighting does not change, thus allowing the bright pupil effect to be prevalent. In addition, for daytime operation, the intensity of the IR present in sunlight far exceeds that of most artificial illuminators, thus hiding the bright spot in the eyes and causing the bright pupil effect-based eye detector to fail.

4.2 System and Hardware

A high-level overview of the system described in this chapter that performs the task of person identification for vehicle personalization is shown in Fig. 9. At the front end is the video hardware that produces a video of the driver that is to be identified. As discussed in previous sections, uncontrolled illumination is a major problem for face detection and recognition systems. Therefore, in order to achieve high recognition rates during the day as well as at night, in uncontrolled environments, this system operates in the NIR frequency spectrum, with NIR illumination provided by an array of IR LEDs. Since CCD and CMOS image sensors are sensitive to NIR, which is just above the frequency range that is visible by the human eye, the driver may be illuminated by an IR source non-invasively and video recorded by a camera that is sensitive in the NIR band. The IR source compensates irregular illumination conditions by providing additional illumination when there is not enough ambient light and filling in shadow regions when there are deep cast or attached shadows on the face. Ambient light outside of the NIR band can be removed by an optical filter to further reduce the ambient illumination variation.

A commercially available IEEE1394a camera, Point Grey Research Flea2 with Sony ICX424AL monochrome CCD, was used for imaging. Figure 10 shows the spectral sensitivity characteristics of the CCD. Since the sensitivity of the CCD decreases

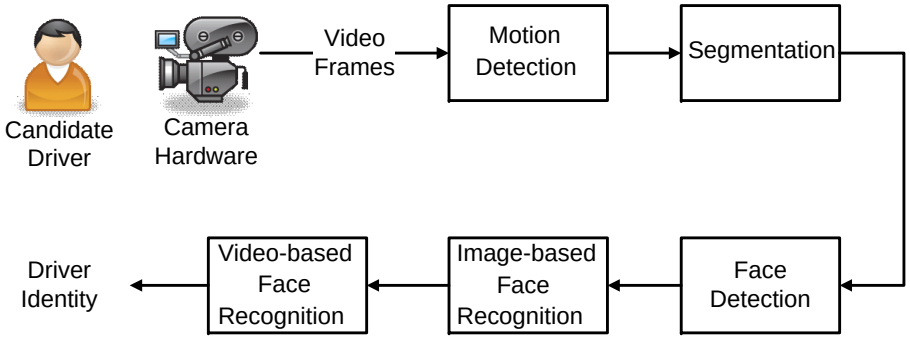


Figure 9: System overview flowchart.

as the wavelength is increased in the NIR band, the minimum wavelength that is invisible to human eyes is preferred. It is commonly reported that humans cannot see beyond 700 nm, but self-conducted visual tests in the laboratory proved the threshold to be approximately 850 nm. Therefore, NIR LEDs with a high output at 880 nm and an approximate half-power bandwidth of 60 nm were selected (Fairchild Semiconductor QED223). The camera lens was fitted with an 880 nm optical band pass filter with a filter response, as shown in Fig. 11. For illumination, a rectangular array of 42 LEDs in seven parallel networks was placed around the camera, as shown in Fig. 12(a). The overall light output from the circuit can be controlled by a DIP switch on the front of the mount to selectively enable any of the seven networks of LEDs. Figure 12(b) shows the experimental setup of the camera and illuminator in a vehicle. Monochrome video is captured at 640x480 pixel resolution. The available I/O ports on the camera are used to drive the LEDs synchronously with the camera shutter.

During video capture, the LED array is pulsed by a control signal from the camera to produce an alternate sequence of the *illuminated* frames (called "I-frames") and *ambient* frames (called "A-frames"). The illuminated and ambient frames are used in image differencing to make face recognition more robust to extreme illumination variations and shadows. Under ideal conditions, with no movement of the face, the *difference* between the illuminated and ambient frame (called "D-frame") will be

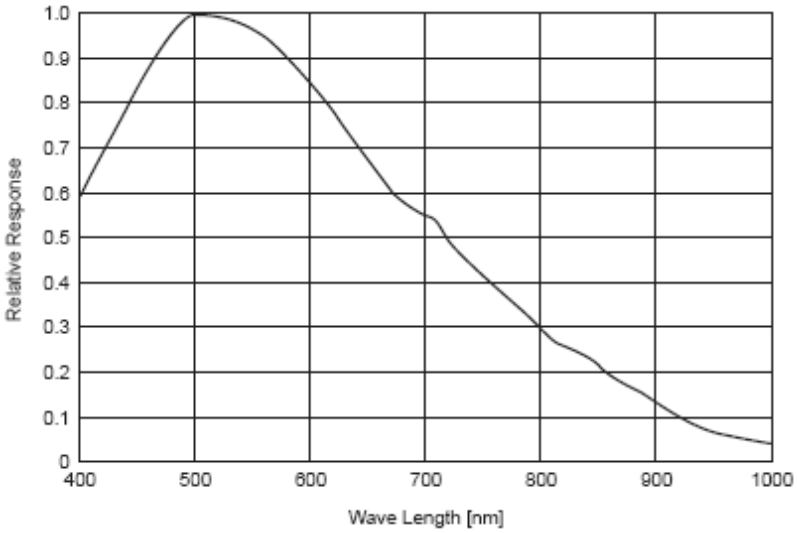


Figure 10: CCD response of the Flea2 camera.

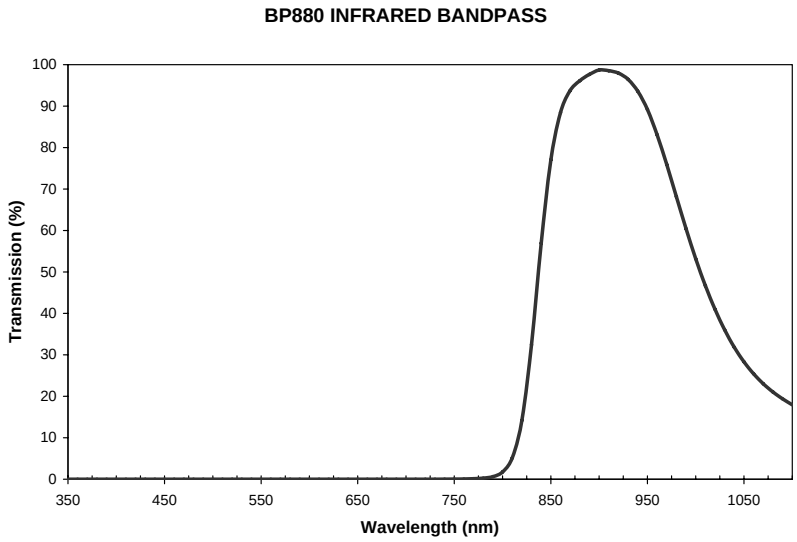
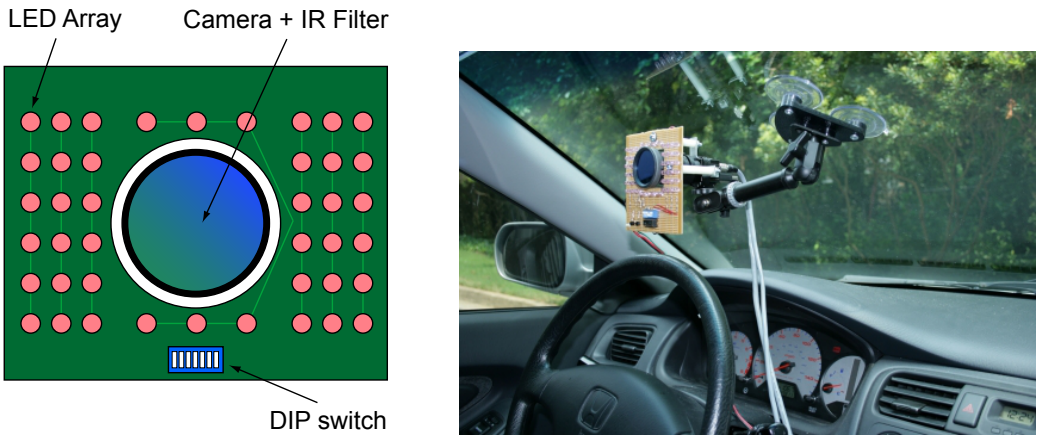


Figure 11: Band pass filter response.



(a) (b)

Figure 12: (a) Illustration of hardware components (b) Camera and illuminator installed in a vehicle.

similar to a night-time image that is captured with the LED array turned on. In other words, the image differencing method will give face images that are illuminated only by the LEDs, thereby removing the unknown and highly varying ambient IR light as well as shadows that are cast by visors, mirrors, and other objects. As a result, the difference frame will have much less illumination variation compared to a face that is illuminated with a combination of outdoor light plus the IR-LED array. With image differencing, since the dynamic range of the difference image will be smaller than that of the original image, a higher bit depth is necessary in the original image. Therefore, 12-bits per pixel were used for image capture. Example frames are shown in Fig. 13. Figure 13(a) shows a ambient frame with a strong shadow that is cast by the car roof and Fig. 13(b) shows an illuminated frame. The difference frame that is formed by a point-wise pixel subtraction of the image in (a) from that in (b) is shown in Fig. 13(c), and Fig. 13(d) shows the same difference image after it has been scaled to increase the dynamic range of the pixels. Note that the difference frame is similar to what we would expect to find at night with the only source of IR illumination coming from the LEDs. Also note that in the difference frame, part of the shirt is very dark. This is due to the fact that the pixel intensities of both the

(a) Ambient frame (b) Illuminated frame (c) difference frame (d) normalized difference frame

Figure 13: Example ambient, illuminated and difference frames are shown. Only the image in (d) was normalized to be visualized in proper contrast.

illuminated frame and ambient frames in this area are saturated and the difference is close to 0. Therefore, saturation should be avoided, if possible, by adjusting the camera exposure. Also, it was observed that gamma correction in the camera should be disabled in order to avoid nonlinear processing prior to image differencing. In the difference frame, negative values were set to 0.

Referring to Fig. 9, the IR video frames from the IR imaging system are processed through several steps to perform face recognition of the driver. First, the image differencing between illuminated and ambient frames is performed. Then, the difference frames are processed to separate the foreground objects, i.e., the candidate driver(s), from the background. Next, the illuminated frames containing detected foreground objects are checked for motion between frames. Still or near-still groups of frames are sent on for further processing, while all others are rejected. The reason for discarding frames that have motion is that when the image-differencing method takes place, even a small amount of motion may severely affect the difference image and make it difficult to perform face recognition. For those frames that have little or no motion, the modified version of the boosted classifier proposed by Viola and Jones [71, 110] is used to find the face of the driver only on the foreground region in the difference

frame. Then, preprocessing is done on the difference frames, such as image transformations and intensity normalization. Finally, linear discriminant analysis (LDA) is performed to determine the identity of the driver for each frame. At the end of some period of time, a decision is then made as to who the driver is. Each of these steps is described in greater detail in the following sections.

4.3 Foreground/Background Segmentation

A foreground/background segmentation method is applied as an initial step toward face localization. Since there is no standard definition of what is classified as foreground or background, application-specific definitions need to be created. For this vehicle personalization application, the area seen by the camera before the driver enters the vehicle is defined as the background, and everything else is classified as foreground.

The image-differencing method enables robust segmentation by applying a threshold to individual pixels. When considering a single frame, e.g., an illuminated or an ambient frame, there is no prior knowledge relating the intensities of the foreground and background areas. Depending on current illumination conditions, background areas may appear brighter than foreground areas or vice versa, making it difficult to achieve appropriate conditions for segmentation. However, when considering a difference frame obtained from consecutive illuminated and ambient frames, there exists prior knowledge about the pixel intensities, i.e., foreground pixels appear brighter than background pixels. Ideally, in a static environment, the difference frame exaggerates the areas that are largely affected by the illumination from the LEDs. The reason for this phenomenon can be explained using the physics of light. It is observed that the intensity of light from the illuminators attenuates at a rate of $\frac{1}{R^2}$, where R is the radial distance from the illuminators to the area being illuminated [79]. Since foreground areas are generally closer to the illumination source than background, they

appear brighter. A threshold method can then be applied to classify pixel intensities as background or foreground accordingly. If the intensity of a pixel is higher than the threshold, then a pixel is classified as a foreground pixel; otherwise, it is classified as a background pixel.

The initial step is to find a threshold value for comparison to each pixel in the difference frame. The range of background pixel intensities can vary depending on the camera settings and interior of the vehicle. But, there is a high probability that the range of the intensities does not change over the length of the video. For this reason, the maximum intensity value of the first difference frame in the video, i.e., the difference frame of the background scene, is a good candidate for the threshold. But the maximum intensity is not robust because it is easily affected by noise generated from the camera, trembling of the vehicle, or outside moving illumination sources. Consequently, 1.1 times the 99.9 percentile is used as the threshold.

The threshold value is used for the foreground/background segmentation of the incoming difference frames from the video. At first, the difference frame is uniformly downsampled by 8 to reduce computation time. A decrease in spatial resolution is acceptable considering that the purpose of the segmentation process is only to find approximate locations of candidate faces. The downsampled difference frame is converted to binary after application of the threshold. A median filter is then applied, followed by dilation and erosion to the binary image. An example showing a typical sequence of images passing through this process is shown in Fig. 14.

In the median filter we used, each output pixel contains the median value from a 3-by-3 neighborhood centered on the corresponding pixel in the thresholded binary image. And the structuring elements for dilation and erosion operations were circular disks of radius 3. For further improvement, we could use one pass of a separable recursive median filter of window width 3 or 5 instead of the non-recursive two-dimensional median filter.

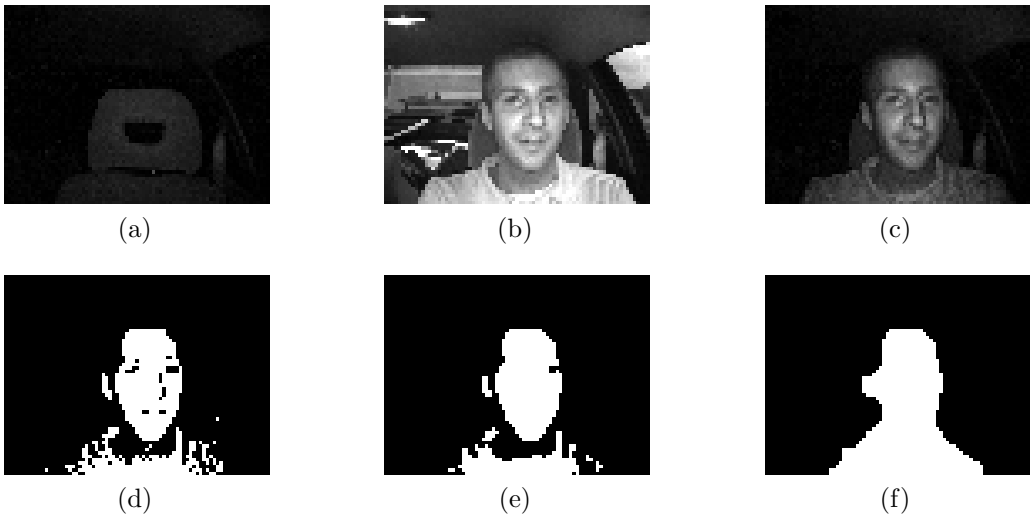


Figure 14: Example images in the processing steps of the foreground/background segmentation are presented. (a) The first difference frame with a background scene (b) The current illuminated frame (c) The current difference frame (d) Thresholding (e) Median filter (f) Dilation and erosion.

In addition, the median filter applied to binary data is a max of mins, or min of maxes, corresponding to a majority logic gate where an AND gate corresponds to a min operator and an OR gate corresponds to a max operator. Erosions and dilations applied to binary images are also maxes and mins over the structuring elements. Therefore, a simpler and faster method can be implemented combining the median, dilation and erosion operations by considering the composition of these three logic operators, which is another logic operator.

4.4 Motion Detection

Although the image-differencing method is used to solve many of the illumination related problems, it is very sensitive to motion between frames. In fact, if motion occurs, the difference image is likely to contain random artifacts; hence, a motion detector is required to determine the existence of motion between frames. Methods used in MPEG-2, MPEG-4, Divx, and other motion-based codecs (compressor-decompressor) serve to objectively quantify motion in a frame set; but these methods have a very

high degree of computational complexity. The goal of the motion detector implemented here is to suggest when to keep or reject frames using simple metrics. The flowchart of motion detection is shown in Fig. 15. Motion detection runs on only the illuminated frames since they are taken with better illumination conditions than ambient frames.

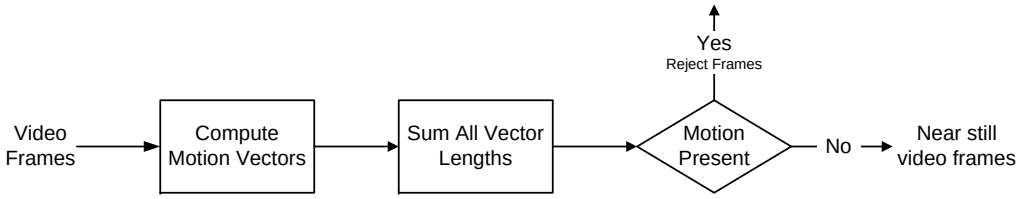


Figure 15: Motion detection flowchart.

Given a frame pair, one frame is chosen as the reference frame, while the other is chosen as the comparison frame. Motion vectors are computed by correlating areas between the reference and comparison frame. As illustrated in Fig. 16, an $M \times N$ block centered at a point (x, y) in the reference frame, referred to as a template block, is correlated with corresponding blocks within a $2M \times 2N$ search window in the comparison frame, which is also centered at point (x, y). The best match is determined by the maximum of the normalized cross correlation [68]:

$$\gamma(u,v) = \frac{\sum_{x,y}[f(x,y) - \bar{f}_{u,v}][t(x-u,y-v) - \bar{t}]}{\sqrt{\sum_{x,y}[f(x,y) - \bar{f}_{u,v}]^2 \sum_{x,y}[t(x-u,y-v) - \bar{t}]^2}} \tag{1}$$

where f is the image within the $2M \times 2N$ search window, t is the template block, $\bar{t}$ is the mean of the template, and $\bar{f}$ is the mean of f under the template. For the purpose of this work, $M = N = 30$ pixels. The distance between the location of the largest correlation coefficient and the center of the search window defines both the direction and amount of translation for the given block.

Once the driver is settled inside the vehicle, it is safe to assume that the driver's face will eventually be located near the center of the frame. Therefore, the normalized

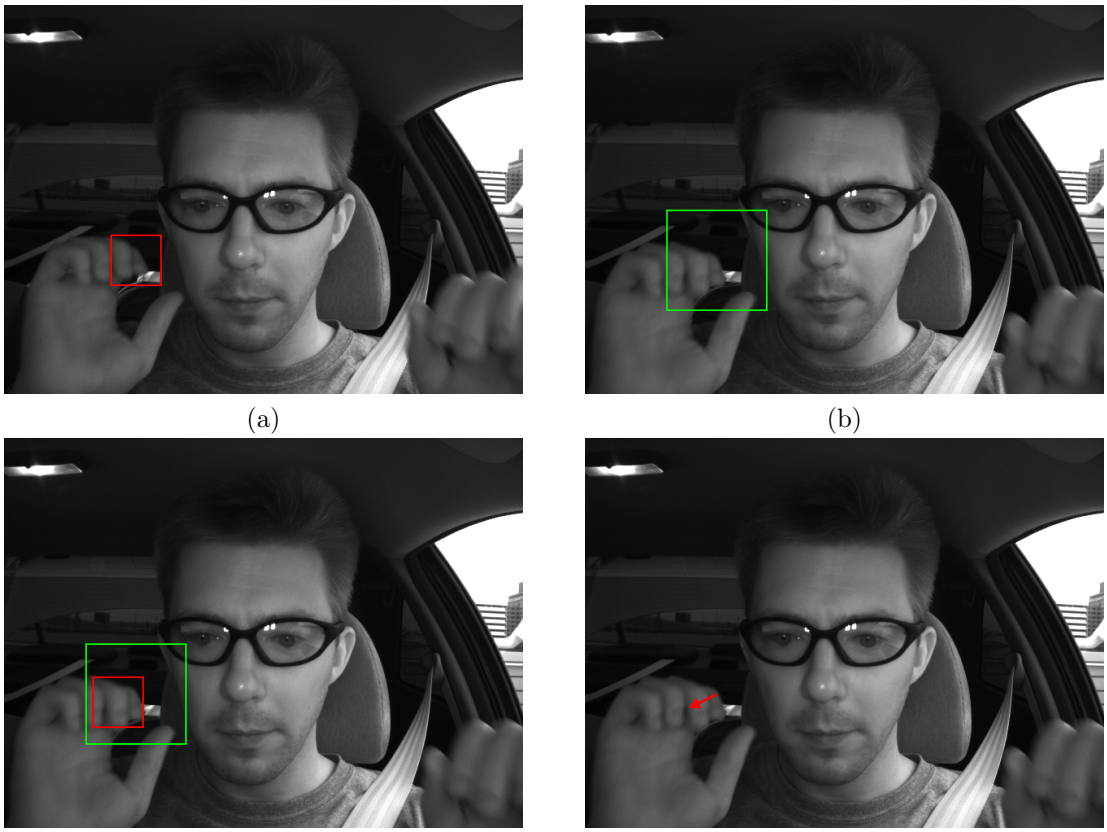
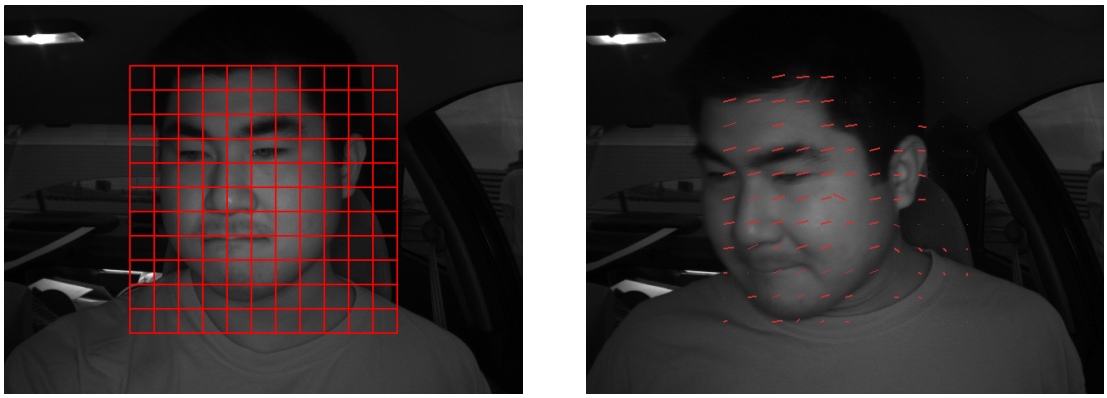


Figure 16: Finding a motion vector. (a) Reference frame with the template block (b) Comparison frame with the corresponding search window (c) Best match between the template block and the search window (d) Computed motion vector.

cross correlation given in Eq. (1) will be focused on finding motion vectors near the center of the frame. As a result, motion vectors are computed within the area near the center of the frame, as illustrated in Fig. 17, where the grid represents the locations of the template blocks. Each block is correlated with its corresponding search window to generate motion vectors. The collection of these motion vectors is referred to as the motion field. An 11×11 grid of template blocks is used in this implementation of the motion detector.

Once the motion field has been determined, the lengths of all the vectors are summed together for use as a metric, β, to determine the amount of motion within



(a) (b)

Figure 17: Motion fields. (a) A grid of template blocks (b) A motion field generated by computing motion vectors for each template block in (a).

the frame.

$$\beta = \sum_{i \in MV} \ell(i) \tag{2}$$

where MV is the collection of motion vectors and $\ell(i)$ is the computed length of the i^{th} motion vector using the L^2 norm.

The training set for the motion detector consisted of 1,000 frame pairs that contained little to no motion within each pair. The magnitude of the motion field for each pair was computed using Eq. (2) and its distribution over the training set is shown in Fig. 18. The exponential distribution shown in Fig. 19 was estimated based on the mean of the distribution of the training set. The formula for the exponential distribution is:

$$p(X) = \frac{1}{\mu} e^{\frac{x}{\mu}} \tag{3}$$

where μ is the sample mean of the data set. After visually inspecting the results on a small subset of the test data, the value corresponding to the 80% percentile (τ) in the CDF (cumulative distribution function) for the exponential distribution was chosen as a threshold to classify the presence of motion between the frames. The empirical values used are:

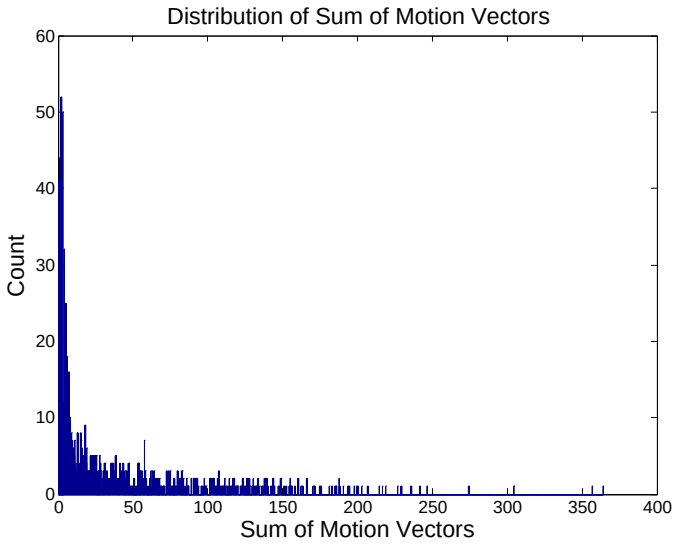


Figure 18: Actual distribution of the sum of lengths of motion vectors.

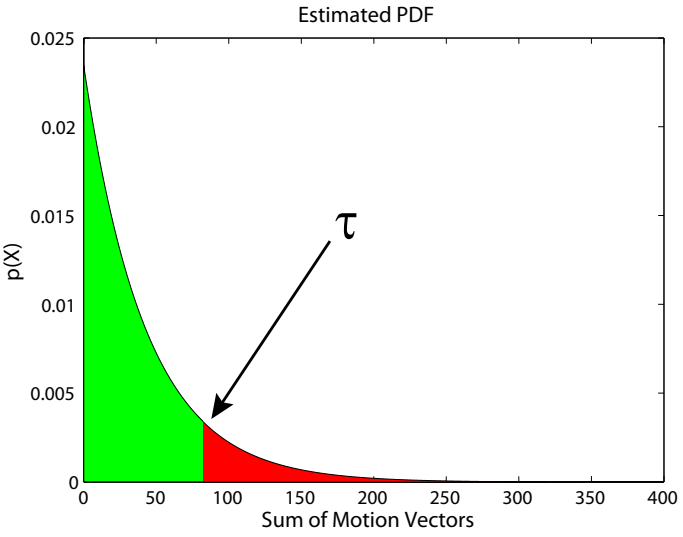


Figure 19: Estimated exponential distribution based on the sample data.

Near Still	if $\beta \leq \tau$
Unknown or motion present	if $\beta > \tau$

If the motion field is classified as still, the corresponding reference frame is forwarded to the face detector. Otherwise, it is rejected.

4.5 Face Detection

The face detector implemented in this system finds the face of the driver using the difference frame. The search range of the face detector was narrowed down by the foreground/background segmentation. Face detection is then performed using the modified version of the boosted classifier proposed by Viola and Jones [71,111]. After detection, the candidate face region is forwarded to the face recognition stage for further processing. With the help of the image-differencing method, this face detector is able to produce good results even in conditions of low or non-uniform illumination. The fowchart of face detection is shown in Fig. 20.

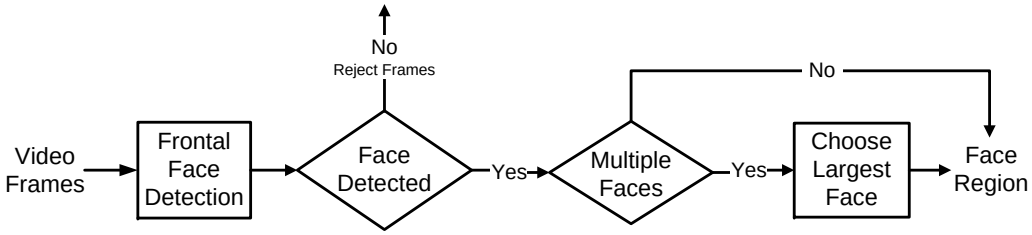


Figure 20: Face detection flowchart.

Figure 21 shows face detection results on an example ambient, illuminated, and difference frame triplet. The face detector failed to detect the frontal face in the ambient frame, and while the face in the illuminated frame was detected, the detected face size was smaller than the actual face. On the other hand, the face detector was successful when the difference frame was used. It was demonstrated experimentally, in fact, that higher face detection rates are achieved using difference frames compared

(a) (b) (c)

Figure 21: Face detection using Viola-Jones face detector [111]. In the figures above, face detection is performed on (a) an ambient frame with no success (b) an illuminated frame (c) a difference frame. In (b) and (c), the eyes are located based on size and orientation of the detected face region.

with ambient and illuminated frames. Specifically, using 4302 ambient and illuminated frame pairs with frontal faces from 10 video sequences, face detection results of ambient, illuminated, and difference frames were manually determined. The error rates are reported in Table 5. Here, a missed detection means that the face detector fails to detect a frontal face, and a false detection means that the detected region is not a face. All of the missed detections and false detections on the 4302 frames were counted and the error rates reported. As we see from the table, the difference frames have the lowest missed detection rates as well as the lowest false detection rates. Not surprisingly, the ambient frames have the highest missed detection rates and the highest false detection rates. From these results it is evident that LED illumination improves the face detection rate and the image-differencing method further improves the performance. When a single face is detected in the difference frame, it is cropped and passed to the face recognition stage for further processing. If multiple faces are detected, the largest face region is considered to be the face of interest and is passed on to the next stage.

Table 5: Error rates of face detection for ambient, illuminated and difference frames.

Types of frames	Ambient	Illuminated	Difference
Missed detection rate	0.1104	0.0463	0.0140
False detection rate	0.0122	0.0101	0.0094

4.6 Face Recognition

Among the many approaches used for face recognition, appearance-based subspace methods are among the most popular, primarily due to their success in controlled or semi-controlled environments and their computational simplicity. Two popular appearance-based subspace analysis methods are Eigenface and Fisherface. Eigenface is equivalent to principal component analysis (PCA) [108] and Fisherface is a combination of PCA and linear discriminant analysis (LDA) [13]. PCA performs dimension reduction by finding a set of representative projection vectors such that a projection of a sample set retains most of the information of the original sample set. On the other hand, LDA uses the class information to find a set of vectors that maximize the between-class scatter while minimizing the within-class scatter [33]. Fisherface methods embody face recognition systems by using personal identities as class labels. Prior research [13] shows that the Fisherface algorithms display higher accuracy in recognizing faces under variable illumination conditions compared to the Eigenface algorithm. If the training data set contains frontal face frames under various illumination conditions for each subject, the subspace corresponding to the illumination variation is minimized from the LDA space because LDA has the ability to maximize the ratio of between-class scatter to within-class scatter. The success of LDA under variable illumination is based on the fact that frames of a Lambertian surface under varying illumination lie in a 3D linear subspace of the image space when cast shadows are ignored [94]. The Lambertian surface images approximately lie in a 9D linear subspace when attached shadows are taken into consideration [8].

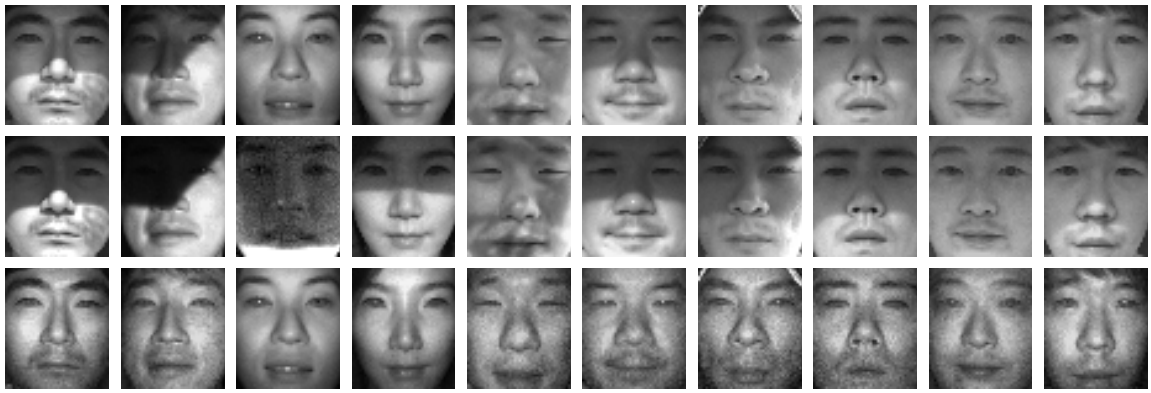
All the images in the training and test data sets go through a preprocessing stage.

This stage includes image registration, masking, and pixel value normalization. Image registration is performed by rotating and scaling the face images in an attempt to relocate the eyes to predefined positions. For the training set, eye locations are manually found and the face images are rotated and scaled according to the eye locations. For the probe and gallery images in the test set, face regions located by the face detector are registered without rotating. Masking is then used to consider only the pixels inside the face boundary. The pixel values are then normalized to have zero mean and a unit standard deviation. Before normalization, the pixel intensities are clipped at a level corresponding to the 99% percentile of the largest intensity value in an effort to reduce the noise from specular reflectance. Specular reflectance depends on the position of the illumination source relative to the face and should be prevented in order to reduce dependency on illumination source factors.

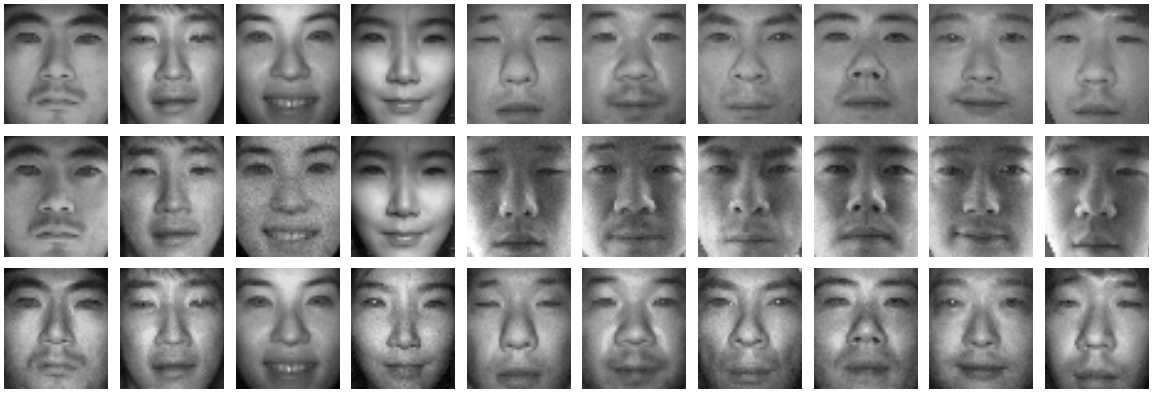
Two experiments were performed to demonstrate the effectiveness of the image differencing on face recognition with Fisherface in highly variable illumination conditions. In the first experiment, face recognition rates are reported on face frames with and without shadows cast by other objects, such as a car roof or a sun visor, to show robustness under various shadow conditions. In the second experiment, face recognition rates are reported on face images illuminated with three different LED configurations: 6, 18, and 24 LEDs, to show robustness to changes in the intensity of the active illumination source.

For the first experiment, near-frontal face frames of 40 subjects are manually selected from the video stream (automatic detection of frontal face images is a subject for the second phase of the development of this system). The frames are divided into two subsets: face frames with shadows and face frames without shadows. The shadow subset consists of 1885 triplets of face frames, where each triplet is made up of an illuminated frame, an ambient frame, and a difference frame. Similarly, the no-shadow subset consists of 1786 frame triplets. The total number of triplets for each subject

varies. The eyes are located manually for each frame in the triplet sharing the same eye locations. Face frames are registered, masked, and normalized in the preprocessing step. Figure 22 (a) shows sample face triplets in the shadow subset and Fig. 22 (b) shows sample face frame triplets in the no-shadow subset. In Fig. 22, each column is a triplet corresponding to a subject. The first row corresponds to illuminated frames, the second row corresponds to ambient frames, and the third row corresponds to the difference frames. For every sample frame, it is evident that most of the shadow is removed in its corresponding difference frame. Face recognition is then conducted multiple times with randomly chosen frames and the average recognition rates are reported for the cases of ambient, illuminated, and difference frames. The procedure for this experiment is as follows: one frame for each of the 40 subjects is randomly selected as a gallery frame from a particular subset. Similarly, another frame for each of the 40 subjects is randomly selected as a probe frame from a particular subset. Hence, the gallery and probe sets have 40 frames each referring to the 40 subjects. Since there are two subsets (shadow and no-shadow) available for selecting the gallery and probe frames, there are effectively four ways to collectively generate gallery and probe sets, as shown in the first and second columns of the subtables in Table 6. 10,000 iterations of Fisherface recognition are performed on each of the four permutations of the gallery and probe sets. After every iteration, a new gallery and probe set is generated by randomly choosing frames from the subsets. The average recognition rate of each permutation is reported in Table 6. The entire experiment is repeated for different choices of the training data sets. The CMU PIE database [99] and CBSR NIR database [69] are selected as candidate training data sets. The CMU PIE database consists of more than 40,000 frames of 68 subjects from various ethnic groups and also includes variations in both pose and illumination. Thirteen Sony DXC 9000 (3 CCD, progressive scan) cameras with gain and gamma correction disabled were used for data acquisition. The CBSR NIR database has 3,940 NIR face frames of 197



(a)



(b)

Figure 22: In (a) and (b), each column is a face frame triplet corresponding to a subject. Each row corresponds to a type of frame; Row 1: Illuminated frames, Row 2: Ambient frames, Row 3: Difference frame obtained by applying the image differencing method on Rows 1 and 2. Two groups of face triplets are shown above, (a) Shadow on the face of the subject, (b) No shadow on the face of the driver.

people of Asian ethnicity taken with NIR illumination fixed at 850 nm wavelength. The first candidate training set was built using only frontal face frames from the CMU PIE database, which accounted to approximately 108 frames per subject or 7,372 frames in total. The second candidate training set was built using face frames without glasses from the CBSR NIR database, which amounted to approximately 17 frames per subject or 3,329 frames in total.

The two subtables in Table 6 show that the image-differencing method enables a face recognition system that is robust to illumination variations. For the ambient frames, the face recognition rate is high when the probe set and gallery set is built from

Table 6: Face recognition experiment results using a variety of face images with and without shadows.

Gallery Set	Probe Set	Type of Frames		
		Ambient	Illuminated	Difference
No shadow	No shadow	0.9987	1.0000	0.9987
Shadow	Shadow	0.8261	0.8499	0.9729
No shadow	Shadow	0.2656	0.4388	0.8933
Shadow	No shadow	0.2739	0.5901	0.9537

(a) Face recognition rates using CBSR NIR database for training.

Gallery Set	Probe Set	Type of Frames		
		Ambient	Illuminated	Difference
no shadow	no shadow	0.9993	0.9987	0.9965
shadow	shadow	0.7998	0.8313	0.9443
no shadow	shadow	0.3361	0.3548	0.8752
shadow	no shadow	0.2602	0.4259	0.8261

(b) Face recognition rates using CMU PIE database for training.

the same subset, i.e., the shadow subset or the no-shadow subset. The face recognition rate is low when the probe and gallery set are built from different subsets. It is observed that LED illumination itself improves recognition rates from the fact that recognition rates of illuminated frames are higher than those of the ambient frames for all cases. The recognition result of difference frames shows significant improvement over illuminated frames especially when comparing shadow and no-shadow frames. Training with the CBSR NIR database showed better results than training with the CMU PIE database because the CBSR NIR database is made up of NIR frames and the CMU PIE database is made up of visible light frames.

The difference in the face recognition rates is further analyzed using the one-way analysis of variance (ANOVA). The objective of the ANOVA test is to verify that the mean face recognition rates achieved when using the illuminated, the ambient, and

difference frames are significantly different at the 0.05 confidence interval. As shown in Table 6, each of the four permuations output three recognition rates corresponding to the different types of frames: the illuminated, the ambient, and the difference frame. To prepare the data for computing ANOVA, the recognition rates of the four permutations are averaged for each type of frame over all experiments. Table 7 (a) shows the resultant ANOVA table when CBSR NIR database is used for training and Table 7 (b) shows the resultant ANOVA table when CMU PIE database is used for training. Observing the results in Table 7 (a) and (b), it is evident that the H_0 hypothesis can be easily rejected because the mean recognition rates recorded when using the illuminated, the ambient, and the image differencing are in fact significantly different. This is also in accordance with the result in Table 7 (a) and (b) that 'Prob>F' is less than 0.05, and, as a consequence, rejects the hypothesis.

Table 7: Analysis of variance for experiment 1.

Source	SS	df	MS	F	Prob>F
Columns	681.215	2	340.607	276435.38	0
Error	363961	29997	0.001		
Total	718.175	29999			

(a) ANOVA Table for CBSR NIR database

Source	SS	df	MS	F	Prob>F
Columns	555.102	2	277.551	206053.16	0
Error	40.406	29997	0.001		
Total	595.507	29999			

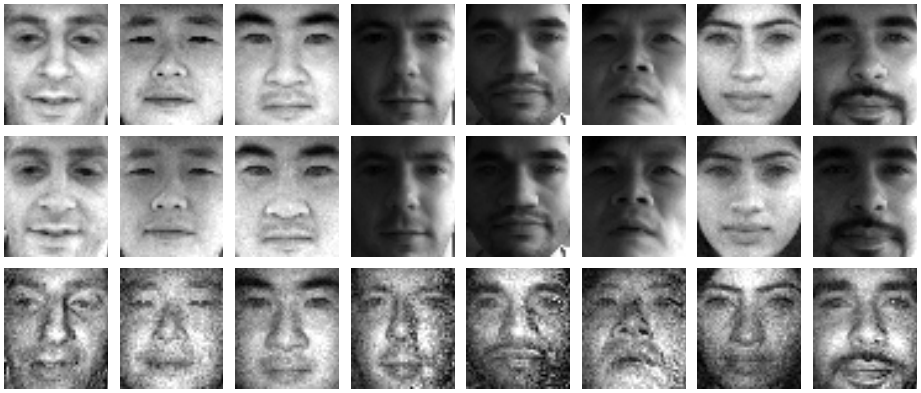
(b) ANOVA Table for CMU PIE database

The difference in the recognition rates is further analyzed using the 2-way Analysis of Variance (ANOVA). In the 2-way ANOVA, two tests are performed; the first test is a 'column-wise' test and the second test is a 'row-wise' test. In the column-wise test, the recognition rates for a particular experiment are tested for consistency between

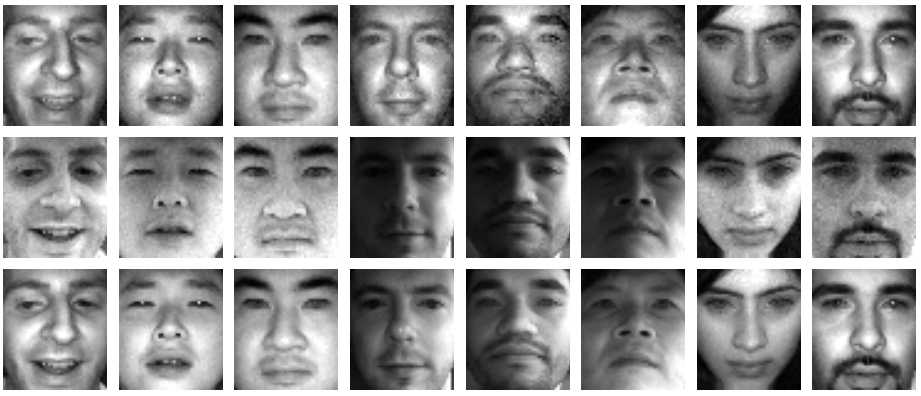
the two databases, and, in the row-wise test, the consistency in the recognition rates for the different experiments is tested. The 2-way ANOVA is used to verify the rejection of the H_0 hypothesis at the 0.05 confidence level across both, the choices of the different experiment setups (row-wise test), and, the choices of the training databases (column-wise test). Table 7 (a) shows the data used for computing ANOVA and Table 7 (b) shows the results of experiment using ANOVA. From the results, it is evident that the H_0 hypothesis can be rejected for the row-wise test because the mean recognition rates recorded for the various experiment setups are significantly different at a 0.05 confidence interval. This result is further supported using Tukey's procedure (the T-Method) [29]. It is also evident that the H_0 hypothesis cannot be rejected at the 0.05 confidence interval for the column-wise test. This implies that the differences in the recognition rates achieved by running the various experiments are more significant that the differences achieved by running the same experiment using different training databases.

The second experiment follows a procedure similar to the first experiment. The test frames are divided into three subsets corresponding to three different LED configurations: 6, 18, and 24 LEDs. The illumination variations in the test set are also increased by including horizontally mirrored frames in each subset. With the mirrored frames included, the six LED subset now consists of 2338 frame triplets, the 18 LED subset consists of 2236 frame triplets, and the 24 LED subset consists of 2210 frames triplets. Figure 23 shows sample face frames in the 6, 18, and 24 LED subsets. For every sample frame, it is evident that most of the shadow is removed in its corresponding difference frame.

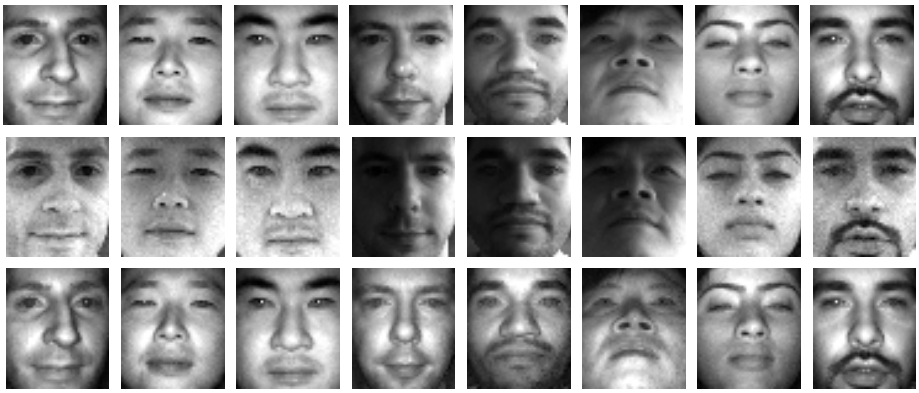
Face recognition is then conducted multiple times with randomly chosen frames, and the average recognition rates are reported for the cases of ambient, illuminated, and difference frames. The procedure for this is as follows: one frame for each of the 10 subjects is randomly selected as a gallery frame from a particular subset.



(a)



(b)



(c)

Figure 23: In (a), (b) and (c), each column is a face frame triplet corresponding to a subject. Each row corresponds to a type of frame; Row 1: illuminated frames, Row 2: ambient frames, Row 3: difference frame obtained by applying the image differencing method on Rows 1 and 2. Three groups of face triplets are shown above, (a) Face illuminated with 6 LEDs, (b) Face illuminated with 18 LEDs, and (c) Face illuminated with 24 LEDs.

Similarly, another frame for each of the 40 subjects is randomly selected as a probe frame from a particular subset. Hence, the gallery and probe sets have 10 frames, each referring to the 40 subjects. Since there are three subsets (6, 18, and 24 LED configuration) available for selecting the gallery and probe sets, there are effectively nine ways to collectively generate gallery and probe sets, as shown in first two columns of the subtables of Table 8. Table 8 (a) reports average recognition rates when the CBSR NIR database is used to build the training set and Table 8 (b) reports average recognition rates when the CMU PIE database is used to build the training set. For both cases, the image differencing shows an increase in the recognition rates compared to the illuminated and ambient frames. There does not appear to be a relationship between active illumination intensities and recognition rates. This result is surprising considering that there appears to be a significant amount of noise present in the 6 LED subset, but the recognition performance seems to remain unaffected.

The difference in the face recognition rates is further analyzed using the 1-way Analysis of Variance (ANOVA). The objective of the ANOVA test is to verify that the mean face recognition rates achieved when using the illuminated, the ambient and the image-differencing method are significantly different at the 0.05 confidence interval. As shown in Table 6, each of the 4 permutations output 3 recognition rates corresponding to the different types of frames; the illuminated, the ambient and the Difference frame. To prepare the data for computing ANOVA, the recognition rates of the 4 permutations are averaged for each type of frame over all experiments. Table 7 (a) shows the resultant ANOVA table when CBSR NIR database is used for training and Table 7 (b) shows the resultant ANOVA table when CMU PIE database is used for training. Observing the results in Table 7 (a) and (b), it is evident that the H_0 hypothesis can be easily rejected because the mean recognition rates recorded when using the illuminated, ambient and the image-differencing method are in fact significantly different. This is also in accordance with the result in Tables 7 (a) and

Table 8: Face recognition experiment results using face images with a variety of active illumination settings.

Gallery Set	Probe Set	Type of Frames		
		Ambient	Illuminated	Difference
6 LEDs	6 LEDs	0.7225	0.7753	0.8989
6 LEDs	18 LEDs	0.6955	0.6825	0.8597
6 LEDs	24 LEDs	0.6824	0.6725	0.8870
18 LEDs	6 LEDs	0.6801	0.7480	0.8733
18 LEDs	18 LEDs	0.6856	0.7547	0.8582
18 LEDs	24 LEDs	0.6787	0.7617	0.8760
24 LEDs	6 LEDs	0.6885	0.7493	0.8866
24 LEDs	18 LEDs	0.6785	0.7690	0.8868
24 LEDs	24 LEDs	0.6935	0.7745	0.9035

(a) Face recognition rates using CBSR NIR database for training.

Gallery Set	Probe Set	Type of Frames		
		Ambient	Illuminated	Difference
6 LEDs	6 LEDs	0.8454	0.8659	0.9619
6 LEDs	18 LEDs	0.8282	0.7686	0.9373
6 LEDs	24 LEDs	0.8258	0.8073	0.9289
18 LEDs	6 LEDs	0.8294	0.8489	0.9239
18 LEDs	18 LEDs	0.8347	0.9032	0.9330
18 LEDs	24 LEDs	0.8342	0.9061	0.9151
24 LEDs	6 LEDs	0.8222	0.8546	0.9492
24 LEDs	18 LEDs	0.8131	0.9068	0.9477
24 LEDs	24 LEDs	0.8279	0.9472	0.9443

(b) Face recognition rates using CMU PIE database for training.

(b) that 'Prob>F' is less than 0.05, and, as a consequence, rejects the hypothesis.

The differences in the recognition rates for the second experiment are also analyzed using the one-way ANOVA. On the same lines as the first experiment, the one-way ANOVA is used to verify the rejection of the H_0 hypothesis at the 0.05 confidence

interval. As shown in Table 8, each of the nine permutations output three recognition rates corresponding to the different types of frames: the illuminated, the ambient, and the difference frame. The data for ANOVA is prepared by computing the average recognition rates of the nine permutations for each of the three types of frames for all experiments. Table 9 (a) shows the resultant ANOVA table when the CBSR NIR database is used for training and Table 9 (b) shows the resultant ANOVA table when the CMU PIE database is used for training. The results in Table 9 (a) and (b) clearly show that the H_0 hypothesis can be rejected again because the mean recognition rates recorded when using the illuminated, ambient and difference frames are again significantly different, and, as a consequence, 'Prob>F' is less than 0.05.

Table 9: Analysis of variance for experiment 2.

Source	SS	df	MS	F	Prob>F
Columns	3.134	2	1.56678	285.9	0
Error	164.387	29997	0.00548		
Total	167.52	29999			

(a) ANOVA Table for CBSR NIR database

Source	SS	df	MS	F	Prob>F
Columns	1.056	2	0.52779	131.5	0
Error	120.4	29997	0.00401		
Total	121.456	29999			

(b) ANOVA Table for CMU PIE database

The aftermath of the two experiments above shows that not only are the average face recognition rates higher when image differencing is used in contrast to the illuminated and ambient frames, but the average rates are also significantly different as proved by ANOVA. This goes to show that the image differencing plays an important role in building this robust face recognition-based personalization system.

4.7 Pose Clustering

The last module is the video-based face recognition with pose clustering. The gallery videos are processed through the modules explained above: image differencing, foreground segmentation, motion detection, face detection, motion interpolation, preprocessing, and projection onto the LDA subspace. Then the projected LDA features are clustered into 10 different poses. Then, the cluster centroids are stored in the system and represent different poses of the driver which improve the robustness of the system to the pose variation. A probe video of an unknown driver is then fed to the system and also processed through the same modules. Then the identity of each detected frontal face image is decided by the nearest neighbor algorithm. The identity of the gallery cluster centroid that has the minimum distance to the probe image in the LDA subspace determines the identity of the probe image. The final decision on the probe video is made by majority voting of the image-based recognition results. Figure 24 shows an example of the pose clustering.

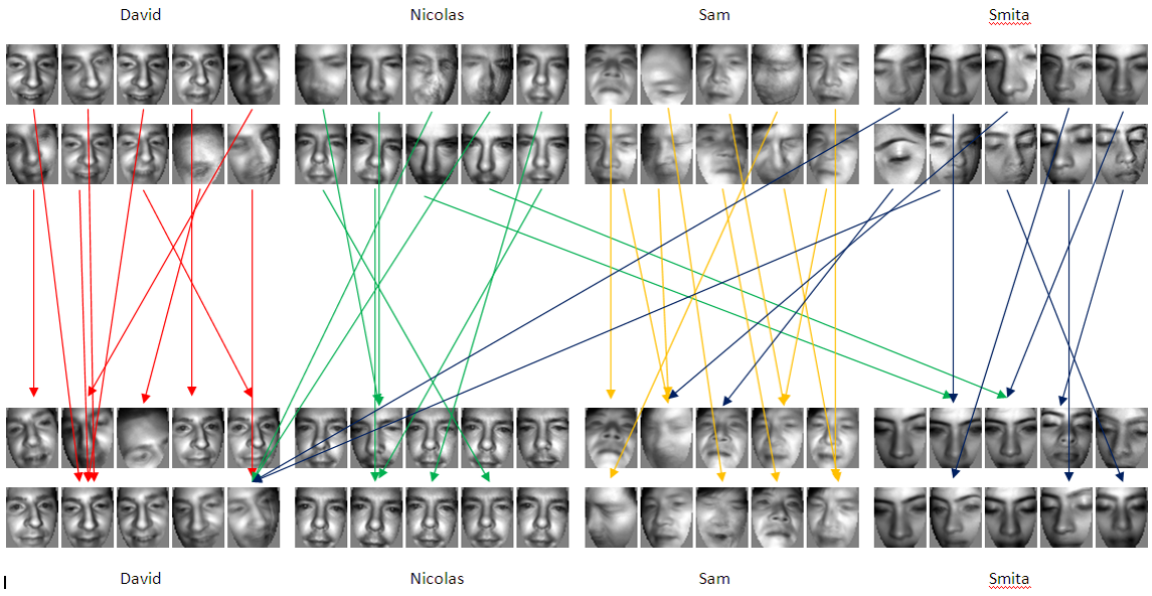


Figure 24: Clustering example.

4.8 Performance Evaluations

A NIR face video database of outdoor vehicular scenario were collected with the camera and LEDs shown in Fig. 12. Videos were recorded in extremely challenging outdoor illumination and shadowing conditions including various weather condition, direct sunlight and cast shadows. The number of LEDs turned on for illuminated frames were varied including 6, 12, 18 and 24 LEDs. The database is composed of 195 videos of 40 drivers in vehicles. There is one gallery video for each driver in the database and the rest of videos are probe videos. For gallery videos, drivers were asked to move their faces up, down, to the left, and to the right slowly to record various head pose variation. To simulate real vehicular scenario, probe videos were taken while drivers get into the car and start engine naturally without any specific user cooperation. The length of each video is approximately from 20 to 30 seconds long. Videos were originally saved in PGM file format and the total size of the database is 197 GB. Illuminated frames and ambient frames were encoded in AVI video file format separately and the total size of encoded videos is 7.8 GB. In real system, procedure similar to the gallery videos in the database will be performed. If the driver selects the new user registration option, the system will instruct the driver to move the face up, down, to the left, and to the right slowly, record the video, and register the user with recorded face images.

The performance of the end-to-end system was evaluated with the NIR face video database. Figures 22 and 23 show sample face images of the 40 subjects in the data set undergoing different illumination variations, for example, non-uniform shadow, intensity variations from led configurations, and so on. For the experiment setup, four subjects from the data set of 40 are randomly selected to constitute a family, while a fifth subject is randomly selected as an intruder. The gallery set for this experiment contains a gallery video for each member of the family. The probe set contains a randomly selected probe video of the same family members along with a randomly

selected probe video of an unknown person posing as the intruder. As a result, the gallery set contains four videos and the probe set contains five videos. Since there are 40 subjects in the video data set, the number of combinations in generating a four-person gallery set and a corresponding five-person probe set is enormous. Due to time constraints, experiments were performed on only 200 of these random combinations. Each combination constituted to a new experiment with a new gallery and probe set. The 200 experiments were further divided into two groups of 100 experiments each. The first group of experiments was considered 'inter-ethnic' experiments and the second group of experiments was considered 'intra-ethnic' experiments. In the inter-ethnic experiment, the family consists of two subjects of Asian ethnicity and two subjects of other ethnicities. In the intra-ethnic experiments, the family consists of four subjects of the same ethnicity, which is the Asian ethnicity for this project. Table 10 shows the recognition rates of the family members and the intruder for the two groups of 100 experiments. The results are reported for the three different frame types: ambient, illuminated, and difference frames. Table 10 summarizes the experimental design and parameter choices.

Table 10: Experimental design and parameter choices.

Experimental setting	method/value
selection of subjects in probe set	random
number of subjects in probe set	4
selection of gallery videos	fixed
selection of subjects in gallery set	subjects in probe set + random intruder
number of subjects in gallery set	5
selection of probe videos	random
number of experiments	200

The results in Table 11 clearly show that the difference frame acquired by using the image differencing enables a more accurate subject recognition system compared to the cases of the illuminated and the ambient frames for both family types. The results for detecting an intruder do not seem to be noticeably affected by the different

types of frames. Since the difference frame is one of the fundamental building blocks of this end-to-end system, further analysis on the results in Table 10 is performed only for the case of the difference frame using one-way ANOVA. For the inter-ethnic experiment, the recognition rates of the family members are averaged for each of the 100 experiments to prepare the data for ANOVA. Since the recognition rates for each family member are nearly identical, the averaged rate is not very different from that of each member, making it appropriate for use with ANOVA. The objective of the one-way test here is to verify that the mean recognition rates of the family members and the intruders are significantly different, implying the rejection of the H_0 hypothesis at the specified confidence interval. The results of the one-way test are reported in Table 12. A similar ANOVA test is also performed for the intra-ethnic experiment and the results are reported in Table 13.

Table 11: Recognition rates of 200 experimental unit iterations are reported for all possible combinations of family types, probe types and frame types.

Family Type	Probe Type	Type of Frames		
		Ambient	Illuminated	Difference
Inter-ethnicity	Family member	0.6575	0.55	0.965
	Intruder	0.07	0.03	0.12
Intra-ethnicity	Family member	0.325	0.2825	0.905
	Intruder	0.18	0.02	0.11

Table 12: Analysis of variance on the recognition results of the inter-ethnic family.

Source	SS	df	MS	F	Prob>F
Columns	35.7013	1	35.7013	642.87	0
Error	11.3125	198	0.0571		
Total	47.0138	199			

Since the 'Prob>F' column in both Table 12 and 13 is less than 0.05, the H_0 hypothesis can be easily rejected at the 0.05 confidence interval, supporting the results in Table 10 that the mean recognition rates for the members of both family types

Table 13: analysis of variance on the recognition results of the intra-ethnic family.

Source	SS	df	MS	F	Prob>F
Columns	31.6012	1	31.6012	537.66	0
Error	11.637	198	0.0588		
Total	43.2387	199			

are significantly different from the intruders. As a result, the implemented system is capable of easily recognizing the members in its data set while also being able to distinguish them from potential intruders or subjects unknown to the data set.

The performance of the end-to-end video-based face recognition system with pose clustering is evaluated with the video dataset. In this case, all 40 subjects in the database are included in the gallery set and the probe set to evaluate more accurate face recognition rates. The gallery set contains the 40 head rotation gallery videos of 40 subjects. And the probe set contains randomly chosen probe videos of 40 subjects. Due to time constraints, experiments were performed on 200 of these random combinations. Table 14 summarizes the experimental design and parameter choices.

Table 14: Experimental design and parameter choices.

Experimental setting	method/value
selection of subjects in probe set	fixed
number of subjects in probe set	40
selection of gallery videos	fixed
selection of subjects in gallery set	fixed
number of subjects in gallery set	40
selection of probe videos	random
number of experiments	200

Table 15 shows the experimental results of image- and video-based recognition of the proposed end-to-end system compared with the state-of-the-art illumination invariant feature methods for face recognition including Local Binary Patterns (LBP) [51] and gradientfaces [126]. The experimental results show that the proposed method

outperforms both illumination invariant feature methods. Additionally, the illumination invariant features methods can be applied to the output images of the proposed system to further increase face recognition rates.

Table 15: Face recognition rate comparison.

Methods	Image-based	Video-based
LBP	0.691	0.772
Gradientfaces	0.719	0.823
Proposed	0.881	0.934
Proposed + LBP	0.923	0.961
Proposed + Gradientfaces	0.929	0.966

Finally, the system performance without the foreground/background segmentation including the morphological filter is evaluated. As described in Section 4.3, the foreground/background segmentation reduces the search range of the following face detection module for computational efficiency and doesn't affect the final face recognition rate. Experimental test of 10 example videos shows that the average time reduction for each frame is 0.43 ms which is not significant in the rate of 30 fps. The performance of the end-to-end system without the foreground/background segmentation is evaluated and the face recognition rate remains the same as in Table 15. Therefore, the foreground/background segmentation can be removed safely without affecting the system performance.

4.9 Conclusion

This chapter presents a system of practical technologies to implement an illumination-robust, consumer-grade biometric system based on face recognition to be used in the automotive market and ultimately result in very low-cost, easy-to-deploy solutions enabling a wide variety of applications that can benefit from personalization. The image differencing method with an active illumination control was presented and proved to produce images independent of the ambient illumination. Foreground/background

segmentation, motion detection, face detection and face recognition modules in the end-to-end system were presented and the performance improvements on the modules with the use of difference frames were shown with test results. Test results on the end-to-end system with test videos taken in the extremely challenging illumination and shadowing conditions demonstrate highly accurate face recognition in the vehicular application scenario.

The current solution uses only frontal face images and does not deal with occlusion problems explicitly. The system could be improved by considering face images with other face poses and also considering occlusion problems. More advanced decision methods can further improve the accuracy in the decision on the subject identification.

CHAPTER V

IMAGE ALIGNMENT AND IMAGE FUSION FOR ACTIVE NEAR INFRARED IMAGE DIFFERENCING

5.1 Overview

We have shown that the NIR image differencing introduced in the previous chapter successfully removes the ambient illumination effect and improves the face recognition rate. However, there are two critical limitations in the active image differencing. First, the methods assume that there is no motion between I-frames and the respective A-frames. Therefore, any motion between frames can result in severe artifacts in the D-frame. Second, the differencing operation tends to amplify the effect of sensor noise and the face recognition performance is deteriorated by the noise when the gallery or probe images are taken under strong ambient illumination, especially with direct sunlight.

For the motion problem, Zou, *et al.*, proposed a motion interpolation approach [135]. They captured an I-frame and an A-frame alternately. Then the motion between two I-frames is estimated using the optical flow estimation method by Black and Anandan [15], and a virtual I-frame is computed using the motion estimate to interpolate between I-frames. The difference frame between this virtual I-frame, and the A-frame captured between the above two I-frames is used for recognition. However, since the motion between an I-frame and an A-frame is estimated from the motion between two I-frames, there is a possibility of errors when the face is not moving at the same speed in the same direction. Additionally, interpolation on I-frames will smooth out the details in I-frame and also in resulting D-frame. And the optical flow method by Black and Anandan is slow and often not viable in real-time applications.

For the outdoor noise problem, the authors in [120] proposed an enhanced NIR imaging where a narrow band NIR laser generator is used to increase the relative NIR light intensity of active NIR illumination to sunlight. However, they reported that their method is not perfect when the ambient NIR lighting is very strong such as direct sunlight, and the noise produced by the differentiation has certain influence on the results.

The problem of active NIR image differencing is related to the use and combination of flash and non-flash images in computational photography in the visual spectrum. While active NIR image differencing methods focus on getting ambient illumination-free difference images for automatic face recognition, most of flash/no-flash methods focus on improving the visual quality of the no-flash image using the paired flash image. Various methods of enhancing color, reducing noise, and reducing shadows have been proposed. Dicarlo, *et al.*, introduced an active imaging method to measure ambient illumination using flash and non-flash image pairs [30]. Petschnigg, *et al.*, proposed the joint bilateral filter combining flash and non-flash images to achieve better exposure and color balance and to reduce noise [81]. In [38], Eisemann and Durand presented an alternative algorithm that shares some of the same basic concepts of [81]. Agrawal, *et al.*, presented a gradient projection and flash-exposure sampling scheme to remove photography artifacts [5]. In [31], shadows from color images were removed using flash/no-flash image edges. Sun, *et al.*, proposed an approach for foreground layer extraction using flash/no-flash image pairs [104]. Zhuo, *et al.*, developed an image deblurring approach to remove camera motion blur using a pair of blurred and flash images [134]. In [91], an iterative improvement of the guided image filter for flash/no-flash photography was presented. Li, *et al.*, introduced a hand-held multispectral camera to capture a pair of blurred image and NIR flash image simultaneously and analyze the correlation between the pair of images for blind motion deblurring [70]. Mikami, *et al.*, captured color and near-infrared images with

different exposure times for image enhancement under extremely low-light scene [78]. Yoon, *et al.*, presented an image enhancement method with flash and no-flash pairs based on adaptive total variation minimization [121].

However, these works do not attempt to combine exposures for moving subjects. When operating on human faces, some motion must be allowed for between frames and since the methods cited are based on the pixel-to-pixel correspondence, misalignment between images can cause artifacts and decrease performance. When an algorithm involves the difference between two pixels, even a very small misalignment can lead to significant artifacts in the result.

This chapter proposes new parametric image alignment methods which directly align face regions in the I-frame and A-frame to increase the efficiency and the accuracy, and a new image fusion method to reduce the noise in the D-frame. This research is motivated by the desire to have reliable face recognition in automobiles. Therefore, the face recognition system must allow natural movement of drivers and needs to be robust on noise in outdoor environment. The new image alignment algorithms are based on the parametric image alignment method proposed by Lucas and Kanade [75] and its inverse compositional variation proposed by Baker and Matthews [7]. To make the algorithm work on the I-frame and A-frame pairs under different illumination conditions, we define a new error minimization criterion and derive a new formula. And also we propose a pre-computation scheme for the inverse compositional algorithm to keep the advantage of fast calculation of the inverse compositional algorithm. From these methods, more accurate image alignment is achieved. In addition, warping and interpolation can be applied to A-frames to avoid smoothing on I-frames. Another problem with difference signals is that noise tends to be amplified. Therefore, in this work, we propose an algorithm for reducing the noise in the D-frame using detail information from the I-frame.

The rest of this chapter is organized as follows: the new parametric image alignment methods based on forward additive and inverse compositional Lucas-Kanade are presented in Section 5.2. Section 5.3 describes the image fusion method for reducing the noise and improving the image quality in the difference image. The details and results of the experiments carried out on the face video dataset of outdoor vehicular scenario are presented in Section 5.4. Finally, Section 5.5 concludes this chapter.

5.2 Image Alignment

Figure 25 and 26 illustrate the motion problem in active NIR image differencing. Figure 25 shows an example of an A-frame, an I-frame and a D-frame where there is no motion between the A-frame and the I-frame. And Fig. 26 shows an example of an A-frame, an I-frame and a D-frame where there is motion between the A-frame and the I-frame. The D-frame in Fig. 26 (c) shows artifacts introduced by the motion between the A-frame and the I-frame. These examples underscore the importance of accurate image alignment or motion compensation.

(a) A-frame (b) I-frame (c) D-frame

Figure 25: Example of images without motion.

(a) A-frame (b) I-frame (c) D-frame

Figure 26: Example of images with motion.

In this section, we introduce new parametric image alignment methods for active image differencing based on the forward additive and inverse compositional Lucas-Kanade image alignment methods. The subsection 5.2.1 summarizes the forward additive and inverse compositional Lucas-Kanade image alignment methods. Then, in the subsection 5.2.2, we propose a new error minimization criterion with non-positive error function which modifies the Lucas-Kanade methods so that the face region in A-frame can be directly aligned to the face region in I-frame.

5.2.1 Lucas-Kanade Image Alignment

5.2.1.1 Forward Additive Algorithm

The original Lucas-Kanade algorithm is a parametric and iterative image alignment method based on gradient descent [75]. It finds an image region in an input image $I(\mathbf{x})$ that best matches a template image $T(\mathbf{x})$, where $\mathbf{x} = (x, y)^{\mathrm{T}}$ represents the pixel coordinates. Here, the notations and derivations of equations are presented (based on [7]).

The vector $\mathbf{p} = (p_1, ..., p_n)^{\mathrm{T}}$ is a parameter vector and $\mathbf{W}(\mathbf{x}; \mathbf{p})$ is a warping function that maps pixels in the template T to the locations in the image I. The choice of the warp can be arbitrary including a translation, scaling, affine, piecewise affine or homography. Then the difference image D is the difference between two images, the template T and the warped image $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$:

$$D(\mathbf{x}; \mathbf{p}) = T(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})). \tag{4}$$

The difference image is the error that should be minimized in general image alignment applications. The goal of the Lucas-Kanade algorithm is to find the parameters that minimize the measure of error $\mathrm{E}(\mathbf{p})$:

$$\underset{\mathbf{p}}{\operatorname{argmin}}\, \mathrm{E}(\mathbf{p}) \tag{5}$$

where the measure of error $\mathrm{E}(\mathbf{p})$ is defined by the sum of squares of pixel intensities

in the difference image:

$$\mathrm{E}(\mathbf{p}) = \sum_{\mathbf{x}} \left[D(\mathbf{x}; \mathbf{p}) \right]^2 . \tag{6}$$

Minimizing the measure of error $\mathrm{E}(\mathbf{p})$ is a non-linear optimization problem because the pixel values $I(\mathbf{x})$ are, in general, non-linear in $\mathbf{x}$. In the Lucas-Kanade algorithm, the non-linear optimization is iteratively calculated using the Gauss-Newton gradient descent method. Starting with an initial guess of $\mathbf{p}$, the method proceeds by the iterations

$$\mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p}, \tag{7}$$

where the increment $\Delta\mathbf{p}$ is the solution to minimize the approximated measure of error $\hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p})$:

$$\underset{\Delta\mathbf{p}}{\operatorname{argmin}} \ \hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p}). \tag{8}$$

The approximated measure of error $\hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p})$ is derived as follows. First, the non-linear expression of the difference image $D(\mathbf{x}; \mathbf{p} + \Delta\mathbf{p})$ is linearized by applying a first order Taylor expansion of $I(\mathbf{W}(\mathbf{x}; \mathbf{p} + \Delta\mathbf{p}))$:

$$\hat{D}(\mathbf{x}; \mathbf{p} + \Delta\mathbf{p}) = T(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - \left(\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \right) \Delta\mathbf{p}. \tag{9}$$

Then, from Eq. (6) the measure of error is approximated:

$$\begin{aligned} \hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p}) &= \sum_{\mathbf{x}} \left[\hat{D}(\mathbf{x}; \mathbf{p} + \Delta\mathbf{p}) \right]^2 \\ &= \sum_{\mathbf{x}} \left[T(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})) - \left(\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \right) \Delta\mathbf{p} \right]^2 . \end{aligned} \tag{10}$$

In these equations, the term $\nabla I = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right)$ is the gradient of the image I and $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$ is the *Jacobian* of the warp. The problem of finding the increment $\Delta\mathbf{p}$ that minimizes Eq. (10) is a least squares problem and it can be solved by setting $\frac{\partial \hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p})}{\partial \Delta\mathbf{p}} = 0$:

$$\Delta\mathbf{p} = H^{-1} \sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \right]^{\mathrm{T}} \left[T(\mathbf{x}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})) \right] \tag{11}$$

where H is the *Hessian* matrix:

$$H = \sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \right]^{\mathrm{T}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \right]. \tag{12}$$

The Lucas-Kanade algorithm then consists of iteratively applying Eq. (11) and (7). This algorithm is referred to as the forward additive algorithm since the update in Eq. (7) is forward and additive.

5.2.1.2 Inverse Compositional Algorithm

Inverse compositional Lucas-Kanade algorithm is a computationally more efficient variation of the original forward additive Lucas-Kanade algorithm [7]. It reformulates the update rule in inverse compositional way so that some variables in the iteration are independent of the current parameter $\mathbf{p}$ and can be precomputed. It is proved that the inverse compositional algorithm is equivalent to the forward additive algorithm up to the first order in $\Delta\mathbf{p}$ [7]. The difference image D, the measure of error E to be minimized, and the warp $\mathbf{W}(\mathbf{x};\mathbf{p})$ update in each step are redefined as follows:

$$D(\mathbf{x};\Delta\mathbf{p},\mathbf{p}) = T(\mathbf{W}(\mathbf{x};\Delta\mathbf{p})) - I(\mathbf{W}(\mathbf{x};\mathbf{p})), \tag{13}$$

$$\mathrm{E}(\Delta\mathbf{p},\mathbf{p}) = \sum_{\mathbf{x}} \left[D(\mathbf{x};\Delta\mathbf{p},\mathbf{p}) \right]^2, \tag{14}$$

$$\mathbf{W}(\mathbf{x};\mathbf{p}) \leftarrow \mathbf{W}(\mathbf{x};\mathbf{p}) \circ \mathbf{W}(\mathbf{x};\Delta\mathbf{p})^{-1}. \tag{15}$$

The difference from the forward additive algorithm is that the small increment of the parameters $\Delta\mathbf{p}$ is applied to the warp for the template image T as in Eq. (13). And the increment $\Delta\mathbf{p}$ is from zero and not from the parameters $\mathbf{p}$ of the current iteration as in the forward additive algorithm. The increment $\Delta\mathbf{p}$ that approximately minimizes Eq. (14) is used to update the warp $\mathbf{W}(\mathbf{x};\mathbf{p})$ as in Eq. (15) and then the updated warp is used for the input image $I(\mathbf{W}(\mathbf{x};\mathbf{p}))$ back in Eq. (13). This algorithm is referred to as the inverse compositional algorithm because the warp update in Eq. (15) involves the inversion and the composition. The difference image $D(\mathbf{x};\Delta\mathbf{p},\mathbf{p})$

is linearly approximated by the first order Taylor expansion of the template image $T(\mathbf{W}(\mathbf{x};\Delta\mathbf{p}))$:

$$\hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p}) = T(\mathbf{W}(\mathbf{x};0)) + \left(\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right)\Delta\mathbf{p} - I(\mathbf{W}(\mathbf{x};\mathbf{p})), \tag{16}$$

and the measure of error E is approximated accordingly:

$$\hat{\mathrm{E}}(\Delta\mathbf{p},\mathbf{p}) = \sum_{\mathbf{x}}\left[\hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p})\right]^2. \tag{17}$$

In Eq. (16), the gradient ∇T is evaluated at $\mathbf{x}$ and the Jacobian $\frac{\partial\mathbf{W}}{\partial\mathbf{p}}$ is evaluated at $(\mathbf{x};0)$ where we set $\mathbf{W}(\mathbf{x};\mathbf{0}) = \mathbf{x}$ without loss of generality. Then $\frac{\partial\hat{\mathrm{E}}(\Delta\mathbf{p},\mathbf{p})}{\partial\Delta\mathbf{p}} = 0$ can be solved by replacing $\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})$ with Eq. (16):

$$\Delta\mathbf{p} = -H^{-1}\sum_{\mathbf{x}}\left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} D(\mathbf{x};\mathbf{0},\mathbf{p}), \tag{18}$$

where

$$H = \sum_{\mathbf{x}}\left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}}\left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]. \tag{19}$$

Because the gradient ∇T and the Jacobian $\frac{\partial\mathbf{W}}{\partial\mathbf{p}}$ are independent of $\mathbf{p}$, the gradient ∇T, the Jacobian $\frac{\partial\mathbf{W}}{\partial\mathbf{p}}$, the steepest descent image $\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}$ and the Hessian H can be precomputed in advance before the iteration.

5.2.2 Image Alignment with Nonpositive Error Function

For active illumination applications, we have two images to align: an I-frame that is taken with the active illumination and an A-frame that is taken with only ambient illumination. The template $T(\mathbf{x})$ is set to be a detected face region in the I-frame, and the input image $I(\mathbf{x})$ is set to be the A-frame. In most image alignment applications, the difference image D is the error that should be minimized. But in the active image differencing method, the difference image D is the desired output image. Since the I-frame is taken with additional illumination, the pixel intensities in the I-frame should be greater than or equal to the corresponding pixel intensities in the A-frame assuming that pixel intensities corresponding to the ambient illumination do not

change between the two images. Therefore, positive pixel values are considered as the desired output, and nonpositive pixel values are considered as error or noise. Consequently we modified the measure of error so that only nonpositive intensities are considered as errors.

5.2.2.1 Forward Additive Algorithm

First, we generalize the measure of error term in Eq. (5) by introducing an error function $\rho(x)$:$\mathbb{R} \rightarrow \mathbb{R}$ instead of the Euclidean L2 norm. The goal is then to minimize this measure of error:

$$\mathrm{E}(\mathbf{p}) = \sum_{\mathbf{x}} \rho(D(\mathbf{x}; \mathbf{p})). \tag{20}$$

To penalize negative values in the difference image, we choose the error function as

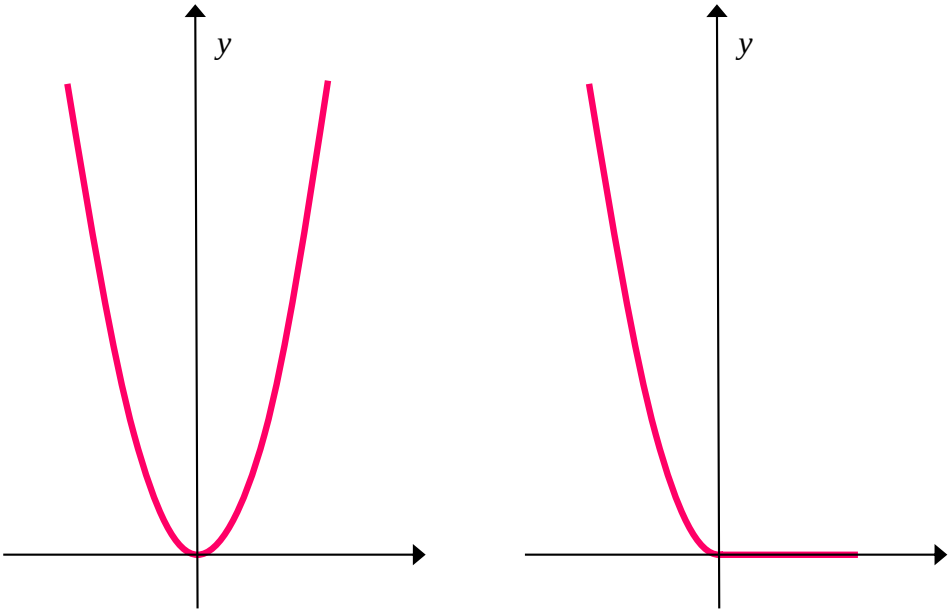
$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & ; \quad x < 0 \\ 0 & ; \quad x \geq 0. \end{cases} \tag{21}$$

For negative values, the L2 norm error function is chosen as in the original Lucas Kanade method. For positive values, the error is set to be zero since the positive values are not errors. The $\frac{1}{2}$ factor is used to simplify the following equations. Figure 27 shows the original L2 norm error function and proposed nonpositive error function. The derivative of the error function $\rho(x)$ is x for negative values of x and zero for non-negative values of x. Then, the first-order Tayler series expansion of

$$\hat{\mathrm{E}}(\mathbf{p} + \Delta\mathbf{p}) = \sum_{\mathbf{x}} \rho\Big(\hat{D}(\mathbf{x}; \mathbf{p} + \Delta\mathbf{p})\Big) \tag{22}$$

is the same as in Eq. (10) except that the function ρ is applied to each term in the sum, which is equivalent to restricting the sum to only those terms that are negative.

To solve the minimization problem in Eq. (8), we calculate the partial derivative of the approximated measure of error with respect to $\Delta\mathbf{p}$, and find the solution that



(a) Original error (b) Proposed error

Figure 27: Nonpositive error function.

makes the equation to be equal to zero:

$$\frac{\partial \hat{\mathrm{E}}(\mathbf{p}+\Delta\mathbf{p})}{\partial \Delta\mathbf{p}} = -\sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^{\mathrm{T}} \rho'\left(\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})\right) \tag{23}$$

$$= -\sum_{\left\{\mathbf{x}:\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})<0\right\}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^{\mathrm{T}} \hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p}).$$

The problem is that this equation cannot be solved analytically because the condition of the summation includes $\Delta\mathbf{p}$ which make the equation non-linear with respect to $\Delta\mathbf{p}$. To solve the equation, we assume that in the condition of the summation,

$$\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p}) \approx D(\mathbf{x};\mathbf{p}) \tag{24}$$

for small $\Delta\mathbf{p}$. Then Eq. (23) can be solved by replacing $\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})$ with Eq. (9) and solving a least squares problem in the same manner as in Eq. (11) and (12):

$$\Delta\mathbf{p} = H^{-1} \sum_{\{\mathbf{x}:D(\mathbf{x};\mathbf{p})<0\}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^{\mathrm{T}} D(\mathbf{x};\mathbf{p}), \tag{25}$$

where

$$H = \sum_{\{\mathbf{x}:D(\mathbf{x};\mathbf{p})<0\}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^{\mathrm{T}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]. \tag{26}$$

Note that the terms in Eq. (25) and (26) are included in the summations only if the difference $D(\mathbf{x};\mathbf{p})$ is negative. That means the updates are driven by the negative values which are the artifacts of the misalignments. Therefore, this algorithm works directly to remove the misalignment artifacts on every iteration.

5.2.2.2 Inverse Compositional Algorithm

The inverse compositional algorithm can also be reformulated to use the nonpositive error function resulting in a more efficient algorithm than the forward-additive nonpositive algorithm. The error measure, E, for the inverse compositional Lucas Kanade image alignment method in Eq. (14) is generalized with the error function $\rho(x)$:

$$\mathrm{E}(\Delta\mathbf{p},\mathbf{p}) = \sum_{\mathbf{x}} \rho(D(\mathbf{x};\Delta\mathbf{p},\mathbf{p})) \tag{27}$$

where we choose the same nonpositive error function $\rho(x)$ as in Eq. (21). Then the partial derivative of the approximated error measure on the parameter increment is approximated as below:

$$\begin{aligned}\frac{\partial\hat{\mathrm{E}}(\Delta\mathbf{p},\mathbf{p})}{\partial\Delta\mathbf{p}} &= \sum_{\mathbf{x}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} \rho'\left(\hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p})\right) \\ &= \sum_{\left\{\mathbf{x}:\hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p})<0\right\}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} \hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p}) \\ &\approx \sum_{\{\mathbf{x}:D(\mathbf{x};\mathbf{0},\mathbf{p})<0\}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} \hat{D}(\mathbf{x};\Delta\mathbf{p},\mathbf{p}),\end{aligned} \tag{28}$$

using the similar approximation in the condition of the summation as in Eq. (24) for small $\Delta\mathbf{p}$. Then $\frac{\partial\hat{\mathrm{E}}(\Delta\mathbf{p},\mathbf{p})}{\partial\Delta\mathbf{p}} = 0$ can be solved by replacing $\hat{D}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})$ with Eq. (16):

$$\Delta\mathbf{p} = -H^{-1} \sum_{\{\mathbf{x}:D(\mathbf{x};\mathbf{0},\mathbf{p})<0\}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} D(\mathbf{x};\mathbf{0},\mathbf{p}), \tag{29}$$

where

$$H = \sum_{\{\mathbf{x}:D(\mathbf{x};\mathbf{0},\mathbf{p})<0\}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^{\mathrm{T}} \left[\nabla T\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]. \tag{30}$$

As in the original inverse compositional Lucas-Kanade algorithm, the gradient ∇T, the Jacobian $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$, and the steepest descent image $\nabla T \frac{\partial \mathbf{W}}{\partial \mathbf{p}}$ are independent of the current parameter $\mathbf{p}$ and can be precomputed and re-used. On the other hand, for the computation of the Hessian H in Eq. (30), the elements of the summation, which are the outer products of the steepest descent image, $\nabla T \frac{\partial \mathbf{W}}{\partial \mathbf{p}}$, with itself, are independent of $\mathbf{p}$, but the condition of the summation depends on $\mathbf{p}$ and the Hessian cannot be precomputed. Therefore, the outer products are precomputed, stored and included in the summation only for the negative pixels in the difference image, $D(\mathbf{x}; \mathbf{0}, \mathbf{p})$, to make the computation of the Hessian more efficient.

Figure 28 shows an example of the results using no motion compensation, using the motion interpolation method by [135], and using our alignment methods with nonpositive error functions in the forward additive and inverse compositional ways. As seen in the figure, the positive difference image with no motion compensation shows severe motion artifacts especially around edges and facial features. The negative difference image with no motion compensation has significant pixel intensities around edges and facial features. The motion interpolation method reduces the motion artifacts on the positive difference image and the pixel intensities on the negative difference image. And our forward additive and inverse compositional methods with nonpositive error function further reduce those noises in both positive and negative difference images. The similarity of the results of forward additive and inverse compositional methods shows the equivalence of the two algorithms. Figure 29 and 30 show the experimental result images after applying face detection and image alignment with nonpositive error function.

5.3 Image Fusion

When motion is compensated by the image alignment methods proposed in the previous section, the D-frame is (ideally) equivalent to the night-time image where the

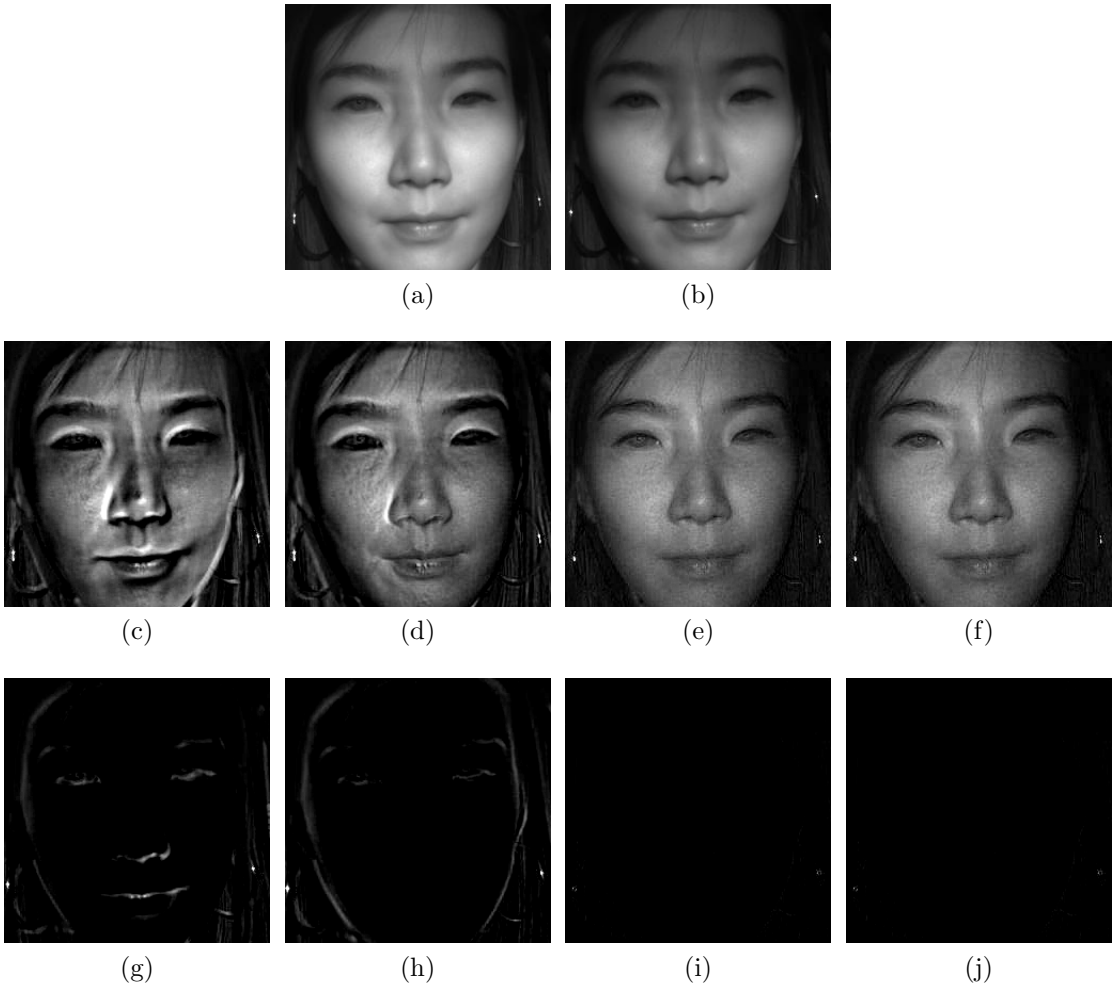


Figure 28: Experimental comparison of image alignment methods: (a) detected face region in the I-frame (b) corresponding face region in the A-frame (c) positive difference image without alignment (d) positive difference image with motion interpolation (e) positive difference image with forward additive alignment method with nonpositive error function (f) positive difference image with inverse compositional alignment method with nonpositive error function (g) negative difference image without alignment (h) negative difference image with motion interpolation (i) negative difference image with forward additive alignment method with nonpositive error function (j) negative difference image with inverse compositional alignment method with nonpositive error function.

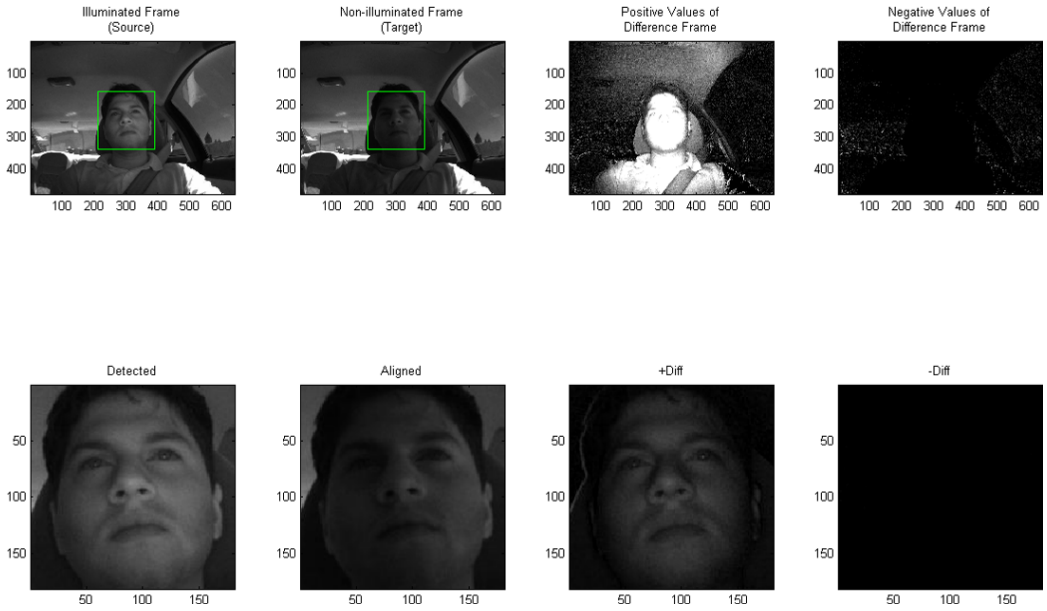


Figure 29: Experimental result 1 of face detection and image alignment.

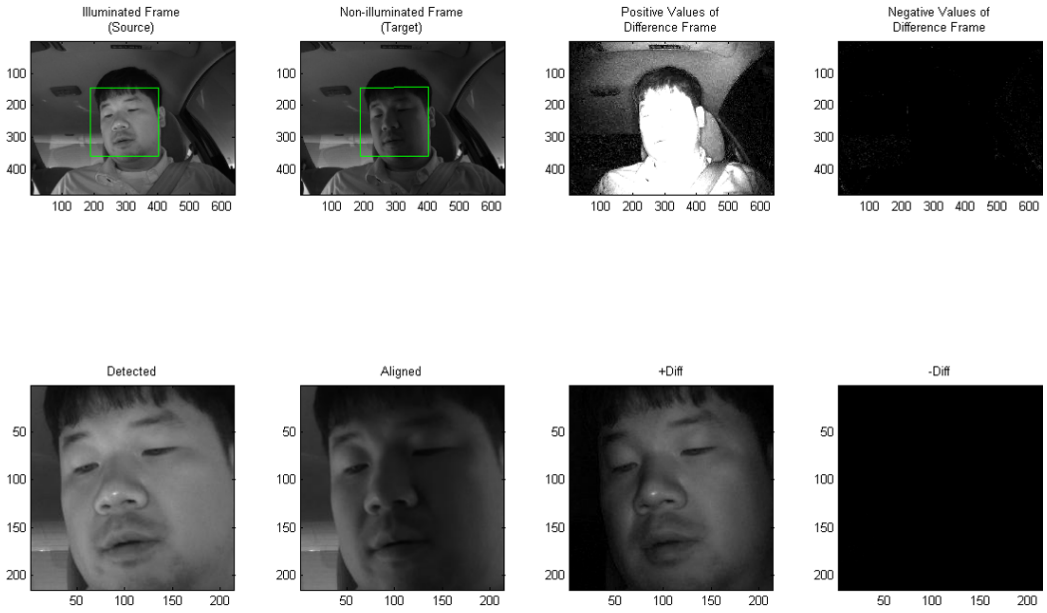


Figure 30: Experimental result 2 of face detection and image alignment.

only illumination is the active LED illumination. However, when the illumination intensity of ambient light is much larger than that of LED illumination, the image quality of the D-frame is poor and the difference frame exhibits noise and has low signal-to-noise ratio (SNR). The main reason for the increased noise in the D-frame is the increased photon shot noise in the I-frame and A-frame. Photon shot noise is caused by statistical quantum fluctuations and is proportional to the pixel intensity. Under bright illumination, photon shot noise dominates the noise behaviour of the sensor [76].

Let S_i and N_i be the signal and noise intensity of a pixel in the I-frame, and S_a and N_a be the signal and noise intensity of the corresponding pixel in the A-frame. Then the total pixel intensity on each frame is the sum of the corresponding signal and noise value:

$$I_i = S_i + N_i, \quad I_a = S_a + N_a. \tag{31}$$

When the signal value, S_i or S_a, is expressed in terms of electrons, the photon shot noise has a standard deviation of

$$\sigma_i = \sqrt{S_i}, \quad \sigma_a = \sqrt{S_a} \tag{32}$$

also with units of electrons [35]. Then the value of the variance of the photon shot noise is equivalent to the signal value (with units of signal squared):

$$\sigma_i^2 = \mathrm{E}\left[N_i^2\right] = S_i, \quad \sigma_a^2 = \mathrm{E}\left[N_a^2\right] = S_a. \tag{33}$$

And the SNR is also equivalent to the signal value:

$$\mathrm{SNR}_i = \frac{\mathrm{E}\left[S_i^2\right]}{\mathrm{E}\left[N_i^2\right]} = \frac{S_i^2}{S_i} = S_i, \quad \mathrm{SNR}_a = \frac{\mathrm{E}\left[S_a^2\right]}{\mathrm{E}\left[N_a^2\right]} = \frac{S_a^2}{S_a} = S_a. \tag{34}$$

If we define the signal, noise and total pixel intensity of the D-frame as follows:

$$S_d = S_i - S_a, \quad N_d = N_i - N_a, \quad I_d = I_i - I_a = S_d + N_d, \tag{35}$$

then, under the assumption that N_i and N_a are independent, the noise variance of the corresponding pixel in the D-frame is

$$\sigma_d^2 = \mathrm{E}\left[N_d^2\right] = \mathrm{E}\left[(N_i - N_a)^2\right] = \mathrm{E}\left[N_i^2 + N_a^2\right] = S_i + S_a = S_d + 2S_a, \tag{36}$$

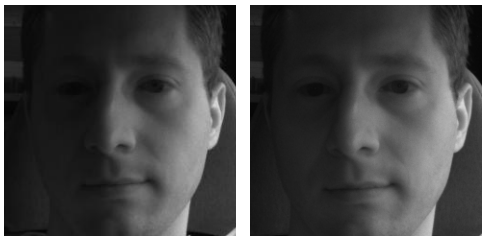
and the SNR of the pixel in the D-frame is

$$\mathrm{SNR}_d = \frac{\mathrm{E}\left[S_d^2\right]}{\mathrm{E}\left[N_d^2\right]} = \frac{S_d^2}{S_d + 2S_a} = \frac{S_d}{1 + 2\frac{S_a}{S_d}}. \tag{37}$$

When there is no ambient illumination ($S_a = 0$), SNR_d is equivalent to the difference signal S_d which is the case of the night-time image. If the difference signal S_d is fixed and the ambient signal S_a is increased, then the noise is increased and SNR_d is decreased. The increased noise can be observed in Fig. 31. The relative intensity of the ambient illumination to LED illumination is large in this example. Therefore, the difference image shows high noise level. Since the ambient illumination is much brighter on left side of the subjects face, the noise is more visible on that side. The SNR_d is further reduced if the camera exposure level is reduced to prevent the possible pixel intensity saturation due to the increased ambient illumination. This can be explained by the reduced S_d with the fixed ratio $\frac{S_a}{S_d}$ in Eq. (37).

One way of dealing with the noise problem caused by the bright ambient illumination is increasing the active NIR illumination power. For example, a narrow band NIR laser generator is used in [120] to overpower outdoor sunlight. However, they reported that their method is not perfect when the ambient NIR lighting is very strong such as direct sunlight, and the noise produced by the differentiation has certain influence on the results. And powerful NIR lasers have other drawbacks, especially when directed at a person's face at close range. Therefore, a different solution is presented here.

One observation is that the D-frame has good illumination levels across the entire face but may exhibit unacceptable levels of noise. On the other hand, the I-frame tends to have good detail quality and low noise but unacceptable variations in illumination. The ideal image for face recognition would combine the advantages of



(a) Ambient (b) Illuminated

(c) Difference (d) Detail (e) Fusion

Figure 31: Example ambient, illuminated, difference, detail and fusion images are shown when the difference image contains visible noise.

the D-frame and I-frame. Since the illumination is largely a low-pass phenomenon, combining the illumination (low-pass spatial components) from the D-frame with the detail (high-pass spatial components) from the I-frame may result in better recognition performance. Here we introduce image fusion methods to combine low frequency of the D-frame and high frequency of the I-frame. We de-noise and transfer detail to merge the controlled illumination of the D-frame with the high-frequency detail in the I-frame.

The high-frequency detail from the I-frame is computed as the following ratio:

$$Q_i = \frac{I_i}{\hat{I}_i}, \tag{38}$$

where I_i is the pixel intensity of the I-frame and $\hat{I}_i$ is the smoothed version of I_i. Then the image fusion is the product of the smoothed version of the D-frame and the detail from the I-frame:

$$I_{fusion} = \hat{I}_d Q_i. \tag{39}$$

To compute the smoothed version of the I-frame and the D-frame, we apply an edge-preserving smoothing filter such as the bilateral filtering [106]. The bilateral filter is designed to average together pixels that are spatially near on another and have similar intensity values. It combines a classic low-pass filter with an edge-stopping function that attenuates the filter kernel weights when the intensity difference between pixels is large. In the notation of [34], the bilateral filter computes the value of pixel p for the I-frame I_i as:

$$\hat{I}_{i,p} = \frac{1}{k(p)} \sum_{p' \in \Omega} g_d\left(p' - p\right) g_r\left(I_{i,p} - I_{i,p'}\right) I_{i,p'}, \tag{40}$$

where $k(p)$ is a normalization term:

$$k(p) = \sum_{p' \in \Omega} g_d\left(p' - p\right) g_r\left(I_{i,p} - I_{i,p'}\right). \tag{41}$$

The function g_d sets the weight in the spatial domain based on the distance between the pxiels, while the edge-stopping function g_r sets the weight on the range based on intensity differences. Typically, both functions are Gaussian with widths controlled by the standard deviation parameters σ_d and σ_r respectively. By experiments, we set σ_d to be 4 pixels and σ_r to be 0.2 when pixel values are normalized to [0.0 1.0]. The smoothed version of the D-frame is computed in the same manner with the same value of σ_d and σ_r. Figure 31 shows an example of the detail image and resulted fusion image. The detail in the I-frame is transferred to the fusion image through the detail image and the low-pass spatial components are transferred from the D-frame while the noise in the D-frame is reduced.

5.4 Experiments

We used the same NIR vehicular-scenario face video dataset that was introduced in the previous chapter. Face regions in I-frames of the videos are detected using a modified version [72] of the boosted classifier proposed by Viola Jones [112]. Corresponding face regions in A-frames are found and aligned by our proposed methods and the difference

face images are calculated. A warping function for the image alignment is modeled as a single affine transformation throughout the whole face image. Piecewise or block-based affine transformation can model more complicated and deformable face motion. However, it requires computationally expensive motion interpolation between pieces or blocks and small deformation between frames makes the single affine transformation approximates the motion with enough precision. The motion interpolation method is also applied to consecutive I-frames for comparison. Then the difference face images are preprocessed to mask the image to consider only pixels inside the face boundary. The pixel values are then normalized to have zero mean and a unit standard deviation. For recognition, two popular dimension reduction methods are applied: Principal Component Analysis (PCA) [109] and Linear Discriminant Analysis (LDA) [13]. PCA and LDA subspaces are trained by CBSR NIR database [69] which has 3,940 NIR face images of 197 subjects.

To test the effectiveness of the proposed direct image alignment methods on face recognition, image and video-based face identification experiments were performed. The gallery set contains the 40 head rotation gallery videos of 40 subjects. And the probe set contains randomly chosen probe videos of 40 subjects. The experimental setting is the same as in Section 4.8. Then each found face image in each probe video is compared with all the found face images in the 40 gallery videos, and the face identity is determined by the minimum distance between two images in three subspace domains: raw, PCA and LDA. Then the same experiment is repeated 200 times with other randomly selected probe videos. The results in Table 16 (a) show that active NIR image differencing improves face recognition performance, and motion interpolation method using Black and Anandan optical flow method further improves the result. Furthermore, our proposed direct image alignment methods based on forward additive and inverse compositional Lucas-Kanade outperform both prior methods in

the raw, PCA, and LDA domains. For the video-based face identification experiment, pose clustering method is applied. Table 16 (b) shows the results and it also shows that our methods outperform the previous methods in the raw, PCA, and LDA domains.

Table 16: Face recognition experiment results. MI-BA stands for motion interpolation method using Black and Anandan optical flow method. NPE-FA stands for the forward additive image alignment with the non-positive error function. NPE-IC stands for the inverse compositional image alignment with the non-positive error function.

Type	RAW	PCA	LDA
A-frame	0.508	0.521	0.586
I-frame	0.629	0.639	0.656
D-frame	0.700	0.726	0.751
MI-BA	0.726	0.753	0.769
NPE-FA	**0.742**	**0.762**	**0.819**
NPE-IC	**0.742**	0.761	**0.819**

(a) Image-based

Type	RAW	PCA	LDA
A-frame	0.533	0.552	0.601
I-frame	0.640	0.676	0.709
D-frame	0.739	0.748	0.792
MI-BA	0.812	0.871	0.882
NPE-FA	**0.845**	**0.898**	**0.932**
NPE-IC	**0.845**	**0.898**	**0.932**

(b) Video-based

Face recognition results improve further when the motion estimation is combined with image fusion. Table 17 shows that a consistent 3% improvement in recognition accuracy results when motion estimation is combined with image fusion.

Table 17: Face recognition experiment results with image fusion.

Type	RAW	PCA	LDA
D-frame	0.729	0.759	0.764
MI-BA	0.749	0.789	0.795
NPE-FA	**0.788**	**0.798**	**0.849**
NPE-IC	0.787	**0.798**	**0.849**

(a) Image-based

Type	RAW	PCA	LDA
D-frame	0.766	0.770	0.828
MI-BA	0.839	0.909	0.912
NPE-FA	**0.890**	**0.939**	**0.961**
NPE-IC	0.889	0.938	**0.961**

(b) Video-based

5.5 Conclusion

Active NIR image differencing improves robustness of face recognition by generating images only lit by active NIR illumination. However, motion between frames

introduces artifacts and strong ambient illumination introduces noise in the difference frame. This chapter proposed methods for image alignment and noise reduction. First, we proposed image alignment methods which directly align face images in the illuminated frame and the ambient frame. The methods can compensate for non-linear motions and are based on forward additive and inverse compositional Lucas-Kanade image alignment with a modified error criteria. Second, we proposed an image fusion method to reduce noise in the difference frame and include the detail information from the illuminated frame. Extensive experiments on a face video dataset from an outdoor vehicular scenario show that the face recognition performance increases by both of the image alignment and image fusion methods. We note that since both NIR and visible lights are reflected on the surface and share the same physical properties, the proposed direct image alignment and image fusion methods can also be applied to flash and no-flash image pair applications.

The end-to-end system is composed of submodules introduced in Chapter 4 and 5. Each submodule was designed to be computationally efficient for embedded systems. However, submodules need to be selectively chosen if the hardware specification is more restricted and very limited. Image differencing, cascade face detection and LDA dimension reduction methods do not involve iterative algorithms and are highly efficient. Those submodules are indispensable parts of the system and should be included. On the other hand, correlation based motion detection, morphological filter based background subtraction, K-means based pose clustering and Lucas-Kanade based image alignment methods require more computational resources compared to the benefit they provide. Therefore, those can be removed from the system. Image fusion and illumination robust features are efficient methods but those are additional supplements for the system. Therefore, those can be included optionally.

CHAPTER VI

CONCLUSION

In this dissertation, we have developed a system of practical technologies to implement an illumination-robust, consumer-grade biometric system based on face recognition to be used in the automotive market and ultimately result in very low-cost, easy-to-deploy solutions enabling a wide variety of applications that can benefit from personalization.

First, we presented an end-to-end face recognition system using NIR illumination and camera system. The key advantage of NIR over visible light is that it is invisible to the human eye. Compared to the previously described scenario, NIR is easily available from the sun during the day since it is an abundant source. At night however, an NIR illuminator can be used to provide the controlled artificial illumination without bothering the driver. This feature laid the foundation for the end product to be non-intrusive. The system consists of three stages; face detection, eye detection and face recognition. The performance of each module and the end-to-end system was tested by the NIR dataset taken in indoor simulating the vehicular environment.

Second, we improved the NIR face recognition system by introducing the image differencing method. Providing illumination sufficient to overcome the shadows cast in full sun is impractical. Therefore, we proposed active NIR image differencing which takes the difference between successive image frames, one illuminated and one not illuminated. In ideal condition when the camera image processing pipeline is linear on pixel intensities and there is no motion between frames, the image differencing method removes the effect of the ambient illumination and yields ambient illumination free face images. We developed an end-to-end face recognition system including the active

NIR image differencing, foreground/background segmentation, motion detection, face detection, pose clustering and face recognition. And it was shown that the image differencing method makes the modules more robust to the ambient illumination variation. Vehicular application videos were taken in extremely challenging outdoor illumination and shadowing conditions and used to test each module. Extensive test results of vehicular scenario were provided to evaluate the end-to-end system.

Lastly, we addressed several aspects of the problem in active NIR image differencing which are motion artifact and noise in the difference frame, namely how to efficiently and more accurately align the illuminated frame and ambient frame, and how to combine information in the difference frame and the illuminated frame. We proposed image alignment methods which directly align face images in the illuminated frame and the ambient frame. The methods can compensate for non-linear motions and are based on forward additive and inverse compositional Lucas-Kanade image alignment with a modified error criteria. Then we proposed an image fusion method to reduce noise in the difference frame and include the detail information from the illuminated frame. Extensive experiments on the face video dataset of the outdoor vehicular scenario showed that the face recognition performance increased by both of the image alignment and image fusion methods.

The current solution uses only frontal face images and does not deal with occlusion problems explicitly. The system could be improved by considering face images with other face poses and also considering occlusion problems. More advanced decision methods can further improve the accuracy in the decision on the subject identification. And we note that since both NIR and visible lights are reflected on the surface and share the same physical properties, the proposed direct image alignment and image fusion methods can also be applied to flash and no-flash image pair applications.

Publications, submissions, and presentations based on this work so far are as follows:

- J. Kang, D. Anderson and M. Hayes, "Image Alignment and Image Fusion for Face Recognition with Active Near Infrared Image Differencing," *Optics Express*. [submitted]

- J. Kang, D. Anderson and M. Hayes, "Face Recognition for Vehicle Personalization with Near Infrared Frame Differencing," *IEEE Transactions on Consumer Electronics*. [submitted]

- J. Kang, D. Anderson, and M. Hayes, "Direct image alignment for active near infrared image differencing," in *Advanced Concepts for Intelligent Vision Systems*, pp. 334-344, Springer, 2015.

- J. Kang, D. Anderson, and M. Hayes, "Face recognition in vehicles with near infrared frame differencing," in *IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, pp. 358-363, 2015.

- J. Kang, and M. Hayes, "Face recognition for vehicle personalization with near-ir frame differencing and pose clustering," in *IEEE International Conference on Consumer Electronics*, pp. 455-456, 2015.

- H. Park, J. Choo, B. Drake, and J. Kang, "Linear discriminant analysis for data with subcluster structure," in *IEEE International Conference on Pattern Recognition*, pp. 1-4, 2008

- J. Kang, A. Borkar, A. Yeung, N. Nong, M. Smith, and M. Hayes, "Short wavelength infrared face recognition for personalization," in *IEEE International Conference on Image Processing*, pp. 2757-2760, 2006.

REFERENCES

[1] "Equinox." http://www.equinoxsensors.com. Accessed: 2016-05-28.

[2] "Inverse square law." http://hyperphysics.phy-astr.gsu.edu/hbase/forces/isq.html. Accessed: 2016-05-28.

[3] "Mobile phone restrictions fact sheet." https://www.fmcsa.dot.gov/driver-safety/distracted-driving/mobile-phone-restrictions-fact-sheet. Accessed: 2016-05-28.

[4] Adini, Y., Moses, Y., and Ullman, S., "Face recognition: the problem of compensating for changes in illumination direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, 1997.

[5] Agrawal, A., Raskar, R., Nayar, S. K., and Li, Y., "Removing photography artifacts using gradient projection and flash-exposure sampling," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 828–835, 2005.

[6] Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., and others, "The banca database and evaluation protocol," in *Audio-and Video-Based Biometric Person Authentication*, pp. 625–638, Springer, 2003.

[7] Baker, S. and Matthews, I., "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[8] Basri, R. and Jacobs, D., "Lambertian reflectance and linear subspaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003.

[9] Basri, R. and Jacobs, D. W., "Lambertian reflectance and linear subspaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003.

[10] Batur, A. and Hayes III, M., "Segmented Linear Subspaces for Illumination-Robust Face Recognition," *International Journal of Computer Vision*, vol. 57, no. 1, pp. 49–66, 2004.

[11] Bebis, G., Gyaourova, A., Singh, S., and Pavlidis, I., "Face recognition by fusing thermal infrared and visible imagery," *Image and Vision Computing*, vol. 24, no. 7, pp. 727–742, 2006.

[12] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J., "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," vol. 19, pp. 711–720, Jul. 1997.

[13] Belhumeur, P., Hespanha, J., Kriegman, D., and others, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[14] Belhumeur, P. and Kriegman, D., "What Is the Set of Images of an Object Under All Possible Illumination Conditions?," *International Journal of Computer Vision*, vol. 28, no. 3, pp. 245–260, 1998.

[15] Black, M. J. and Anandan, P., "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer vision and image understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[16] Blanz, V. and Vetter, T., "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003.

[17] Blass, E., "Toyota's remote-entry wrist watch." September 2005.

[18] Bowyer, K. W., Chang, K., and Flynn, P., "A survey of approaches to three-dimensional face recognition," in *null*, pp. 358–361, IEEE, 2004.

[19] Bowyer, K. W., Chang, K., and Flynn, P., "A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition," *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[20] Chan, C. H., Zou, X., Poh, N., and Kittler, J., "Illumination invariant face recognition: a survey," *Face Recognition in Adverse Conditions*, pp. 147–166, 2014.

[21] Chang, K. I., Bowyer, K. W., and Flynn, P. J., "An evaluation of multimodal 2d+ 3d face biometrics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 4, pp. 619–624, 2005.

[22] Chen, C.-P. and Chen, C.-S., "Lighting normalization with generic intrinsic illumination subspace for face recognition," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1089–1096, IEEE, 2005.

[23] Chen, H. F., Belhumeur, P. N., and Jacobs, D. W., "In search of illumination invariants," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 254–261, IEEE, 2000.

[24] Chen, T., Yin, W., Zhou, X. S., Comaniciu, D., and Huang, T. S., "Illumination normalization for face recognition and uneven background correction using total variation based image models," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 532–539, IEEE, 2005.

[25] Chen, T., Yin, W., Zhou, X. S., Comaniciu, D., and Huang, T. S., "Total variation models for variable lighting face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 9, pp. 1519–1524, 2006.

[26] Chen, W., Er, M. J., and Wu, S., "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 2, pp. 458–466, 2006.

[27] Chen, X., Flynn, P., and Bowyer, W., "Visible-light and infrared face recognition," in *Workshop on Multimodal User Authentication*, p. 48, Citeseer, 2003.

[28] Cornsweet, T. and Crane, H., "Accurate two-dimensional eye tracker using first and fourth Purkinje images," *J. Opt. Soc. Am*, vol. 63, no. 8, pp. 921–928, 1973.

[29] Devore, J. and others, *Probability and statistics for engineering and the sciences*. Duxbury, 2000.

[30] DiCarlo, J. M., Xiao, F., and Wandell, B. A., "Illuminating illumination," in *Color and Imaging Conference*, pp. 27–34, Society for Imaging Science and Technology, 2001.

[31] Drew, M. S., Lu, C., and Finlayson, G. D., "Removing shadows using flash/noflash image edges," in *IEEE International Conference on Multimedia and Expo*, pp. 257–260, IEEE, 2006.

[32] Du, S. and Ward, R., "Wavelet-based illumination normalization for face recognition," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2, pp. II–954, IEEE, 2005.

[33] Duda, R., Hart, P., and Stork, D., *Pattern Classification*. Wiley-Interscience, 2000.

[34] Durand, F. and Dorsey, J., "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM transactions on graphics (TOG)*, vol. 21, pp. 257–266, ACM, 2002.

[35] Eastman Kodak Company, Rochester, New York, *CCD Image Sensor Noise Sources*, Aug. 2001.

[36] Ebisawa, Y., "Improved video-based eye-gaze detection method," *Instrumentation and Measurement, IEEE Transactions on*, vol. 47, no. 4, pp. 948–955, 1998.

[37] Ebisawa, Y. and Satoh, S., "Effectiveness of pupil area detection technique using two light sources and image difference method," *Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE*, pp. 1268–1269, 1993.

[38] Eisemann, E. and Durand, F., "Flash photography enhancement via intrinsic relighting," *ACM transactions on graphics*, vol. 23, no. 3, pp. 673–678, 2004.

[39] Farokhi, S., Shamsuddin, S. M., Sheikh, U., Flusser, J., Khansari, M., and Jafari-Khouzani, K., "Near infrared face recognition by combining zernike moments and undecimated discrete wavelet transform," *Digital Signal Processing*, vol. 31, pp. 13–27, 2014.

[40] Gao, Y. and Leung, M. K., "Face recognition using line edge map," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 6, pp. 764–779, 2002.

[41] Garea, E., Heydi, L., Vazquez, M., Kittler, J., and Messer, K., "An illumination insensitive representation for face verification in the frequency domain," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 215–218, IEEE, 2006.

[42] Georghiades, A., Belhumeur, P., and Kriegman, D., "From few to many: illumination cone models for face recognition under variable lighting and pose," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 6, pp. 643–660, 2001.

[43] Ghiass, R. S., Arandjelović, O., Bendada, A., and Maldague, X., "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, vol. 47, no. 9, pp. 2807–2824, 2014.

[44] Gonzalez, R. and Woods, R., *Digital Image Processing*. Prentice Hall, third edition, 2007.

[45] Gross, R. and Brajovic, V., "An image preprocessing algorithm for illumination invariant face recognition," in *Audio-and Video-Based Biometric Person Authentication*, pp. 10–18, Springer, 2003.

[46] Hallinan, P. W., "A low-dimensional representation of human faces for arbitrary lighting conditions," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pp. 995–999, IEEE, 1994.

[47] Hansen, D. and Hammoud, R., "An improved likelihood model for eye tracking," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 220–230, 2007.

[48] Haro, A., Flickner, M., and Essa, I., "Detecting and tracking eyes by using their physiologicalproperties, dynamics, and appearance," *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, 2000.

[49] Haro, A., Flickner, M., and Essa, I., "Detection and tracking eyes by using their physiological properties, dynamics, and appearance," pp. 163–168, 2000.

[50] Heo, J., Savvides, M., and Vijayakumar, B., "Illumination tolerant face recognition using phase-only support vector machines in the frequency domain," in *Pattern Recognition and Image Analysis*, pp. 66–73, Springer, 2005.

[51] Heusch, G., Rodriguez, Y., and Marcel, S., "Local binary patterns as an image preprocessing for face authentication," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 6–pp, IEEE, 2006.

[52] Hizem, W., Krichen, E., Ni, Y., Dorizzi, B., and Garcia-Salicetti, S., "Specific sensors for face recognition," in *Advances in Biometrics*, pp. 47–54, Springer, 2005.

[53] Hornberg, A., *Handbook of Machine Vision*. Wiley-VCH, 2007.

[54] Hutchinson, T., White Jr, K., Martin, W., Reichert, K., and Frey, L., "Human-computer interaction using eye-gaze input," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, no. 6, pp. 1527–1534, 1989.

[55] Job, A., "Driving without car keys.".

[56] Jobson, D. J., Rahman, Z.-u., and Woodell, G. A., "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *Image Processing, IEEE Transactions on*, vol. 6, no. 7, pp. 965–976, 1997.

[57] Jobson, D. J., Rahman, Z.-u., and Woodell, G. A., "Properties and performance of a center/surround retinex," *Image Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 451–462, 1997.

[58] Jobson, D., Rahman, Z., Woodell, G., Center, N., and Hampton, V., "Properties and performance of a center/surround retinex," *Image Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 451–462, 1997.

[59] Kang, J., Borkar, A., Yeung, A., Nong, N., Smith, M., and Hayes, M., "Short Wavelength Infrared Face Recognition for Personalization," *Image Processing, 2006 IEEE International Conference on*, pp. 2757–2760, October 2006.

[60] Kawato, S. and Ohya, J., "Automatic skin-color distribution extraction for face detection and tracking," vol. II, pp. 1415–1418, 2000.

[61] Kittler, J., Hilton, A., Hamouz, M., and Illingworth, J., "3d assisted face recognition: A survey of 3d imaging, modelling and recognition approachest," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pp. 114–114, IEEE, 2005.

[62] Kong, S. G., Heo, J., Abidi, B. R., Paik, J., and Abidi, M. A., "Recent advances in visual and infrared face recognitiona review," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 103–135, 2005.

[63] Kong, S., Heo, J., Abidi, B., Paik, J., and Abidi, M., "Recent advances in visual and infrared face recognition: a review," *Computer Vision and Image Understanding*, vol. 97, pp. 103–135, Jan. 2005.

[64] Kothari, R. and Mitchell, J., "Detection of eye locations in unconstrained visual images," p. 19A8, 1996.

[65] Kriegman, D. J. and Belhumeur, P. N., "What shadows reveal about object structure," *JOSA A*, vol. 18, no. 8, pp. 1804–1813, 2001.

[66] Lai, Z.-R., Dai, D.-Q., Ren, C.-X., and Huang, K.-K., "Multiscale logarithm difference edgemaps for face recognition against varying lighting conditions," *Image Processing, IEEE Transactions on*, vol. 24, no. 6, pp. 1735–1747, 2015.

[67] Lee, K.-C., Ho, J., and Kriegman, D., "Nine points of light: Acquiring subspaces for face recognition under variable lighting," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–519, IEEE, 2001.

[68] Lewis, J., "Fast normalized cross-correlation," *Vision Interface*, pp. 120–123, 1995.

[69] Li, S., Chu, R., Liao, S., and Zhang, L., "Illumination Invariant Face Recognition Using Near-Infrared Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[70] Li, W., Zhang, J., and Dai, Q.-h., "Robust blind motion deblurring using near-infrared flash image," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1394–1413, 2013.

[71] Lienhart, R. and Maydt, J., "An extended set of Haar-like features for rapid object detection," *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002.

[72] Lienhart, R. and Maydt, J., "An extended set of haar-like features for rapid object detection," in *Proceedings of International Conference on Image Processing*, vol. 1, pp. I–900, IEEE, 2002.

[73] Liu, D.-H., Lam, K.-M., and Shen, L.-S., "Illumination invariant face recognition," *Pattern Recognition*, vol. 38, no. 10, pp. 1705–1716, 2005.

[74] Liu, Z.-f., You, Z.-s., Jain, A. K., and Wang, Y.-q., "Face detection and facial feature extraction in color image," p. 126, 2003.

[75] Lucas, B. D. and Kanade, T., "An iterative image registration technique with an application to stereo vision," in *International joint conference on artificial intelligence*, vol. 81, pp. 674–679, 1981.

[76] MacDonald, L. W., *Digital heritage: applying digital imaging to cultural heritage*. Routledge, 2006.

[77] Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G., "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, vol. 964, pp. 965–966, Citeseer, 1999.

[78] Mikami, T., Sugimura, D., and Hamamoto, T., "Capturing color and near-infrared images with different exposure times for image enhancement under extremely low-light scene," in *IEEE International Conference on Image Processing*, pp. 669–673, IEEE, 2014.

[79] Newman, P., "More on brightness as a function of distance." January 2006.

[80] Pan, Z., Healey, G., Prasad, M., and Tromberg, B., "Face recognition in hyperspectral images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1552–1560, 2003.

[81] Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., and Toyama, K., "Digital photography with flash and no-flash image pairs," *ACM transactions on graphics*, vol. 23, no. 3, pp. 664–672, 2004.

[82] Prabhakar, S., Pankanti, S., and Jain, A. K., "Biometric recognition: Security and privacy concerns," *IEEE Security & Privacy*, no. 2, pp. 33–42, 2003.

[83] Prasad, M. and Tromberg, B., "Face Recognition in Hyperspectral Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1552–1560, December 2003.

[84] Qing, L., Shan, S., Chen, X., and Gao, W., "Face recognition under varying lighting based on the probabilistic model of gabor phase," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3, pp. 1139–1142, IEEE, 2006.

[85] Ramamoorthi, R., "Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1322–1333, 2002.

[86] Ramamoorthi, R. and Hanrahan, P., "On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object," *JOSA A*, vol. 18, no. 10, pp. 2448–2459, 2001.

[87] Savvides, M., Abiantun, R., Heo, J., Park, S., Xie, C., and Vijayakumar, B., "Partial & holistic face recognition on frgc-ii data using support vector machine," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pp. 48–48, IEEE, 2006.

[88] Savvides, M., Kumar, B., and Khosla, P. K., "Eigenphases vs eigenfaces," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. oo810–oo813, IEEE, 2004.

[89] Savvides, M., Kumar, B., and Khosla, P. K., "" corefaces"-robust shift invariant pca based correlation filter for illumination tolerant face recognition," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–834, IEEE, 2004.

[90] Scheenstra, A., Ruifrok, A., and Veltkamp, R. C., "A survey of 3d face recognition methods," in *Audio-and Video-Based Biometric Person Authentication*, pp. 891–899, Springer, 2005.

[91] Seo, H. J. and Milanfar, P., "Iteratively merging information from a pair of flash/no-flash images using nonlinear diffusion," in *IEEE International Conference on Computer Vision Workshops*, pp. 1324–1331, IEEE, 2011.

[92] Seow, M.-J., Valaparla, D., and Asari, V. K., "Neural network based skin color model for face detection," pp. 141–145, 2003.

[93] Shan, S., Gao, W., Cao, B., and Zhao, D., "Illumination normalization for robust face recognition against varying lighting conditions," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pp. 157–164, IEEE, 2003.

[94] Shashua, A., *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, Massachusetts Institute of Technology, 1992.

[95] Shashua, A. and Riklin-Raviv, T., "The quotient image: class-based rerendering and recognition with varying illuminations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 129–139, 2001.

[96] Shashua, A., "On photometric issues in 3d visual recognition from a single 2d image," *International Journal of Computer Vision*, vol. 21, no. 1-2, pp. 99–122, 1997.

[97] Short, J., Kittler, J., and Messer, K., "A comparison of photometric normalisation algorithms for face verification," in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 254–259, IEEE, 2004.

[98] Sim, T., Baker, S., and Bsat, M., "The cmu pose, illumination, and expression database," vol. 25, pp. 1615–1618, Dec. 2003.

[99] Sim, T., Baker, S., and Bsat, M., "The CMU pose, illumination, and expression database," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1615–1618, 2003.

[100] Sim, T. and Kanade, T., "Combining models and exemplars for face recognition: An illuminating example," in *Proceedings of the CVPR 2001 Workshop on Models versus Exemplars in Computer Vision*, vol. 1, 2001.

[101] Socolinsky, D. A. and Selinger, A., "A comparative analysis of face recognition performance with visible and thermal infrared imagery," tech. rep., DTIC Document, 2002.

[102] Socolinsky, D. A. and Selinger, A., "Thermal face recognition in an operational scenario," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–1012, IEEE, 2004.

[103] Socolinsky, D. A., Selinger, A., and Neuheisel, J. D., "Face recognition with visible and thermal infrared imagery," *Computer vision and image understanding*, vol. 91, no. 1, pp. 72–114, 2003.

[104] Sun, J., Sun, J., Kang, S. B., Xu, Z.-B., Tang, X., and Shum, H.-Y., "Flash cut: Foreground extraction with flash and no-flash image pairs," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[105] Thomas Yang, C.-H., Lai, S.-H., and Chang, L.-W., "Robust face matching under different lighting conditions," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 2, pp. 149–152, IEEE, 2002.

[106] Tomasi, C. and Manduchi, R., "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*, pp. 839–846, IEEE, 1998.

[107] Turk, M. and Pentland, A., "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[108] Turk, M., Pentland, A., Vision, Group, M., of Technology, M. I., and Laboratory, M., *Eigenfaces for Recognition*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[109] Turk, M. and Pentland, A., "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[110] Viola, P. and Jones, M., "Rapid object detection using a boosted cascade of simple features," *Proc. CVPR*, vol. 1, pp. 511–518, 2001.

[111] Viola, P. and Jones, M., "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[112] Viola, P. and Jones, M., "Rapid object detection using a boosted cascade of simple features," in *Proceedings of Computer Vision and Pattern Recognition*, vol. 1, pp. I–511, IEEE, 2001.

[113] Wang, B., Li, W., Yang, W., and Liao, Q., "Illumination normalization based on weber's law with application to face recognition," *Signal Processing Letters, IEEE*, vol. 18, no. 8, pp. 462–465, 2011.

[114] Wang, H., Li, S., and Wang, Y., "Generalized quotient image," *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2.

[115] Wang, H., Li, S. Z., and Wang, Y., "Face recognition under varying lighting conditions using self quotient image," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 819–824, IEEE, 2004.

[116] Wei, S.-D. and Lai, S.-H., "Robust face recognition under lighting variations," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, pp. 354–357, IEEE, 2004.

[117] Wilder, J., Phillips, P. J., Jiang, C., and Wiener, S., "Comparison of visible and infra-red imagery for face recognition," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pp. 182–187, IEEE, 1996.

[118] Xie, C., Savvides, M., and Kumar, B. V., "Quaternion correlation filters for face recognition in wavelet domain.," in *ICASSP (2)*, pp. 85–88, 2005.

[119] Xie, X. and Lam, K.-M., "An efficient illumination normalization method for face recognition," *Pattern Recognition Letters*, vol. 27, no. 6, pp. 609–617, 2006.

[120] Yi, D., Liu, R., Chu, R., Wang, R., Liu, D., and Li, S. Z., "Outdoor face recognition using enhanced near infrared imaging," in *Advances in Biometrics*, pp. 415–423, Springer, 2007.

[121] Yoon, S. M., Lee, Y. J., Yoon, G.-J., and Yoon, J., "Adaptive total variation minimization-based image enhancement from flash and no-flash pairs," *The Scientific World Journal*, vol. 2014, 2014.

[122] Yu, Z., Cheng, D., Khalil, I., Kay, J., and Heckmann, D., "Theme issue on adaptation and personalization for ubiquitous computing," *Personal and Ubiquitous Computing*, vol. 16, no. 5, pp. 467–468, 2012.

[123] Yukhin, A. and Klimov, A., "Methods and systems for detecting and recognizing an object based on 3d image data," 2007. US Patent 7,174,033.

[124] Zhang, L. and Samaras, D., "Face recognition under variable lighting using harmonic image exemplars," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, pp. I–19, IEEE, 2003.

[125] Zhang, L. and Samaras, D., "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 351–363, 2006.

[126] Zhang, T., Tang, Y. Y., Fang, B., Shang, Z., and Liu, X., "Face recognition under varying illumination using gradientfaces," *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2599–2606, 2009.

[127] Zhang, Y., Tian, J., He, X., and Yang, X., "Mqi based face recognition under uneven illumination," in *Advances in Biometrics*, pp. 290–298, Springer, 2007.

[128] Zhao, S. and Grigat, R., "An Automatic Face Recognition System in the Near Infrared Spectrum," *Proceedings of the 4th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2005)*, pp. 437–444, July 2005.

[129] ZHAO, W., CHELLAPPA, R., PHILLIPS, P., and ROSENFELD, A., "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[130] Zhao, W. and Chellappa, R., "Illumination-insensitive face recognition using symmetric shape-from-shading," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 286–293, IEEE, 2000.

[131] Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M., and Nahavandi, S., "Recent advances on singlemodal and multimodal face recognition: A survey," *Human-Machine Systems, IEEE Transactions on*, vol. 44, no. 6, pp. 701–716, 2014.

[132] Zhou, S. K., Aggarwal, G., Chellappa, R., and Jacobs, D. W., "Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, pp. 230–245, 2007.

[133] Zhu, Z. and Ji, Q., "Robust real-time eye detection and tracking under variable lighting conditions and various face orientations," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 124–154, 2005.

[134] Zhuo, S., Guo, D., and Sim, T., "Robust flash deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2440–2447, IEEE, 2010.

[135] Zou, X., Kittler, J., and Messer, K., "Motion compensation for face recognition based on active differential imaging," in *Advances in Biometrics*, pp. 39–48, Springer, 2007.