

2019-05-14

# Detecting Abnormalities in Thermal Pattern of Faces for Healthcare Applications

Ejindu, Oluchukwu Roseline

---

Ejindu, O. R. (2019). Detecting Abnormalities in Thermal Pattern of Faces for Healthcare Applications (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from <https://prism.ucalgary.ca>.  
<http://hdl.handle.net/1880/110344>

*Downloaded from PRISM Repository, University of Calgary*

UNIVERSITY OF CALGARY

Detecting Abnormalities in Thermal Pattern of Faces for Healthcare Applications

by

Oluchukwu Roseline Ejindu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL & COMPUTER ENGINEERING

CALGARY, ALBERTA

MAY, 2019

© Oluchukwu Roseline Ejindu 2019

## Abstract

In this work, we propose a novel method of applying deep learning technique in thermal image processing and analysis for healthcare application. It addresses detection of abnormal thermal patterns, thus identifying, in particular, patterns of elevated temperature that indicate fever, hypothermia and related abnormalities. Temperature estimation is performed based on the analysis of regions-of-interest from the thermal images of human faces. Another focus of this work is to investigate thermal effects of alcohol intoxication. We applied the deep learning approach on 16,000 usable images of 40 subjects from a publicly available Drunk-Sober database. Two Convolutional Neural Network architectures were investigated for the task of processing of two regions of interest - the forehead and the eyes. The accuracy of the neural network classifiers to predict subject's insobriety using the forehead and eye regions-of-interest reached 95.5% and 96.67%, respectively, compared to the best known results on the same data using a non-deep neural networks. To boost the accuracy of classification, both the feature-level and the score-level fusion were applied as well, thus improving the accuracy to 96.92%.

## Acknowledgements

First and foremost, I would like to appreciate my supervisor Dr. Svetlana Yanushkevich, for her supervision and guidance towards the progress of this work. Her patience was profound in every stage of this program and her mentorship and guidance has opened a whole new world for me.

I would also like to express my profound gratitude to my friends and family members especially my husband, Nelson Obi-Alago, for supporting me throughout this journey. Your support and encouragement gave me the strength I needed to pursue this degree. And to my wonderful children, Olivia and Alexander Obi-Alago, thank you for being my distraction and my determination.

Lastly, I would like to acknowledge institutions that provided the financial resources for this research - MITACS, Calgary Drop-in Center and the Queen Elizabeth II scholarship.

# Table of Contents

Abstract . . . . .	ii
Acknowledgements . . . . .	iii
Table of Contents . . . . .	iv
List of Tables . . . . .	vi
List of Figures and Illustrations . . . . .	vii
List of Symbols, Abbreviations and Nomenclature . . . . .	viii
1 Introduction . . . . .	1
1.1 Motivation . . . . .	1
1.2 Background . . . . .	3
1.3 Objectives . . . . .	4
1.4 Contributions . . . . .	6
1.5 Thesis Outline . . . . .	7
2 Literature Review . . . . .	9
2.1 Thermal Imagery for Health Diagnosis . . . . .	9
2.2 Thermal Imagery for Detection of Insobriety . . . . .	11
2.3 Face Detection . . . . .	12
2.4 Thermal Image Registration . . . . .	13
2.5 Classifiers . . . . .	14
2.5.1 Support Vector Machine . . . . .	14
2.5.2 Deep Learning . . . . .	15
2.5.3 Training Methodologies . . . . .	17
2.6 Fusion Methodologies . . . . .	18
2.7 Conclusion . . . . .	19
3 Face Thermal Patterns Analysis . . . . .	21
3.1 Infrared Thermography . . . . .	21
3.2 Working Principle of an Infrared Camera . . . . .	23
3.3 Temperature Extraction and Classification . . . . .	24
3.4 Experiment . . . . .	25
3.5 Analysis of Experimental Result . . . . .	27
3.6 Conclusion . . . . .	28
4 Insobriety Detection using Thermal images and Machine Learning . . . . .	31
4.1 Data sets for the experiment . . . . .	31
4.2 The proposed system . . . . .	32
4.3 Face Detection and Region-of-Interest Extraction . . . . .	35
4.4 Pre-Processing . . . . .	38
4.5 Implementation using Support Vector Machines . . . . .	39
4.5.1 Feature Extraction using Local Difference Patterns . . . . .	39
4.5.2 Support Vector Machine for classification . . . . .	40
4.5.3 Experiments . . . . .	43
4.6 Convolutional Neural Networks . . . . .	45
4.6.1 Layers of the CNN . . . . .	46
4.6.2 Error Function . . . . .	53

4.6.3	Proposed Insobriety Detection System using CNN . . . . .	56
4.6.4	Experiments . . . . .	57
4.6.5	Fusion . . . . .	64
4.6.6	Performance evaluation and comparison to other works . . . . .	65
4.6.7	Real-life Experiments . . . . .	68
4.6.8	Results . . . . .	70
4.7	Conclusion . . . . .	70
5	Conclusion and Future work . . . . .	75
5.1	Summary of Results . . . . .	75
5.2	Suggestions for future work . . . . .	76
	Bibliography . . . . .	78
A	Copyright Permissions . . . . .	86
A.1	UPatras Database . . . . .	86
B	Ethics Approval . . . . .	87

## List of Tables

3.1	Comparison of the Temperature measurement from the Camera and the Thermometer. . . . .	30
4.1	Comparison of the accuracy of the proposed approach with other methods . . . . .	44
4.2	Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Number of epochs= 30; filter size = $5 \times 5$ ; Learning Rate = 0.0001) . . . . .	59
4.3	Tuning of hyper-parameters of the network built with forehead Images: 3 parameters are constant (Training/Testing Set = 90/10; filter size = $5 \times 5$ ; Learning Rate = 0.0001) . . . . .	61
4.4	Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Training/Testing Set = 90/10; Number of Epochs = 30; Learning Rate = 0.0001) . . . . .	61
4.5	Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Training/Testing Set = 90/10; filter size = $5 \times 5$ ; Number of Epochs = 30) . . . . .	61
4.6	Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Number of epochs= 30; filter size = $3 \times 3$ ; Learning Rate = 0.0001) . . . . .	62
4.7	Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; filter size = $3 \times 3$ ; Learning Rate = 0.0001) . . . . .	63
4.8	Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; Number of Epochs = 30; Learning Rate = 0.0001) . . . . .	63
4.9	Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; filter size = $5 \times 5$ ; Number of Epochs = 30) . . . . .	63
4.10	Comparison of the accuracy for different methods of sobriety classification . . . . .	68

## List of Figures and Illustrations

1.1	Architecture of the proposed system for thermal image analysis . . . . .	5
3.1	Flow chart of the proposed algorithm for fever detection . . . . .	26
3.2	Infrared Camera reading versus forehead thermometer reading, in Farenheit . . . .	28
3.3	Bland-Altman Plot of the temperature measures by the thermometer and the thermal camera, in Farenheit . . . . .	29
4.1	Overall system architecture of the proposed insobriety detection system. Input images are fed into the system for face detection, followed by pre-processing, feature extraction and classification. This architecture is used for both training on database images, and testing on probe images. . . . .	33
4.2	Rectangular templates used in Viola-Jones algorithm:(a) and (b) are two-rectangle feature, (c) is a three rectangular feature, and (d) is a four-rectangle feature. The illustration is inspired by [?] . . . . .	36
4.3	Integral Image using the sum of the pixels. The value of the integral pixel in location 1 is the sum of the pixel values in A, the value in location 2 is the sum of the pixels in A and B. Similarly, the values for locations 2 and 4 are the sum of A and C, and the sum of A,B, C and D, respectively . . . . .	37
4.4	Creation of a simple LBP . . . . .	39
4.5	Histograms of the LDP for Drunk and Sober subjects, respectively . . . . .	40
4.6	Support Vector Machine illustration from [78] . . . . .	42
4.7	Confusion Matrix for the classification of subjects . . . . .	44
4.8	Architecture of a Convolutional Neural Network . . . . .	46
4.9	Conv Layer:A $3 \times 3 \times 1 \times 2$ Conv layer with $S = 1$ , $P = 1$ is applied to a $3 \times 3 \times 1$ input matrix, each filter convolves the input separately, resulting in a $3 \times 3 \times 2$ output matrix . . . . .	48
4.10	Conv Layer:A $3 \times 3 \times 2 \times 1$ Conv layer with $S = 1$ , $P = 1$ is applied to a $3 \times 3 \times 2$ input matrix. Each convolves its corresponding channel, and are summed to form a $3 \times 3 \times 1$ output . . . . .	48
4.11	Pooling Layer: A $2 \times 2$ pooling layer with a stride of 2 and pad of 0 is applied to a $4 \times 4$ input matrix, resulting in a $2 \times 2$ output matrix . . . . .	50
4.12	Rectified Linear Unit that zeros any negative inputs . . . . .	51
4.13	Log-loss when true label as explained in [84] . . . . .	54
4.14	CNN Training in Matlab: The accuracy increases as the training goes through the epochs and updates the weight accordingly while the loss is going down to zero . .	60
4.15	Architecture of the system showing the stages where fusion was implemented . . .	64
4.16	Confusion Matrix for each experiment for 40 iterations . . . . .	66
4.17	Example of a subject standing in front of the infrared camera as the infrared image recording was taken . . . . .	70
4.18	The Interface of the application after the thermal image is uploaded . . . . .	71
4.19	The Interface of the application with the necessary information for the user . . . .	72



## List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
2D	Two-Dimensional
3D	Three-Dimensional
BP	Back Propagation
BNorm	Batch Normalization
CNN	Convolutional Neural Network
Conv Layer	Convolution Layer
DI	Drop-in center
FNR	False Negative Rate
FPR	False Positive Rate
FC	Fully Connected
GAN	Generative Adversarial Network
GEI	Gait Energy Image
LR	Learning Rate
Relu	Rectified Linear Unit
SGD	Stochastic Gradient
SVM	Support Vector Machine
TPR	True Positive Rate
Ucalgary	University of Calgary
UPatras	University of Patras

# Chapter 1

## Introduction

This thesis is a result of the feasibility study conducted in collaboration with the Calgary Drop-In & Rehab Centre Society in Calgary (also called the DI Centre). This study was focused on the task of detecting potential medical emergencies of the DI Centre clients in a non-invasive, access check-in manner, using thermal body metrics (a type of biometrics) at distance, i.e. via infrared camera. Two problems within this study were investigated:

- How accurate would be the body temperature estimation (via face area) using infrared camera and a particular access point setup given certain restrictions?
- Would the usage of thermal video be helpful in detecting abnormal temperatures caused by fever or other medical condition such as insobriety?

### 1.1 Motivation

The Calgary DI Centre is one of the largest homeless shelters in North America. It has been using biometrics such as fingerprints for the access point for few years. The Centre serves about 2,000 homeless people per day, and identity management and security is in paramount demand there. The shelter has multiple groups of clients, and providing appropriate services to various groups by identifying the known or occasional clients (such as families fleeing domestic violence, trade workers who need temporary shelter at night, or drug addicts and heavily intoxicated occasional visitors) is an important part of the shelter entrance and security system.

As of 2018-2019, the IT Department of the DI Centre is conducting a pilot project on using face biometrics (face recognition) at a prototype 'kiosk'. Face recognition of frequent client (for now, volunteers only) is performed at this one prototype setup using access to Microsoft face

recognition tools. The IT Department is considering adding other ways to assist in client services, such as detecting early signs of illnesses or insobriety. One of the choices for a relevant study was to investigate potential use of thermal camera that could be later installed in the same prototype 'kiosk'. Usage of early detection of extreme fevers, hypothermia (frostbite) or insobriety would allow the DI Centre personnel to assist client and call medical emergency service as needed.

Homelessness is a social phenomenon related to human dynamics in the changing world, economical and financial issues, as well as potential threats such as epidemics and natural disasters [1]. Homelessness is a part of a more complex problem: according to the UN, over 1 billion people live in inadequate shelter and over a 100 million people live in conditions classified as homelessness [2].

Shelters and Drop-In Centres are social services or charities that provide services to homeless people, as well as people who are temporarily dislocated due to emergencies or natural disasters. Such Centres need a highly efficient infrastructure for managing clients and protecting personnel. However, the majority of Shelter clients do not possess any IDs, or carry IDs that may not be validated. Using biometric-based identity management is one of the solutions to this problem. For shelter clients, biometric IDs can empower them to exercise their fundamental economic and social rights, as well as enabling Shelter governing agencies to validate and streamline their services.

The DI Centre formulated their current problems and challenges in the area of secure access to the building, as well as building facilities security as related to clients' activities and access to services:

- What biometric solutions can be recommended and integrated in the DI into access control such that those technologies are acceptable, non-invasive, usable, and privacy and trust preserving? Current fingerprint scanning deployed in the DI does not seem to satisfy the above criteria; for example, fingerprint identification has low accuracy because some of the clients have worn-out or unclean fingers.
- How to enhance or update the existing technologies that would improve the safety

of citizens and personnel? Currently, no solutions are known for risk assessment or warning in the DI, except for manual analysis of video-feed from CCTV cameras by security officers. In addition, the DI center faces problems as huge as unexpected deaths of clients, so an early warning system is of utmost importance in the center to reduce the problem of unexpected instant deaths of the clients of the shelter.

This thesis reports the preliminary results of a pilot project initiated by the Calgary Drop-In & Rehab Centre Society, and conducted in cooperation with the Biometric Technology Laboratory, University of Calgary, Canada. This study focused in the following tasks:

- Body temperature estimation using thermal imaging of face of a subject, acquired using an infrared camera in a particular access point setup given certain restrictions.
- Feasibility study detecting abnormal thermal pattern caused by certain medical condition such as insobriety.

## 1.2 Background

Face biometrics is a non-invasive, contactless, and commonly accepted type of biometrics for access points and surveillance networks. The challenges of facial recognition include environmental variations (luminance, day and night time), pose, accessories etc. However, face imaging in visible and cross-spectrum (such as thermal) bands, can provide useful features assessed at a distance, such as body temperature, facial expression, age, gender and, possibly, identify drug and alcohol abuse cases [3].

Infrared Thermography (IRT), also known as thermal imaging, is the science of detecting the infrared radiation emitted from a body and extracting the temperature value. IRT is a biometric technology that is contactless and poses lesser privacy concerns compared to video and photograph recording, since thermal images does not allow directly identify subject based on the appearance. IRT as a form of biometrics has been used for monitoring a patient's health status [4].

Outbreaks of contagious diseases such as severe acute respiratory syndrome (SARS) have occurred in some countries, reinforcing intensive IRT based screening at the borders especially at international airports [5]. IRT screening was used for detection of febrile clients because fever is one of the most noticeable signs of infections such as avian influenza, influenza A virus subtype H1N1, and SARS [6].

Recently, a body of research on using thermal image analysis of face to detect alcohol intoxication appeared starting with [7].

This thesis revisits the major results achieved in this area, and specifically focuses on detecting thermal abnormalities such as elevated temperature, assessing the accuracy of measurements given the restricted data, and, finally, focusing on thermal pattern indicative to insobriety, given available data. We further investigate the feasibility of the above approaches for the specific application in the Calgary DI Centre.

### 1.3 Objectives

The main hypothesis (‘thesis’) of this study is formulated as follows: “The thermal body pattern abnormalities caused by certain medical conditions, which are correlated to specific temperature distributions, can be detected and analyzed using infrared images/video of subject’s face in close proximity to the thermal camera. The applications involves detecting abnormal temperatures caused by fever or other medical condition such as insobriety”.

The main objectives of the thesis are as follows:

- Design a semi-automated system capable of a feasible estimation of the subject facial temperature. The automation is important for scalability to real-life implementations, where no manual intervention is possible.
- Create an implementation which continuously records infrared images of a subject face and predicts the level of sobriety.

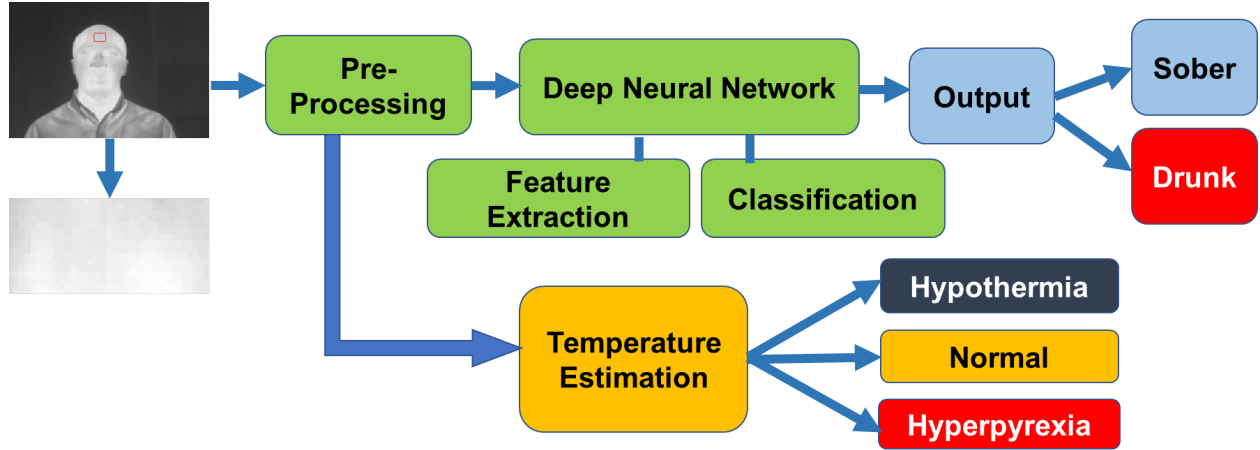


Figure 1.1: Architecture of the proposed system for thermal image analysis

- Utilize state-of-the-art techniques, such as feature extractors and classifiers based on machine learning, in particular, Convolutional Neural Network (CNN) and Support Vector Machines (SVM).
- Test the proposed approach on an available data sets such as the University of Patras ‘Sober-Drunk’ Database. Organize the experiments and present the results in a way that is unbiased, and can be easily compared with by other researchers.

Figure 1.1 shows the general architecture of the system we are proposing for the analysis of thermal images.

This system includes the following components:

- The input data is an infrared image of the subject’s face acquired at close proximity to the infrared camera.
- The images are passed to the Pre-processing stage that ensures the frames are uniform and normalized.
- The next stage is two-fold: (a) a temperature estimation using the curve-fitting model, or (b) feature extraction and classification of sobriety using the deep neural network.

- the output data is the estimated temperature (which can be further used to classify the subjects as having normal temperature, febrile or hypothermic), and the subject label (either ‘sober’ or ‘drunk’).

## 1.4 Contributions

The contributions made during the course of this research are reflected throughout the following publications:

- O. Obi-Alago, S. Yanushkevich, “Using Thermal Signature Abnormalities in health-care,” in 19th Annual Alberta Biomedical Engineering Conference, 2018.
- O. Obi-Alago, P.Kozlow, A. Noor, S.C. Eastwood, H.M. Wetherley, and S. Yanushkevich, “Pilot Project: Biometric-Enabled Risk Assessment for City Emergency Shelters,” the IEEE Transactions on Computational Social Systems, submitted.
- O. Obi-Alago, S. Yanushkevich, “Thermal Image Processing and Classification using Deep Neural Networks,” in the Eight International Conference on Emerging Security Technologies, UK 2019, submitted.

The first publication mainly focused on developing the system for detecting illness and also alcohol detection using CNN. Two networks was built using two regions-of-interest on the images, and the accuracies of classification was estimated.

The second paper describes a multi-biometric enabled identity management and risk assessment system for clients that are part of an emergency shelter. As part of a smart-city concept, this system proposes to improve authentication by fusing many data channels such as gait, face, and thermal imaging. For this system an individual’s physiological as well as behavioral features can be used to grant permission to access certain services. Based on Bayesian decision networks the proposed system is capable of providing a risk assessment which is based on available biometric data.

In the third publication, an extra step was taken to fuse the results from the first paper in order to develop a more accurate system for classification. This yielded a higher classification accuracy than the individual networks proposed in the first paper.

## 1.5 Thesis Outline

The rest of this thesis is organized into the following four chapters. Chapter 2: Literature Review contains a relevant literature review on face detection and extraction methods, the use of IRT in detection fever and sobriety of individuals and then the classification methodologies on gait recognition, virtual modeling, databases available, and a detailed description about some common algorithms and techniques currently being used. Chapter 3, Face Thermal Patterns Analysis, describes how infrared images can be used to detect a sick subjects and the experiments that was conducted on this approach. Chapter 4, Insobriety Detection using Thermal images and Machine Learning, describes the proposed system of detecting sobriety in subjects. In addition, an application was developed and experimentation was conducted to test the feasibility of the system. Lastly, Chapter 5: Conclusions & Future work summarizes important observations made throughout the study, and presents ideas for future work.





## Chapter 2

### Literature Review

Thermal imaging has been used for decades to analyze temperature distribution of objects of interest. In this section, an overview of the background information and breakthroughs on the use of thermal imaging in healthcare applications such as human body temperature estimation and analysis of thermal pattern of faces will be discussed.

#### 2.1 Thermal Imagery for Health Diagnosis

There is ongoing research on application of thermal, or infrared (IR) cameras to capture and record temperature variations on the skin for medical diagnostic purposes. The human body is homeothermic, i.e self-generating and regulating a stable internal body temperature necessary for survival.

The temperature of the human face is about 300 K, or 36.6°C and, according to Wien Law, radiates as 300 Kelvin black-body (with wave length of 9.5 microns). Recall that Wien's law states that the black-body radiation curve for different temperature peaks at a wavelength is inversely proportional to the temperature. That peak is different for different temperatures as stated by the Planck radiation law, which describes the spectral brightness of black-body radiation as a function of wavelength at any given temperature. It explains the shift of the spectrum of black-body radiation toward shorter wavelengths as temperature increases. In other words, all objects emit a certain amount of black body radiation as a function of their temperature, and the higher an object's temperature, the more infrared radiation is emitted as black-body radiation. It works even in total darkness because ambient light level does not matter.

Thermal sensors, or cameras, help record the temperature distribution of an objects in the camera view that emits infrared light. The focused light is scanned by a phased array of infrared-detector elements. The detector elements create a temperature pattern called a thermogram. A

usual infrared detector resolution, or the number of pixels, is  $480 \times 640$  or less.

The thermogram created by the detector elements is translated into electric impulses that are sent to a signal-processing unit that forms a data for the display.

The thermal cameras that are best for human body thermal pattern analysis operate in the wavelengths range from 7.5 to 13.0 microns (mid- and long-wave infrared). Note that 9.5 microns for 300 Kelvin black body corresponds to that of human skin. Raw thermal images are gray-scale images, usually with 256 levels of intensity. Pixel intensity of the thermal images can be mapped onto the temperature maps. Thermal images does not provide an absolute temperature in every pixel of the image, and rather must be compensated in a software way.

In practice, using as reference a local temperature measurement on the object or a subject is the simplest and most effective solution. Thermal cameras labeled “R” are radiometrically calibrated. Using such cameras enables the capture of absolute temperature in every pixel of an image. They save their images in RJPG (radiometric JPG) format, a .jpg image with temperature data embedded in every pixel. However, even with radiometrically calibrated cameras, the differences in illumination, the properties of the object surface (material, texture), also affect the thermal emissivity. Such bias can be compensated by the software, as well as at the hardware level.

The correspondence between the image pixel intensity and temperature of a subject on the thermogram can be used to extract useful information on the metabolic and vascular activity to detect abnormal changes in physiology.

Thermal imaging is used to study a number of diseases, especially ones that cause inflammation or where blood flow rate changes. Currently, a lot of physicians employ thermal imaging cameras to detect some medical conditions, such as arthritis, repetitive strain injury, muscular pain, and circulatory problems. For example, [8] reported that the surface temperature of an arthritic joint is related to the inflammation on the joint.

Thermal imaging has also been used to diagnose osteoarthritis, which is a type of joint disease that results from breakdown of joint cartilage and underlying bone. In [9]; it was concluded that

there are temperature changes associated with degenerative changes caused by osteoarthritis.

In [10], detection of abnormalities in peripheral circulation using thermography was studied. The paper investigated circulatory disorders, such as Raynaud's phenomenon, or hand arm vibration syndrome, that causes damage to small blood vessels from exposure to vibrating machinery. It was shown that the effect of local blood circulation on skin temperature can be assessed by thermal imaging.

Infrared imaging was also applied to diagnose cancer, especially breast cancer. Paper [11] was one of the first to report the use of surface temperature measurements as a possible tool for breast cancer diagnosis. Paper [12] reviews the progress that has been made in using thermal imaging to detect breast cancer over the past three decades.

## 2.2 Thermal Imagery for Detection of Insobriety

Insobriety is a physiological condition caused by alcohol affect on body function. The use of thermal imagery in detection of alcohol intoxication lies in the fact that the blood flow rate in blood vessels on the face changes when alcohol is consumed and, consequently, causes a change in the temperature distribution [13].

In [7], it was suggested that features derived from the pixel values within regions-of-interest on the face can be used to detect alcohol intoxication. The approach called Local Difference Pattern (LDP) calculated using the pixels on the thermal image of a face was applied to identify intoxicated subjects [14]. This approach considered a difference between the neighboring pixels, and used it for classification. Research by [15] shows that there are specific regions of interest on the face that are better suited for detecting intoxication. It is found that the thermal behavior of forehead and nose regions of the face change after alcohol consumption. The eye region was also investigated in the research on this subject. The temperature distribution on the eyes of sober and drunk persons was studied in [16]. It was observed that the temperature difference between the sclera and the iris increases when an individual consumes alcohol. The information derived from this difference in

temperature was used to classify a subject as being intoxicated or sober.

In a research conducted by [13], the blood vessels on the face was separated and isolated from the rest of the information on the image of the face by applying morphology on the diffused image. To extract the blood vessels from the thermal image, the top-hat transform and anisotropic diffusion was applied to the image. Top-hat transform, which is an operation that extracts small elements and details from images, was used to isolate hot or cold features in an image of a specific size, and a non-linear anisotropic diffusion was applied to enhance the vessels on the images. The increased blood vessel activity was utilized for classification of sobriety levels of subjects.

Neural network classifier was applied by [17] to discriminate thermal images of “sober” and “drunk” subjects. Two approaches based on neural networks were studied. In the first approach, different neural structures were used for different locations in the thermal image of the face, and the convergence capabilities of the network were monitored. A successful convergence characterized the corresponding location of the face as being a good candidate for intoxication identification. According to the second approach, a single neural structure was trained with data from the thermal images of the whole face.

## 2.3 Face Detection

Detection of face in a thermogram is an important part of finding facial temperature abnormalities.

The most common face detection algorithm, applied on video images, is Viola-Jones algorithm [18]. It filters the face using a set of Haar-like features (moving windows). The calculations are simplified by using an approach called an integral image. The feature filter extracts the patterns such as eyes, eyebrows and facial borders. These filters, also called “weak” classifiers, are then boosted with AdaBoost which selects the most discriminative features. Finally, a cascade of classifiers selects the best detected location of the face. This method works extremely fast, but can have low detection rates on rotated images or different lighting conditions.

Another unified model for face detection is presented in [19]. It is based on a search algorithm

that uses a mixture of trees with a shared pool of part of the facial templates. It models every facial landmark as a part and use global mixtures to capture topological changes due to viewpoint.

Authors in [20] proposed a face detection approach specifically for infrared images. They approached the face detection from a statistical point of view. They analyzed the “hot” parts on the face that are the tissue area that have a lot of blood vessels, such as the periorbital and the forehead. In the contrary, the “cold” parts have fewer blood vessels; it includes cheeks and nose. This feature casts the human face as a bimodal distribution entity. The background of a thermal image can also be described by a bimodal distribution. It typically consists of walls (cold objects) and the upper part of the subject’s body dressed in clothes (hot objects). The bimodal distributions of the face and background vary over time, due to thermo-physiological processes and environmental noise respectively. Using these distributions, a face region in a thermal image can be detected.

Deep learning approach such as Convolutional Neural Network (CNN) have also been used for face detection, as suggested in [21]. The CNN cascade operates at multiple resolutions, rejects the background regions in the fast low-resolution stages, and carefully evaluates a small number of challenging candidates in the later, high-resolution stage. To improve localization effectiveness and reduce the number of candidates at later stages, they introduced a CNN-based calibration stage after each detection stage in the cascade.

## 2.4 Thermal Image Registration

Image registration has been a challenging task in the field of image processing. The purpose of image registration is to account for any lateral or vertical movement from the subject that takes place during image collection. There are various techniques available for image registration for medical images and for images in biometric applications. The author in [22] proposed an intra subject image registration of the infrared images. In [23], the authors discussed a Multi-Modality Image registration (MMIR) of visible and infrared facial regions for the accurate detection of the Canthi (eye corner) regions. According to the study, Canthi regions provide the most accurate es-

timates of the core body temperature. The challenge is that it is very difficult to accurately identify Canthi regions from an infrared image. They used a two-step coarse to fine registration for both the visible and infrared images, achieving a reasonable registration accuracy and the temperature reading accuracy of  $0.10 \pm 0.09^{\circ}F$ .

## 2.5 Classifiers

In deep learning, classification is the process of identifying which set or categories a new observation belongs to, on the basis of a training set of data containing observation whose category is known. Classifiers are used to perform the classification and this is done after the extracted features are passed onto them. The classifiers learns from the training data and predicts the class of the new data from the testing set. The choice of the classifier depends on a number of factors: size of the data, whether the data is labeled or not, and if the classes are discrete or continuous.

A simple K-Nearest-Neighbor, for example was not used in our study, because it is generally inefficient for the large data set size [24]. The UPatras database that was used in this work is properly labeled and has thousands of images making the SVM and CNN the preferred classifiers.

### 2.5.1 Support Vector Machine

SVM is a supervised learning model based on the principle of regression analysis. It was first introduced in 1992 [25], and was used to train and classify unseen data. SVM have been used to solve a both pattern recognition and estimation problems. It has also been applied to dependency estimation, forecasting and constructing intelligent machines [26, 27]. SVMs have the capacity to work with very large feature space, due to the generalization principle which is based on the Structural Risk Minimization Theory i.e., the algorithm is based on guaranteed risk bounds of statistical learning theory [28]. SVMs as a classification method also has an application in medical research. K.Polat [29] applied Least Square SVM (LSSVM) in solving the problem of breast cancer diagnosis, with classification accuracy of 98.53%.

Training an SVM is simple, as there is only one parameter for linear SVM, and two in RBF SVM. It does not have the problem of local minima, unlike some other learning classifiers such as neural networks. Rather, SVM calculates a decision boundary based on the parameters selected.

The most common form of SVM is the C-support vector classification (C-SVC), which classifies data with a cost parameter  $C$ . The value of  $C$  is used to determine the number of misclassified data samples and the decision boundary. If some of the outlier samples are left to be misclassified by the SVM, it allows the decision boundary to be smooth and simple, thus classifying other unseen data more accurately. The parameter  $C$  is set by the user, and is generally grid searched in order to generate the most optimal value.  $n$ -SVC is an extension where the parameter  $n$  controls the number of allowed support vectors. A support vector is a data point close to the decision boundary, which SVM uses to optimize the data class separation.

SVM algorithms use a set of mathematical functions called kernels for the transformation of input to the required form. Multiple kernels for SVM have been proposed: linear, Radial Basis Function (RBF), polynomial and Gaussian kernel [25]. The simplest kernel is the linear kernel which is used to classify linear data. When the data set gets more complex (non-linear), RBF kernel is commonly used. It has shown a better classification accuracy than the linear kernel [30]. Polynomial and Gaussian kernel also classifies non-linear data well. It was used in medical research for the classification of normal and malignant tissues [31].

### 2.5.2 Deep Learning

Deep learning (also known as deep structured learning) is a part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Deep learning can be supervised, semi-supervised or unsupervised. Deep learning uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input. The networks have many more layers (increased depth) than their traditional counterpart. The features learned from deep networks have proven to classify better than shallow networks. Types of deep learning networks include deep



auto-encoders, deep Bayesian networks, and convolutional neural network (CNN). CNN is a class of deep neural networks, most commonly applied to analyzing visual imagery. They take into account the spatial information present in images by using layers of filters instead of individual weights. CNN is a variation of the traditional neural network, but it uses filters instead of weights which allows for many more layers because there are less overall parameters to be trained.

CNN was first introduced by Yann LeCun in 1994 when he created LeNet used for automatic document recognition [32]. The LeNet5 architecture explored the fact that image features are distributed across the entire image, and convolutions with learnable parameters are an effective way to extract similar features at multiple locations. At the time there was no GPU to help training, and CPUs were slow, and previous solutions treated each pixel as a separate input of a large multi-layer neural network. LeNet5 proved that those should not be used in the first layer, because images are highly spatially correlated, and using individual pixel of the image as separate input features would not take advantage of these correlations.

CNN became more popular with the introduction of AlexNet [33], a winner of the ImageNet Large Scale Visual REcognition challenge 2012 (ILSVRC). In 2014, another network was reported that was deeper than AlexNet and performed better by 6% [34].

CNN has been used for many applications. Facial Recognition has been done with CNN on large databases and it has generally outperformed the common filter based methods [35, 36, 37, 38, 39]. FaceNet was created by Google, and VGG network were made for face detection and showed over 99% detection rate on Faces-in-the-wild images [40, 41]. CNNs have been used to solve challenging computer vision tasks, especially in object detection and classification, text recognition, sign recognition and scene understanding [42].

In 2016, Microsoft and Google released ResNet and Inception, which is an extension of CNNs called directed acyclic graph neural networks (DAGNNs) [43, 44]. The idea of DAGNNs is that the data does not need to flow in one path through the network, its layers can be passed over during training by inserting "shortcuts". A very deep version with 1,000 layers was created, but it

performed worse than the smaller networks, showing there is a limit to accuracy gains by simply creating a deeper network. ResNet-152 and its extensions currently show superior performance on the ImageNet data set.

A lot of researchers are working on optimizing deep learning and creating extensions and additions to improve the structure and hyper parameters. One major drawback in CNN is having a lot of parameters that need to be fine-tuned, which affects the accuracy of the network. Ensemble networks have been proposed as a solution to this, where each network has slightly different values for the hyper-parameters. Every network classifies a portion of the data, and the results are fused together to produce one answer.

### 2.5.3 Training Methodologies

Training a classifier in CNN is a important. When the network is trained properly, a high accuracy of classification is achieved. The input data is usually separated into a training set and a testing set, such that no same images are put in both the training and testing set, to avoid having false accuracy. Separating the training and testing set data is done in two ways - image level and subject level separation.

In image level separation, all the data is merged and treated as one disregarding the label on the images. Then this merged data is split into training and testing set. The training set is used for training the network and the testing set is used for testing and cross-validation to evaluate the performance of the system.

In subject level separation, the images are split according to their labels. In this scenario, the subject in the training images is completely different from the subject in the testing set. This approach is very suitable for scenarios where the test subjects are most likely not known to the network. In should be noted that in emergency shelter scenario that this study is dedicated to, it is very likely that the probe subject was been seen by the system before (was not included in the training). Therefore, in our study we chose to use the subject level separation. In addition, if the database contains very similar images, subject level separation makes the classification simpler

because it reduces the amount of images to match, unlike in the image level separation where all the images are mixed together.

Cross-validation is a strategy used in designing a more robust solution for the accuracy of a network. The data set is broken into  $K$  sets, where  $K$  value is the number of subjects. Some of the sets are used for training and the remaining set is used for classification. For the UPatras database, there are 40 subjects, and we chose 40-fold cross validation. This is boundary case of validation, also known as “leave-one-out” approach. This method was chosen in this study as it provides a robust accuracy, and allows to estimate accuracy for each subject individually.

Another measurement most commonly used when the class sizes are unbalanced is precision. Precision is the number of correct predictions of one class divided by the total number of images in the class. Given a problem with  $c$  classes, there will be  $c$  precisions. The overall accuracy is then calculated as the average of the precisions. The advantage to using this over classification accuracy is that it properly represents the accuracy of each class, regardless of its size.

## 2.6 Fusion Methodologies

Fusion in pattern recognition is a method of merging results from various classifier, or experimental setups, and allows to achieve a better accuracy. Instead of relying on just one method, fusion gives us an opportunity to exploit the strength of both methods, and increase the accuracy of classification.

Multimodal biometric systems that uses several biometrics (for example, face and iris, gait and fingerprint) generally apply fusion of different types of biometrics at various levels. Single (unimodal) biometrics tend to suffer due to noise sensitive data, intra-class variations, inter-class similarities, etc. The purpose of fusing some unimodal together biometrics is to increase the robustness and resistance to spoof attacks [45]. There are various levels at which fusion can be performed [46]:

- Sensor level - where two or more datasets acquired using different sensors are com-

bined.

- Feature level - which focuses on combining critical features extracted from two or more modalities.
- Score level - where the scores obtained from various matchers/classifiers are combined.
- Decision level - where the decisions of classifiers for various systems are combined.

A lot of research has been ongoing in the field of fusion methodologies in an attempt to improve biometric classification systems. These works include fusing different biometric modalities at different levels. For example, the performance of iris recognition was improved using score level fusion for scores derived for both textural and topological based matchers [47]. [48] proposed to combine the iris and face at features level using Complex Common Vector (CCV) algorithm, and was able to achieve better accuracy. Experimentation by [49] using score and signal level fusion in iris recognition system yielded a better accuracy, although score level fusion performed slightly better than the signal fusion.

In the field of security using facial recognition systems, fusion has also been proven to boost system accuracy. In the research conducted by [50], higher security using face matching was achieved by using feature level fusion. Decision level fusion of fingerprint and finger-vein as presented by [51] shows the increased accuracy. The recognition was done using minutiae points on ridge area, and finger-vein recognition was done using Local Binary Patterns method.

## 2.7 Conclusion

In this chapter, we revised the ongoing research on the use of thermal image for health diagnosis and detecting alcohol intoxication. We discussed the use of thermal images as a non-invasive method of diagnosis of certain thermal abnormalities that may indicate health emergencies and illnesses. The major advantage of infrared imagery is the fact that it is non-invasive and provides a

relatively good accuracy in temperature estimation, as well as thermal regions detection or classification.

The need of properly registered images was also discussed, and different methods of achieving same were presented. The classification of selected types of thermal image abnormalities, such as the one indicative to subject's insobriety, as reported in literature, were reviewed. Two specific approaches were the focus of this review: SVM and CNN. In addition, a fusion approach to achieve a higher accuracy of classification was revised. Fusing different features prior to classification, as well as fusion of few classifier results is generally shown to boost the accuracy of classification.

The reviewed approaches were selected to be used and/or modified for the purpose of the experimental study presented in the next chapters.

## Chapter 3

### Face Thermal Patterns Analysis

This Chapter revisits the approaches to analysis of thermal pattern distribution on human face using remote measurements by infrared (IR), or thermal, sensors known as IR video cameras. Specifically, we focus on detecting regions-of-interest (ROI) for detecting abnormal temperatures such as elevated above normal. These reading will be later used in Chapter for detecting the pattern on the thermal face image indicative to insobriety, in particular.

#### 3.1 Infrared Thermography

IR thermography is a technology that maps the surface, that is, skin, temperature of a body [52]. Use of thermal images for remote measurement of fever have been investigated in medical and health sciences for few decades [53, 54, 55, 56, 57].

There is an evidential correlation between the facial temperature and the health status of a subject. Normal body temperature ranges from about  $36^{\circ}\text{C}$  to about  $37.2^{\circ}\text{C}$ .

Fever is a very well known symptom of some infectious diseases such as influenza, SARs etc. Fever screening based on infrared thermographs (IRT) is an approach that has been implemented during infectious disease pandemics, such as Ebola and Severe Acute Respiratory Syndrome [23].

Fever is also often part of the defensive response of the immune system of the human body, which aims to kill intruding microbiological organisms (e.g. bacteria, viruses etc.) [58]. Normal fever is not so bad as it supports the healing process of the body. However, if the the fever increases above  $43^{\circ}\text{C}$  and also very persistent, it could be dangerous [59]. Suppressing such fever is very important especially for older people in order to protect their cardiovascular organs [60].

The most well-known approach to measure fever is using a mercury-in-glass thermometer [60]. The cheap and affordable standard devices sometimes require up to 4-5minutes to obtain a proper

temperature reading. The reliability of these measurement devices is also dependent on the part of the body the temperature was obtained from. Common areas for temperature measurements are under the arms, mouth and ear region but the axilla region doesn't always give a very precise measurements [61]. There are other areas of the body that temperature reading can be taken from but some of them can be very intrusive, inconvenient or even painful like the rectal region [62]. Not being very patient with the measurement process, because of waiting times of 5 minutes or more, can cause false measurement results because the subject may not be able to remain calm for the duration of the measurement. It becomes extremely challenging to measure fever especially for children because they cannot stay calm for a prolonged length of time. A specialized fever measurement device which uses infrared on the ear drum [63] is very fast and reliable when used properly [64, 65]. However, the measurement is prone to error, for instance, when the tympanic thermometer is held at a wrong angle. Additionally, both devices need direct body contact, which is unhygienic. On the contrary, there are forehead thermometers that can be used without direct body contact, but studies has shown that they have some limitations [66]. Nevertheless, these thermometers offer a fast measurement and are hygienic.

It has become considerably easy to travel through different countries and continents with various climates. This has contributed to the increase in epidemic and pandemic crises, especially in elderly people or people with a weak immune system [67]. Therefore, in airports or other border control points, a quick temperature measurement and tests using thermal cameras have been found beneficial [68]. Due to the increased use of thermal images in detecting illness, there has been an investigation of how reliable the thermal cameras are for fever measurement [54, 55, 56, 57]. In [56], the authors discuss the need of standardization when using infrared cameras.

Since the main goal of this study is investigate the possible solutions for remote, non-invasive ways to assess the subjects' possible health-related thermal status abnormalities, we first embark on measurement of face skin temperature, as the most obvious area of body available at the point of access to a shelter. Specifically, in this study, we collaborated with the IT department of the Calgary

Drop-in Centre which is in the process of the implementation of a kiosk with a face recognition for frequent clients of this shelter. The IT department there was interested to case study the use of IRT, together with the facial recognition system. The goal is to identify clients who might be sick (fever, other illnesses manifested via elevated or abnormally low body temperature) and need to be directed to emergency healthcare services.

Motivated by these applications, this thesis investigates a contactless body temperature measurement approach based on IRT, using a thermal camera. The aim of this development is to automate fever measurement, including automated region-of-interest detection. Unlike the existing contact or semi-contact temperature measurement devices, such contactless measurement would be an ambient screening of a user of an access kiosk point, and the subject would not feel that their privacy is violated. This solution will also help the shelters or other mass access system staff to provide appropriate health services to their clients. This solution will also be embedded in the application for a client identification kiosk (using face biometrics) that is being developed, in particular, in the Calgary Drop-in Centre. Since such kiosk requires a client to have their faces to present to the video camera, an additional biometrics, thermal face video, would allow for the health status check in the same setup as the established authentication process. If a subject's thermal images display a high fever or other abnormal thermal patterns, a health personnel would be dispatched to access the client and provide medical help if needed.

### 3.2 Working Principle of an Infrared Camera

Thermal cameras are sensitive to thermal radiation at the infrared light wavelength. This light is invisible to the naked eye, but can be felt as heat if the intensity is high enough.

All objects emit some kind of infrared radiation, and it's one of the ways that heat is transferred. If you hold your hand over some hot coals on the grill, those coals are emitting a ton of infrared radiation, and the heat is transferring to your hand. The hotter an object is, the more infrared radiation it produces. Thermal cameras can see this radiation and convert it to an image that we



can then see with our eyes, much like how a night vision camera can capture invisible infrared light and convert it to an image that our eyes can see. The amount of radiation a body emits can be calculated using Planck's law:

$$B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1} \quad (3.1)$$

where  $\nu$  is the frequency of radiation,  $T$  is the absolute temperature,  $k_B$  is the Boltzmann constant,  $h$  is the Planck constant, and  $c$  is the speed of light in the medium, whether material or vacuum.

The thermal cameras can be cooled (older technology) or uncooled. Inside of an uncooled thermal camera, there is an array of measuring devices that capture infrared radiation, called microbolometer. Microbolometers do not detect photons. Instead, they pick up on temperature differences by sensing thermal radiation from a distant object. As microbolometers absorb thermal energy, their detector sensors rise in temperature, which in turn alters the electrical resistance of the sensor material. An embedded processor can interpret these changes in resistance and use the data points to generate an image on a display. These arrays do not need any cooling systems. That means they can be integrated into smaller devices, such as night vision goggles, weapons sights and handheld thermal imaging cameras.

### 3.3 Temperature Extraction and Classification

The value of the temperature derived from infrared images is generally evaluated based on the image pixel intensity. The raw infrared images are usually represented by a grey-scale image intensity varying from 0 to 255 (256 levels of intensity). The higher the intensity value, the higher is the temperature; however, this relationship is not linear. To estimate the temperature value, the principle of polynomial curve fitting is generally used.

Curve fitting is a process used in predictive analytics. Its goal is to create a curve that describes a mathematical function that best fits a series of data points. Interpolation is normally used in order to find the exact fit to the data [69]. An alternative way is to create a smoothing function which "fit" the data with a certain approximation. In either case, a single mathematical function is assigned

to the entire body of data, with the goal of fitting all data points into a curve that delineates trends and aids prediction. Fitted curves can be used as an aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among two or more variables. Extrapolation can be further used to extend the fitted curve beyond the range of the observed data. Extrapolation is subject to a degree of uncertainty since it may reflect the method used to construct the curve as much as it reflects the observed data [70].

### 3.4 Experiment

The purpose of the experiments is to test the feasibility of the proposed approach and the selected system architecture (including a thermal camera) for fever detection. The first section describes the task of fever detection, which involves classifying the subject database into three classes. These three classes are: low temperature, normal temperature and high temperature (fever). The precision of the proposed system is compared to the temperature readings from a reasonably precise forehead thermometer which serves as the ground truth. Next, the standard deviation and the variance of our reading is calculated. Finally, we calculate how close the remote reading to the temperature reading from a forehead thermometer.

The data for this experiment were obtained using the frames with images of faces of multiple subjects using the infrared camera Miricle from Thermoteknix Inc. The data were collected under the University of Calgary Research Ethics Committee approval (a Certificate of Approval can be found in the Appendix to this thesis).

The flowchart of the proposed approach to the facial temperature estimation using remote measurements is shown in Figure 3.1.

To test the feasibility of our system in varying environmental condition, we acquired the reading at two different locations and at different times of the day. Multiple thermal images (frames of the infrared video) of total of 55 subjects were collected for this experiment at room temperature. The resulting measurements were grouped into three classes: hypothermia (abnormally low

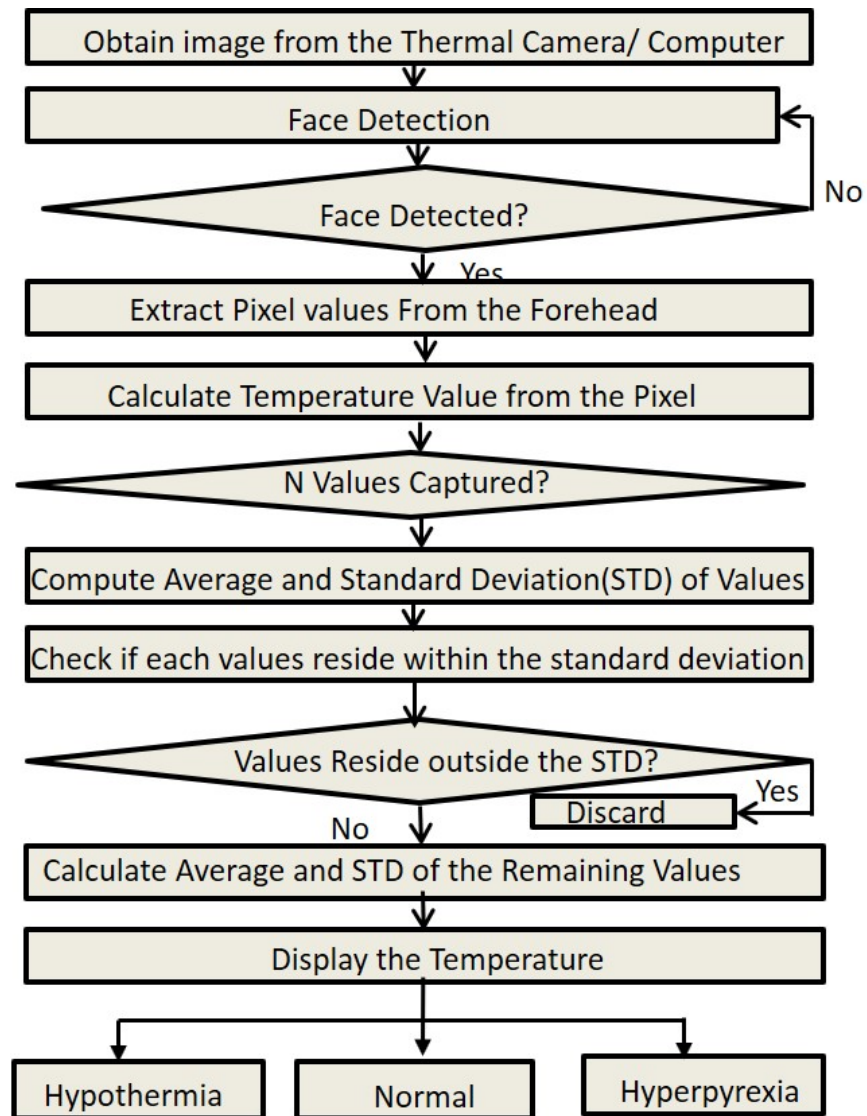


Figure 3.1: Flow chart of the proposed algorithm for fever detection

temperature), normal temperature and hyperpyrexia (high body temperature, or fever). At the time of the thermal video recording, the forehead temperature was measured using a thermometer, to serve as our ground truth for comparison. Table 3.1 shows the value of the average temperature for each subject during one recording session, using both the thermal camera and the thermometer. Figure 4.18 shows the interface of the Matlab-based application we built for the measurement of temperature using the thermal camera.

### 3.5 Analysis of Experimental Result

For analysis of the experimental results, we used the direct comparison approach using the coefficient of determination (also known as R-squared value), as well as a method known as Bland-Altman plot.

The coefficient of determination, or R-squared, is a statistical measure that represents the proportion between the variances of two variables in a regression model. In our case, we compare variances of both ways to measure facial temperature. For example, if the R-squared of a model is 0.50, then approximately half of the observed variation can be explained by the measured variable (temperature).

A graph illustrating coefficient of determination, or R-squared, for the reading from the thermometer and the thermal camera is shown in Figure 3.2. The R-squared value is more than 90% which means that the points are not so far from the fitted line (which indicated where the two values are the same).

Bland-Altman plot, or difference plot, is used in medical research to measure the agreement between two measures [71]. The resulting graph is a scatter plot that represents a difference between the two paired measurements versus the mean of the two measurements. The mean difference is the estimated bias, and the standard deviation  $s$  of the differences measures the random fluctuations around this mean. This plot allows to estimate whether 95% of the data points lies within  $\pm 2s$ , which corresponds to an acceptable error.

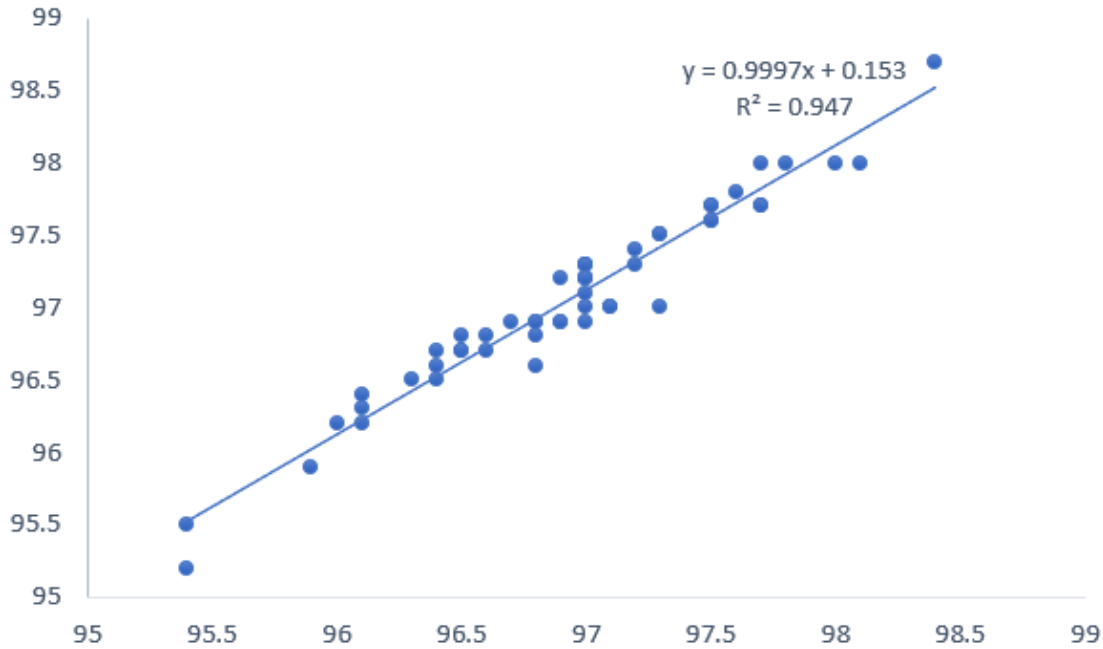


Figure 3.2: Infrared Camera reading versus forehead thermometer reading, in Farenheit

The Bland-Altman plot for our experiment is shown in Figure 3.3. In our case, 95% of our points lie within the  $\pm 2s$  of the mean difference, and this is an acceptable accuracy for estimation of the normal forehead temperature.

The regression analysis using those plots confirm that given 95% confidence, the difference between the contact and contactless measurement lies within mean plus/minus two standard deviations. The 95% limits of agreement can be unreliable estimate for small sample sizes, and, thus, it is important to calculate confidence intervals for 95% limits of confidence. The Bland and Altman's approximate method provides this estimate of interval quite well.

### 3.6 Conclusion

The study presented in this chapter discussed the feasibility of detecting illness and diseases with the infrared camera. An infrared camera Miricle by Thermoteknix Ltd. was used to imitate the environment in the access point at a shelter, Calgary Drop-in center as a case study. The goal was to investigate whether it is possible to conduct such measurements for the aforementioned

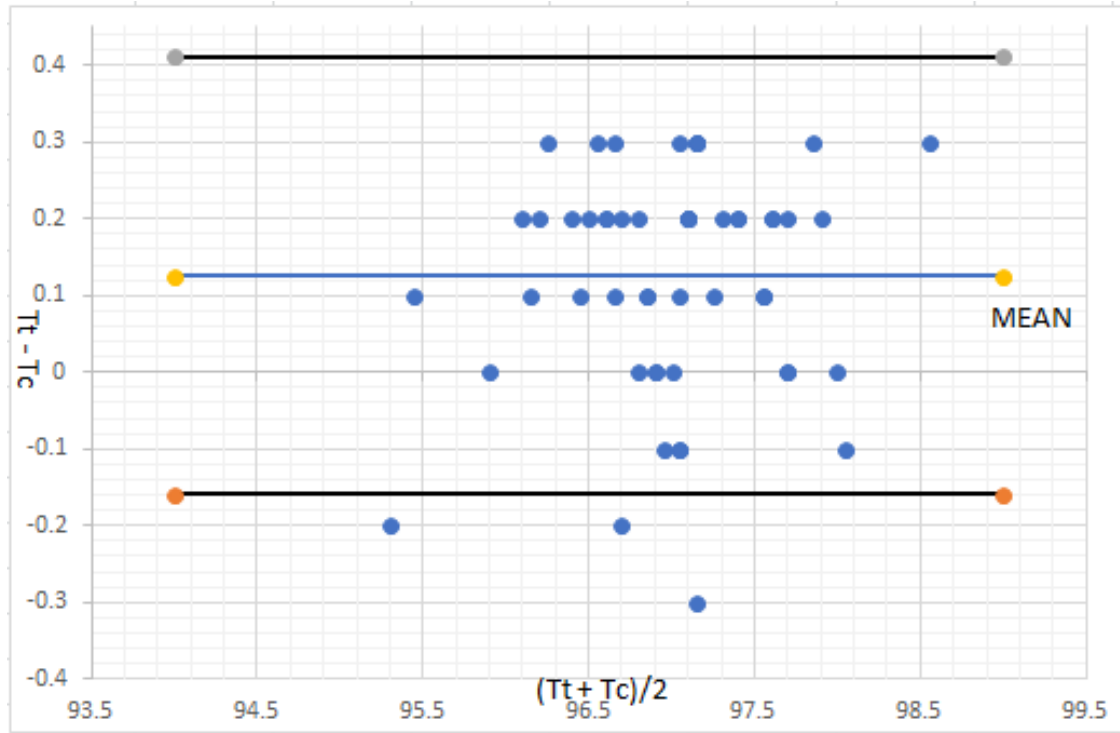


Figure 3.3: Bland-Altman Plot of the temperature measures by the thermometer and the thermal camera, in Farenheit

application, and whether the accuracy of measurement is acceptable.

Using the proposed regression model, we drew the conclusion that there is a strong correlation between the temperature reading from a forehead thermometer with the one using the thermal camera system. In particular, the bland-Altman plot shows that over 95% of our data points lies within two standard deviations, and, thus, it is quite feasible, given the controlled experimental conditions, such as a subject standing in front of camera, in a room-temperature environment. Unfortunately, we were not able to conduct experiment on the subject coming from extreme cold weather, which may influence the measurement accuracy. As well, it was difficult or impossible to have febrile subjects, or subjects with extreme fever. Further research is required to evaluate the thermography based measurement accuracy for extreme environment and subject conditions.

Table 3.1: Comparison of the Temperature measurement from the Camera and the Thermometer.

Number	Thermometer( $^{\circ}F$ )	Thermometer( $^{\circ}C$ )	IR Reading( $^{\circ}F$ )	IR Reading( $^{\circ}C$ )	Ta - Tm	Avg( $^{\circ}F$ )	STD
1	96.3	35.7	96.5	35.8	0.2	96.5	0.04
2	97.0	36.1	97.3	36.3	0.3	97.2	0.09
3	97.1	36.2	97.0	36.1	-0.1	97.1	0.01
4	97.0	36.1	97.3	36.3	0.3	97.2	0.09
5	97.3	36.3	97.0	36.1	-0.3	97.2	0.09
6	96.5	35.8	96.8	36.0	0.3	96.7	0.09
7	97.0	36.1	97.2	36.2	0.2	97.1	0.04
8	96.8	36.0	96.9	36.1	0.1	96.9	0.01
9	96.9	36.1	97.2	36.22	0.3	97.1	0.09
10	96.4	35.8	96.6	35.9	0.2	96.5	0.04
11	98.0	36.7	98.0	36.1	0.0	98.0	0.00
12	97.5	36.4	97.7	36.5	0.2	97.6	0.04
13	97.0	36.1	96.9	36.1	-0.1	96.9	0.01
14	97.5	36.4	97.7	36.5	0.2	97.6	0.04
15	97.2	36.2	97.4	36.3	0.2	97.3	0.04
16	97.0	36.1	97.3	36.3	0.3	97.2	0.09
17	97.0	36.1	97.2	36.2	0.2	97.1	0.04
18	97.1	36.2	97.0	36.1	-0.1	97.1	0.01
19	95.4	35.2	95.5	35.3	0.1	95.5	0.01
20	97.3	36.3	97.5	36.4	0.2	97.4	0.04
21	96.5	35.8	96.7	35.9	0.2	96.6	0.04
22	97.0	36.1	97.1	36.2	0.1	97.1	0.01
23	95.9	35.5	95.9	35.5	0.0	95.9	0.00
24	96.4	35.8	96.7	35.9	0.3	96.6	0.09
25	96.1	35.6	96.4	35.8	0.3	96.3	0.09
26	96.1	35.6	96.2	35.7	0.1	96.2	0.01
27	96.6	35.9	96.7	35.9	0.1	96.7	0.01
28	97.3	36.3	97.5	36.4	0.2	97.4	0.04
29	96.1	35.6	96.3	35.7	0.2	96.2	0.04
30	97.0	36.1	97.3	36.3	0.3	97.2	0.09
31	97.5	36.4	97.6	36.4	0.1	97.6	0.01
32	98.1	36.7	98	36.7	-0.1	98.1	0.01
33	96.8	36.0	96.8	36.0	0.0	96.8	0.00
34	96.4	35.8	96.5	35.8	0.1	96.5	0.01
35	95.4	35.2	95.2	35.1	-0.2	95.3	0.04
36	97.8	36.1	98.0	36.7	0.2	97.9	0.04
37	97.2	36.2	97.3	36.3	0.1	97.3	0.01
38	97.0	36.1	97.0	36.1	0.0	97.0	0.00
39	96.6	35.9	96.8	36.0	0.2	96.7	0.04
40	97.0	36.1	97.2	36.2	0.2	97.1	0.04
41	96.8	36.0	96.9	36.1	0.1	96.9	0.01
42	97.0	36.1	97.2	36.2	0.2	97.1	0.04
43	96.8	35.8	96.6	36.1	-0.2	96.7	0.04
44	97.7	36.5	98.0	36.7	0.3	97.9	0.09
45	97.7	36.5	97.7	36.5	0.0	97.7	0
46	96.0	35.6	96.2	35.7	0.2	96.1	0.04
47	96.9	36.1	96.9	36.1	0.0	96.9	0
48	96.7	35.9	96.9	36.1	0.2	96.8	0.04
49	97.6	36.4	97.8	36.6	0.2	97.7	0.04
50	97.7	36.5	97.7	36.5	0.0	97.7	0
51	97.5	36.4	97.6	36.4	0.1	97.6	0.01
52	98.4	36.9	98.7	37.1	0.3	98.6	0.09
53	96.5	35.8	96.7	35.9	0.2	96.6	0.04
54	97.0	36.1	97.2	36.2	0.2	97.1	0.04
55	96.9	36.1	96.9	36.1	0.0	96.9	0.00

## **Chapter 4**

# **Insobriety Detection using Thermal images and Machine Learning**

This chapter is dedicated to investigation of possible solution to another problem, formulated by the collaborating IT department in the Calgary Drop-in & Rehab Centre Society. The problem in question is to identify clients who may not be sober, i.e. under alcohol influence, and, thus, may need to be directed to the shelter area that is separated from the services to other clients seeking shelter. The identification of such clients might be conducted in a non-invasive way using thermal camera readings, and combined with the existing infrastructure such as a kiosk that already performs a function of face recognition.

This chapter introduces a novel approach to classification of patterns of thermal images of faces with the goal to detect subjects' insobriety. This approach is based on applying the state-of-the-art machine learning technique. Specifically, we apply deep neural networks, to extract features from thermal images of faces, and, consequently, classify the subjects to two groups: "sober" or "drunk". The method proposed in this chapter yielded a classification accuracy that is comparable or superior in some cases to the previous work done by other researchers in this field using different machine learning techniques including non-deep neural networks and Support Vector Machine (SVM) [16, 14].

### **4.1 Data sets for the experiment**

The database used for this study is the UPatras "Sober-Drunk" database [72]. The thermal video frames (images) from this database was used for training the network. The database was collected at the University of Patras in 2012-2013 using FLIR Thermo Vision Micron/A10 infrared camera



(18 mm, f/1.6) contains images of 41 subjects, 31 males and 10 females. Each person consumed four glasses of red wine, 120 ml each (13% vol.), in 1h period (a total of 480 ml of wine, i.e. 62.4 ml of alcohol). The first 50 frames were collected of each subject before alcohol consumption, and the second acquisition of 50 frames was performed 30 minutes after drinking the fourth glass of wine. In each acquisition, a sequence of 50 frames for each person was acquired with a sampling period of 100 ms between the frames. The resolution of the infrared images is  $128 \times 160$  pixels.

This database was used for training the pattern recognition algorithms we used. In addition, a combination of the UPatras database images and real-life subjects (sober) we collected for this study, was used for testing.

## 4.2 The proposed system

The following four sections describe the main components of the proposed approach and system architecture: 1) Input data preparation, 2) Face detection and region of interest extraction, 3) pre-processing, and 4) classification (we applied Convolutional Neural Network (CNN) and SVM and compared the results).

Figure 4.1 shows the overall system architecture of the insobriety detection system.

The overall architecture is described as follows:

- **Input Image:** The input images were prepared using the data from the UPatras database [72] and Matlab program for reading separate files, as provided by the database owners. We also used frames from short infrared videos of real-life subjects we collected in the Biometric Technologies Lab at UofC and in the Calgary Drop-in Centre's IT Department, using Miricle Thermoteknix infrared camera with resolution  $480 \times 640$ .
- **Face Detection and Region of Interest Extraction:** The subject's face is detected in the frame (image), then the ROI such as eyes and the forehead are is cropped, or extracted, and passed on to the next stage for pre-processing.

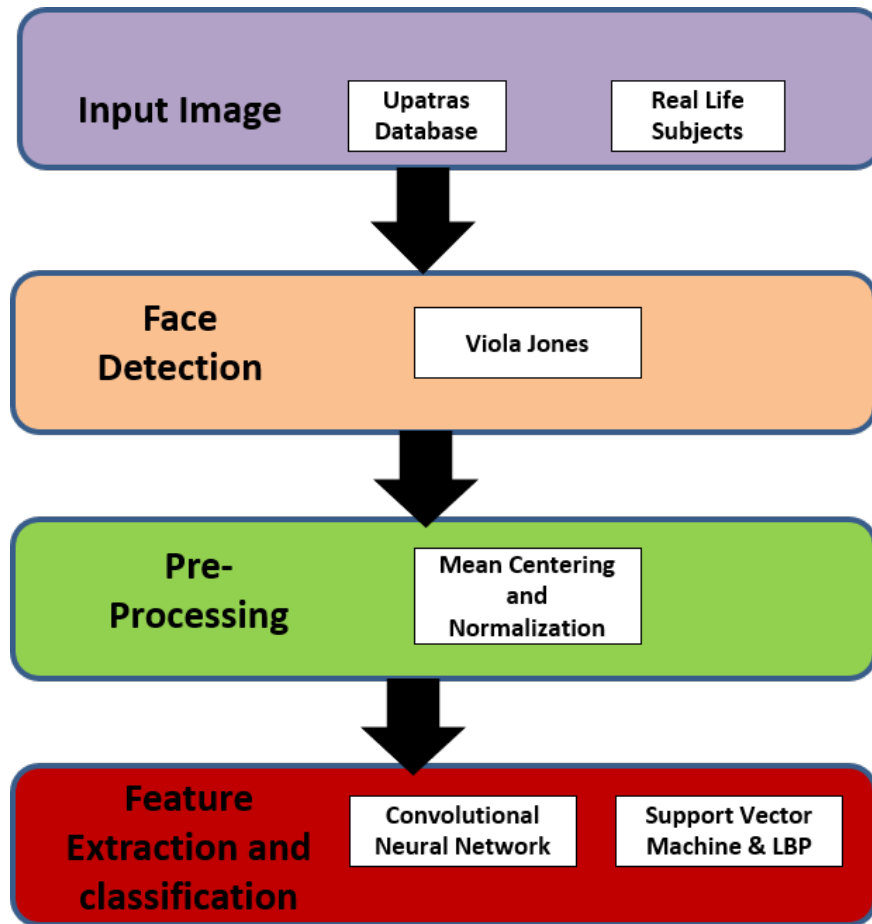


Figure 4.1: Overall system architecture of the proposed insobriety detection system. Input images are fed into the system for face detection, followed by pre-processing, feature extraction and classification. This architecture is used for both training on database images, and testing on probe images.

- **Pre-Processing:** At this stage, the mean centering is performed on the images.

This involves calculating the mean of the frame pixels' intensity, and finding the new intensity value for each pixel by subtracting the original pixel intensity from the mean.

The thermal image pixel values are integers with values between 0 and 255. A general data preparation technique for image data is to subtract the mean value from the pixel values. This approach is called centering, as the distribution of the pixel values is centered on the value of zero. Centering requires that a mean pixel value be calculated prior to subtracting it from the pixel values.

Next, a normalization is performed which involves scaling the data to be uniform. The images are also re-scaled to the same size so it can be acceptable by the CNN. In our experiment, we rescaled the forehead images to  $45 \times 15$ . The images of the eyes were rescaled to  $25 \times 15$ .

- **Feature Extraction and Classification:** This step is performed in our study by using two machine learning techniques: SVM and CNN. We compared the performance of both techniques.

The first approach includes feature extraction using an approach called Local Difference Patterns (LDP). The classification is performed using SVM. In this experiment, we use only the forehead as the ROI.

The second approach involves a CNN to perform both the feature extraction and classification. Two ROIs were used for this experiment, - the eyes and the forehead. A separate network was built for each ROI, since the respective images have different sizes and different feature vectors.

- **Fusion:** In order to improve the overall performance, we applied a fusion of two CNN based classifiers, one using the whole face frames and another using the eye

region. We considered fusion at both the feature level (via concatenation of feature vectors) and score level (using various strategies to combine the classification scores).

In Section 4.6.6, we provide the results of performance comparison of the proposed approach and the works done by other researchers using the same database. The accuracy of the system, especially after the fusion was performed, was comparable to other works. Section 4.6.7 presents the results obtained on the data acquired in the Biometric Technologies Laboratory and in the IT Department in the Calgary Drop-in Center.

### 4.3 Face Detection and Region-of-Interest Extraction

The ROIs on the facial images, or frames, of the subjects used in this experiment were detected using a well-known Viola-Jones face detector [18]. This face detector is commonly used in facial recognition to locate and crop faces in video frames, or images. One of its main features is that the training is quite slow while the detection is fast. The MATLAB implementation of Viola-Jones algorithm was used for this experiment. There are three basic steps involved in this algorithm: 1) Haar-like transform for feature extraction, 2) calculation of integral images, and 3) the AdaBoost feature selection.

1. The feature extraction is done using Haar-like features which is a very simple and efficient 2-dimensional (2D) filter. The 2D filters are “moved” across the image detecting facial features like the eyes, nose, mouth and the edges of the face. There are some properties common to the human face that makes it easy to detect. For example , the eye region of the face is darker than the upper cheeks, also the nose bridge region is brighter than the eyes region. Haar-like features are then calculated as shown in Figure 4.2: the sum of the pixels in the black rectangle is calculated, and then the sum of the pixels in the white rectangle is subtracted from the first

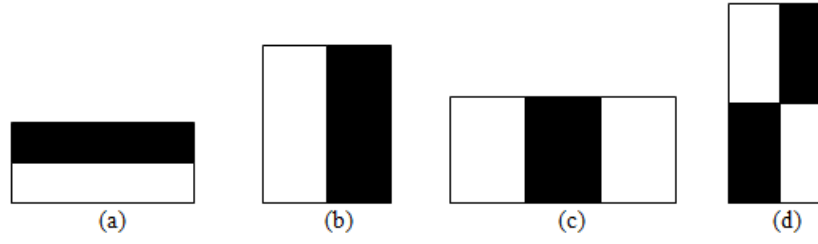


Figure 4.2: Rectangular templates used in Viola-Jones algorithm:(a) and (b) are two-rectangle feature, (c) is a three rectangular feature, and (d) is a four-rectangle feature. The illustration is inspired by [?]

sum. This calculation happens inside a detection window which is moved across the image.

2. This step is based on the algorithm that calculates an integral image in order to speed-up the calculations. The integral image allows for the Haar extractor to be calculated by substituting every pixel point with the sum of all the points to the left and above it. Figure 4.3 shows an integral image obtained for an image pixel. The Integral value at any point  $(x,y)$  in the summed-area table is the sum of all the pixels above and to the left of  $(x,y)$ , inclusive, as stated in the formula below:

$$I(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y') \quad (4.1)$$

In other words, the intensity  $I$  values of the integral pixels are computed using the formula:

$$I(x,y) = I(x,y) + I(x,y-1) + I(x-1,y) - I(x-1,y-1) \quad (4.2)$$

3. The last step is the AdaBoost feature selection which chooses the optimal features by combining a set of simple learning algorithms (also called weak learners) to form a strong classifier. No single feature can accurately classify a face with some minimal error. Each weak learning algorithm determines the optimal threshold for each classification which might reduce the classification error. A certain weight is

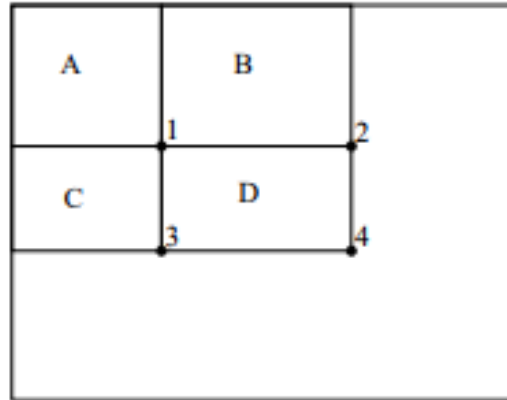


Figure 4.3: Integral Image using the sum of the pixels. The value of the integral pixel in location 1 is the sum of the pixel values in A, the value in location 2 is the sum of the pixels in A and B. Similarly, the values for locations 3 and 4 are the sum of A and C, and the sum of A,B, C and D, respectively

assigned to each weak learner according to its classification accuracy. Learners with high accuracy gets a higher weight whereas learners with low accuracy gets a lower weight. A cascade of classifiers is used for the classification. These classifiers achieve very high classification at a short computation time. This is achieved by ensuring that the boosted classifier's threshold is adjusted so that the false negative rate is reduced. This cascade of classifiers selects a rectangular region that is most likely the one containing a face. The first classifier, which is a simple classifier, eliminates all the regions that are clearly incorrect, and the other set of classifiers, which is more complex, eliminates the other sub regions. The the final result will give a rectangular region bounding the face found in the image.

It is important to mention that Viola-Jones algorithm can return multiple bounding boxes of the face in an image. To correct this error, the minimum size of the face can be selected to reduce the number of incorrect classifications. The minimum image size is typically the smallest possible image that can be used in the experiment. Viola-Jones algorithm is very fast and accurate in face detection.

In our approach, after the face is extracted, the next step is to crop out the ROI on the face.

When using SVM for classification, only the forehead is used as a ROI. When using the CNN, a combination of both face and eyes are used as the ROI. The forehead was chosen as the ROI, because the research conducted in [14] shows that that binary pattern of the pixels on the forehead changes after the consumption of alcohol. As well, after alcohol consumption, the temperature distribution of the sclera and the iris changes: the sclera becomes evidently hotter than the iris [16]. To extract the eye regions, we had to manually draw a triangular box around the region, then map the locations over other frames of the same image. This region is cropped from the facial image, and then normalized to a uniform size to work well with the network.

#### 4.4 Pre-Processing

The major pre-processing techniques performed on the images are mean centering and size normalization. Mean centering involves calculating the mean of each data point and subtracting each pixel from the mean. To achieve this, the mean centered images is divided by the standard deviation of the sample. The formula used to perform this is given below:

$$m_i = \frac{x_i - \mu_{x_i}}{\sigma_{x_i}} \quad (4.3)$$

where  $x_i$  is one data sample,  $m_i$  is the normalized data sample,  $\mu_{x_i}$  is the mean of the sample, and  $\sigma_{x_i}$  is the standard deviation of the sample. After the mean centering, we perform a normalization which involves scaling the data to be uniform. The images were clipped and re-sized in such a way that it will be acceptable to the network. The technique that was used for this step is the bicubic interpolation. For the forehead images, it was re-sized to  $45 \times 15$  which is very close to the size of the cropped image. The initial formula for bicubic interpolation is as follows:

$$p(x,y) = \sum_{i=0}^3 \sum_{j=0}^3 \alpha_{ij} x^i y^j \quad (4.4)$$

where  $x$  and  $y$  are the coordinates of the pixel being interpolated,  $p(x,y)$  is the output, and  $\alpha_{ij}$  is a 16-term coefficient that is a function of the derivatives in the  $x$  and  $y$  directions. The coefficients are found by taking the partial derivatives of the above formula and solving a set of equations.

$ y_1 - x $	$ y_2 - x $	$ y_3 - x $
$ y_8 - x $	$x$	$ y_4 - x $
$ y_7 - x $	$ y_6 - x $	$ y_5 - x $

Figure 4.4: Creation of a simple LBP

Bi-cubic interpolation was chosen as it is generally smoother than other known interpolations such as bi-linear or nearest-neighbor.

## 4.5 Implementation using Support Vector Machines

To form the input for the SVM classifier, we have to extract the thermal image features suggesting alcohol intoxication. The ROI used in this case is the forehead. The feature that were extracted are based on the LDP on the forehead; this feature was shown to work best for thermal images for similar task [14].

### 4.5.1 Feature Extraction using Local Difference Patterns

Local Difference Pattern (LDP) is a modified version of the Local Binary Pattern (LBP).

LBP is a feature extractor that transforms every pixel of an image into a binary pattern by determining their relationship with the neighboring pixel [73]. The LBP was proposed in [74]. It suggested that for each pixel in the image, considered as a centre pixel in its location, a local information can be extracted. The difference between each center pixel intensity and the neighbor pixel intensity is evaluated, and a binary value is assigned to the neighbor pixel depending on the value of the difference. Using a  $3 \times 3$  window size, for a center pixel with intensity  $I_c$ , and the neighboring pixels with intensities  $I_n (n = 1, 2, \dots, 8)$ , an LBP is created as shown in Figure 4.4.



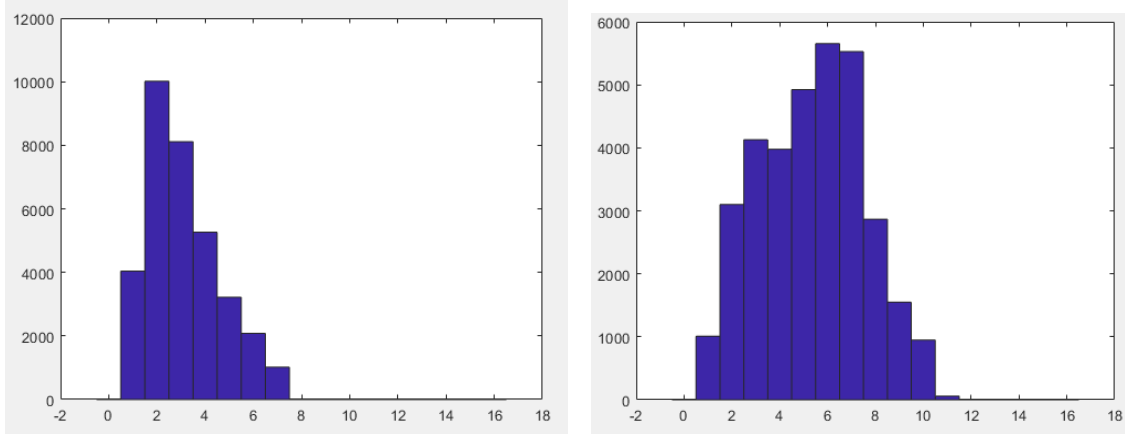


Figure 4.5: Histograms of the LDP for Drunk and Sober subjects, respectively

After the LBP is computed, the binary value of 1 or 0 is assigned to each output pixel.

The method used in this thesis for feature extraction is LDP. This modified feature extraction approach is proposed by [14]. It uses a window size with the sum of the differences of the neighboring pixels to get a difference pattern, LDP. The formula is:

$$z = \sum_{i=1}^8 |y_i - x| \quad (4.5)$$

The LDP is evaluated over the region of interest (forehead) of all the images corresponding to both the “sober” and “drunk” persons. The area of the forehead region is  $15 \times 45$  pixels. All the values obtained are used to form a histogram representing the distribution of pixel intensity differences. Figure 4.5 shows the histogram of difference values of a subject’s “drunk” and “sober” images. The values obtained from this histogram is used for classification using SVM as described in the next section.

#### 4.5.2 Support Vector Machine for classification

SVM is a supervised learning models which is associated with algorithms used for data classification and regression analysis. It requires the data to be labeled, and it can only classify between two classes. To use SVM to classify between more than 2 classes requires combining multiple SVMs together. SVM calculates a decision boundary between the provided classes based on the given data samples and hyper-parameters. SVM can perform non-linear classification using the kernel

function. It implicitly maps the input into a high dimensional feature space. These are functions which takes low dimensional input space and transform it to a higher dimensional space, thereby converting a non-separable problem to a separable one. In many applications, non-linear classifier provides better accuracy, although the linear classifiers have simple training algorithms that are very scalable [75].

Matlab implementation of SVM and the C-SVM library was used for this work [76]. C-SVM has to do with solving a decision boundary which has the maximum space between the two classes [75][77]. The linear case of C-SVM is defined as follows:

$$f(x) = w^T x + b = 0 \quad (4.6)$$

where  $w^T$  is a vector of the weights normal to the decision boundary,  $b$  is the bias vector and  $x$  is the input data size of  $n$ , with corresponding labels  $y$ . The boundary margin is defined as the distance from the decision boundary to the closest points in either class. The closest points are known as support vectors. Figure 4.6 shows the idea of the decision boundary, the margin, and support vectors, which are the circled points closest to the decision boundary.

The distance we want to maximize is the margin which has length  $\frac{2}{\|w\|}$ . Each sides margin has an equal length of  $\frac{1}{\|w\|}$ , which is equivalent to minimizing  $w^T w$ . To add the constraint that the data must fall outside or on the margin:

$$y(w^T x + b) \geq 1 \quad (4.7)$$

When the decision boundaries and the data are not linear, the data is mapped implicitly to a higher dimension using the kernel function.

For decision boundaries and data which are non-linear, the kernel function is used to implicitly map the data to a higher dimension. This is important for data which is not linearly separable. If we define a non-linear function  $\phi$ , which maps  $x$  from its dimension to a higher dimension, we can then define the kernel function  $K$  to be the dot product of two mapped data points. This is done in order to remove the dot product  $\phi(x_i)^T \phi(x_i)$ , which is very computationally expensive, and replace

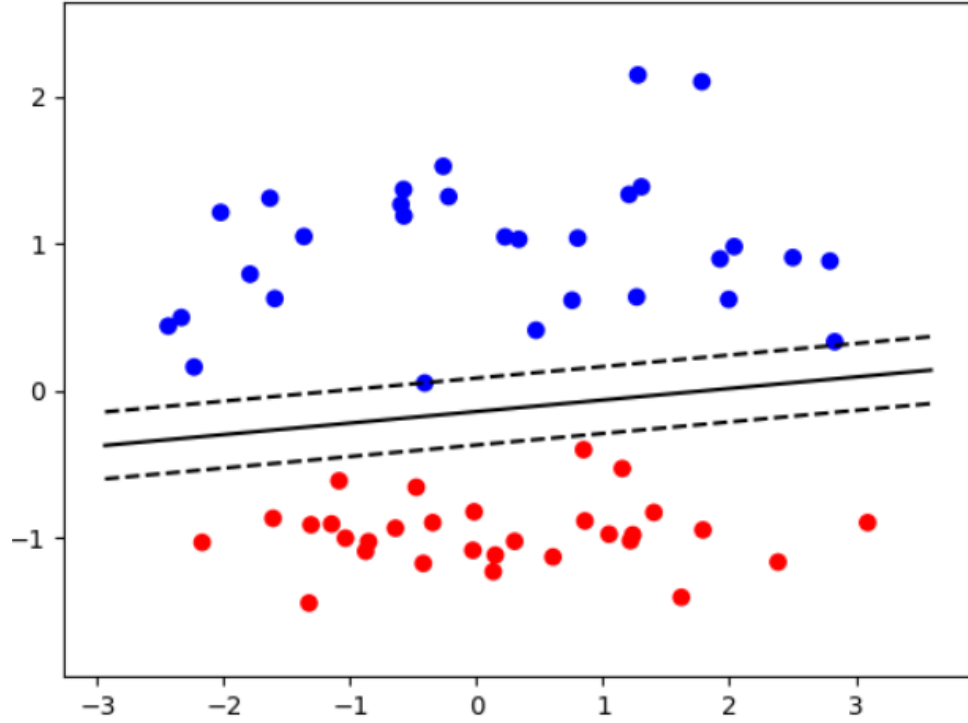


Figure 4.6: Support Vector Machine illustration from [78]

it with a function  $K$  which we can quickly calculate.

The weight vector is defined as a linear combination of the mapped training data  $w = \sum_{i=1}^n \alpha_i y_i \phi(x_i)$ .

The common types of kernel used are the linear, radial basis function (RBF) and the polynomial kernel. Amongst the three types, linear is the simplest. It trains faster and it has only one parameter to optimize. RBF has the two parameters to optimize but it trains slower than the linear. If the optimal parameters are found, then it will train as fast as the linear kernel. In the case of the polynomial kernel, it can perform better than the RBF for certain data types.

RBF kernel has the form

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (4.8)$$

where  $\gamma$  is a hyper-parameter that is supposed to be optimized. As  $\gamma$  increases, the flexibility or curvature of the RBF decision boundary increases, which can lead to over-fitting. When  $\gamma$  is small, the decision boundary is nearly linear. The goal is to separate the data with a plane with the minimum misclassified images, such that the distance from the plane to all points is maximum.

SVM also takes into account outlier points  $z$  which are allowed to be misclassified with a penalty cost  $C$ . If  $C$  is large, there is a larger penalty to misclassified samples. This constraint generally prevents poor generalization due to outlier data points. Thus,  $C$  is the second hyper-parameter that need optimization.

#### 4.5.3 Experiments

The experiment was performed to test the feasibility of the proposed system architecture using SVM for detecting insobriety using thermal frames.

The output is expected to be the input image label corresponding to one of two classes: 'drunk' and 'sober'. The accuracy of this system is compared with the results reported in other works in the same area.

For input data for training the SVM classifier, we used size-normalized images from the UPa-tras database. For testing, we used a combination of database images, and images of subjects we collected in the Biometric Technologies Laboratory.

The Sober-Drunk database collected by a team at the University of Patras, Greece was used for this experiment. 16000 usable frames of thermal recording of 40 subjects provided a total of 4000 images of sober subjects and 12000 by the collectors for each subject: 4 images of the face, 4 for the eye region, 4 for the ears and 4 for the hand. We used the frames capturing the face and the eye region. The first 4 acquisitions of these regions were done before alcohol was taken, and the remaining 12 were done after some alcohol consumption. For each acquisition, 50 sequential frames were taken every 100 ms, and the total time of acquisition was 5 s for the 50 frames. The resolution of the camera that was used in acquisition is  $128 \times 160$  with 16 bits per pixel. The images from the database serves as the input.

All the 16,000 usable frames of all the 40 subjects were used. A leave-one-subject-out cross validation (40-fold cross validation) was used to find the systems accuracy.

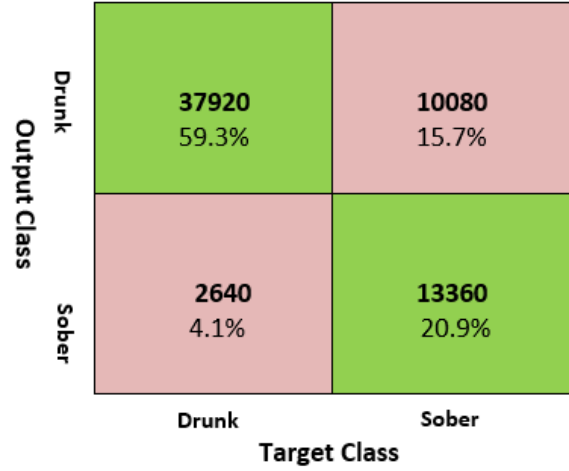


Figure 4.7: Confusion Matrix for the classification of subjects

The accuracy was calculated using the formula:

$$OverallAccuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (4.9)$$

Table 4.1 shows the accuracy of each class compared against previous works.

A confusion matrix for the classification is shown in Figure 4.7. The matrix represents the accuracy of classification for the 40 iterations during the cross-validation. Each subjects has 1600 images that was used for testing in each iteration, which gives a total of 64,000 training/testing images. In the pursuit of improving the accuracy of classification, the experiment on machine learning was conducted in section 4.6.

Table 4.1: Comparison of the accuracy of the proposed approach with other methods

Method	Sober Accur.	Drunk Accur.	Overall Accur.
Local Diff. Pattern [14]	N/A	N/A	85%
SURF Features [79]	N/A	N/A	89.23%
Three-layer NN [80]	93.1%	84.2%	N/A
NN-Fusion Diss.Feat. [81]	N/A	N/A	<b>96.2%</b>
Proposed system	83.5%*	79%	80.13%

The accuracies of the SVM-based are less, compared to the results reported in this area cited in the Table 4.1 . This gives are motivation to apply other techniques such as deep neural networks,

as described in the next section.

## 4.6 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of deep learning which is commonly used in visual images. In this work it is used as both feature extractor and classifier. CNN is similar to a multilayer perceptron which is design to reduce the pre-processing. A CNN has an input layer, normalization layer, pooling layer, fully connected layer and the output layer. The training data for a neural network is appropriately labeled and fed into the network. This input is passed through activation functions and series of weight, and an output score is produced. This output score creates an error margin which is back propagated through the network to readjust the weight. This readjustment of the weight continues until a very low error is achieved. For a regular neural network, if the input size is very large, the size of the network grows exponentially and the computation time is not feasible.

CNN uses 2D filters to extract information from an input image. These filters reduces the size of the input image while preserving the most important information in the image. By doing so, the computational time required to train the network is greatly reduced. The 2D filters slide through the input image and evaluate each section of the image one at a time. The input images is reduced as it moves through the layers of the network. Figure 4.8 shows the overall architecture of a CNN. If the number of the training data increases, it is accommodated by increasing the number of layers in the network. The overall architecture of CNN includes an input layer, activation functions, convolution and pooling layers.

Training a CNN requires the use of stochastic gradient descent (SGD) and back-propagation (BP). A gradient measures the change in the output of a function as the input value changes. SGD aims at minimization of the cost function and to achieve a local minimum. The algorithm sweeps through the training set, and updates the cost function until it converges.

Convolution and fully connected layers are analogous to traditional hidden layers, they perform

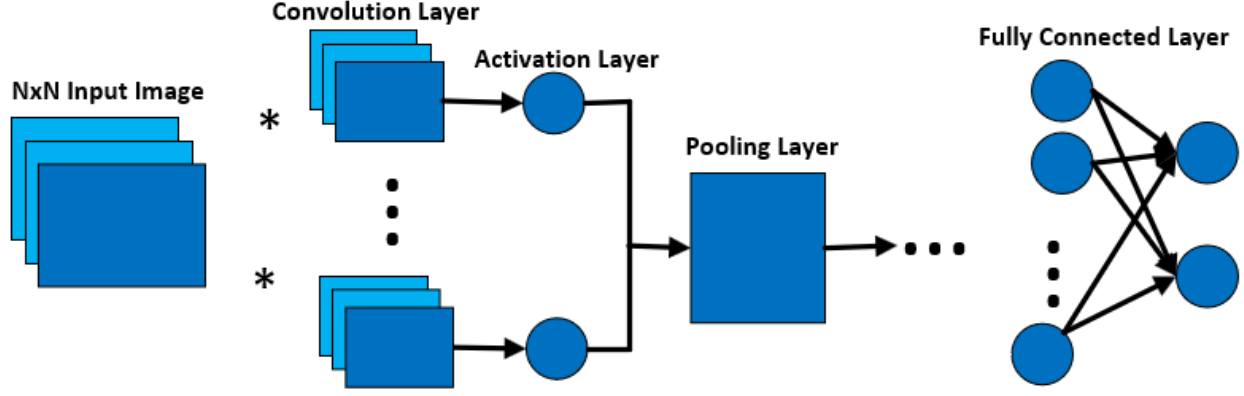


Figure 4.8: Architecture of a Convolutional Neural Network

the feature extraction and are responsible for the majority of the learning. Activation functions are the same as in traditional neural networks. Pooling layers are used to reduce the size of the data, while ideally retaining most of the information learned. Dropout and batch normalization are used to increase the performance of the network by reducing over-fitting and improving convergence times. An overview of the types of layers most commonly used is given in the next subsections.

#### 4.6.1 Layers of the CNN

##### Convolution Layer

The convolution (Conv) layer replaces to the individual weights that is used in the regular neural networks. A Conv layer has a  $P \times P \times D$  set of filters, and take as input a matrix of size  $N \times N \times D$ .  $P$  is the width and height of the filter kernel, usually between 3 and 11,  $D$  is the number of feature channels (for example, a greyscale image has 1 channel, color images have 3). For example, if the input size is an array of  $32 \times 32 \times 1$  pixels and the filter is  $5 \times 5 \times 1$  filter size. The filter will slide through and convolve with the image resulting in a  $28 \times 28 \times 1$  feature map. The response of the  $i^{th}$  Conv layer is expressed as follows:

$$Y_i = W_i X_i + b_i \quad (4.10)$$

where  $Y_i$  is the output of the Conv layer,  $X_i$  is the  $N \times N \times C$  input,  $W_i$  is the  $F \times F \times C \times D$  matrix of layer weights, and  $b$  is the vector biases. There is also the stride and the padding of the filter

that moves across the input. The stride, also known as the down-sampling factor, sets how much pixels the filter will slide through to perform another convolution. It is usually set to 1 or 2. When the stride is 1 then we move the filters to 1 pixel at a time. When the stride is 2 then we move the filters to 2 pixels at a time and so on. Setting this to 1 or 2, will make the output map not to be too small and it will retain the most important features of our image. Padding on the other hand defines how much zero padding will be added around the borders of the image. This usually done because the filters do not fit perfectly into the input image. It is usually set to 0 or 1. By adding the padding, it makes it possible to perform convolution at the borders of the image. This makes the borders of the image as important as the center part of the image. In general, given an input of size  $N \times N \times C$  and a Conv layer size of  $F \times F \times C \times D$ , the output will be of size  $N' \times N' \times D$ . The output size of the first two dimensions,  $N'$ , of a Conv layer can be calculated using the formula:

$$N' = \frac{N - F + 2P}{S} + 1 \quad (4.11)$$

where  $N'$  is the size of the first two dimensions of the output,  $N$  is the size of the first two dimensions of the input layer,  $F$  is the size of the filter kernel,  $P$  is the padding and  $S$  is the stride. For instance if  $F = 3$ ,  $S = 1$  and  $P = 1$ , the output  $N'$  becomes equal to the input  $N$ . If this happens, there will be no down-sampling at the convolution layer, only feature extraction. All the down-sampling will be left to the pooling layer.

Figure 4.9 and 4.10 shows two examples of a Conv layers with two variance of convolution. First, a Conv layer of size  $3 \times 3 \times 1 \times 2$ , with  $S = 1$  and  $P = 1$ , is applied to input with of size  $3 \times 3 \times 1$ , producing an output of size  $3 \times 3 \times 2$ . Each of the filters convolves the one input separately, and produces one output matrix each. Next, a  $3 \times 3 \times 2 \times 1$  Conv layer is applied to an input with two channels, of size  $3 \times 3 \times 2$ . Each channel of the input will be convolved by the corresponding channel of the Conv layer, and will be summed to produce one output. In the general case, both scenarios would be present in one Conv layer.



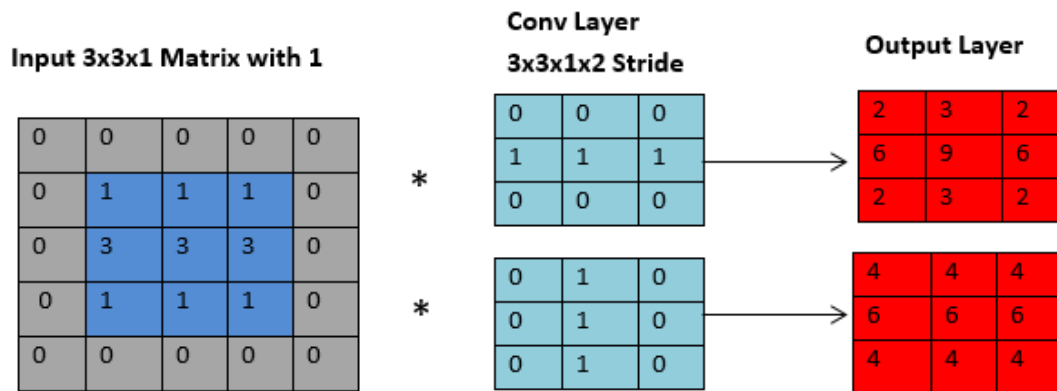


Figure 4.9: Conv Layer: A  $3 \times 3 \times 1 \times 2$  Conv layer with  $S = 1$ ,  $P = 1$  is applied to a  $3 \times 3 \times 1$  input matrix, each filter convolves the input separately, resulting in a  $3 \times 3 \times 2$  output matrix

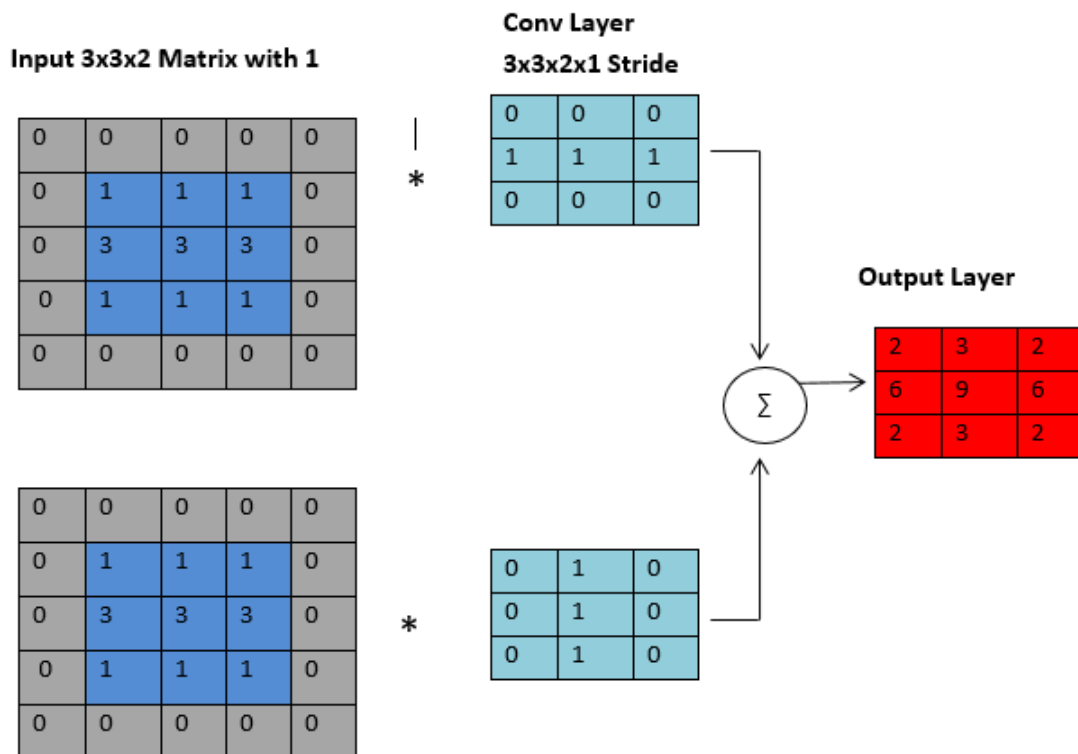


Figure 4.10: Conv Layer: A  $3 \times 3 \times 2 \times 1$  Conv layer with  $S = 1$ ,  $P = 1$  is applied to a  $3 \times 3 \times 2$  input matrix. Each convolves its corresponding channel, and are summed to form a  $3 \times 3 \times 1$  output

## Pooling Layer

The main function of this layer is to reduce the input image size, and at the same time preserve its most important features. This help reduce the computation time and the problem of over-fitting. The most common types of pooling are Max pooling, Average pooling and the Sum pooling. Max pooling is the much more common than the other two types. It extracts only the maximum pixel in the feature map, while average pooling calculates the average of the pixel, and sum pooling calculates the sum of all the pixels in the feature map.

The max pooling layer is needed because if a Conv layer extracts a feature, there is no need to preserve the information around the feature but the feature itself. Pooling layers have the same parameters as the Conv layer: kernel size, or the number of points to be pooled together, stride, and padding. A typical size of the pooling kernel is 2 or 3, with strides of 1 or 2 and padding of 0. Figure 4.11 shows an example of max pooling, average pooling and sum pooling layer.

There are two strides that is commonly used in pooling layer - overlapping pooling or non-overlapping. Overlapping pooling occurs when the stride  $S$  is smaller than the kernel  $F$ , most commonly,  $F = 3$  and  $S = 2$ . This allows input values to be included in multiple pooling operations. Non-overlapping occurs when  $F = S$ , most commonly 2. The pooling layer can be removed completely and replaced with the Conv layers with a higher stride to do the size reduction too [82]. This concept was not tested in this thesis.

## Batch Normalization

This is an optimization technique which normalizes the output of the Conv layer before they are fed into the activation function. This parametrizes the mean and the standard deviation as trainable parameters, to achieve convergence. The BNorm operation is performed on every mini batch. For every image in the mini-batch, it subtracts the per channel mini batch mean, and divides it by the per channel mini batch standard deviation:

$$Y(C) = \gamma \tilde{x}_i + \beta \quad (4.12)$$

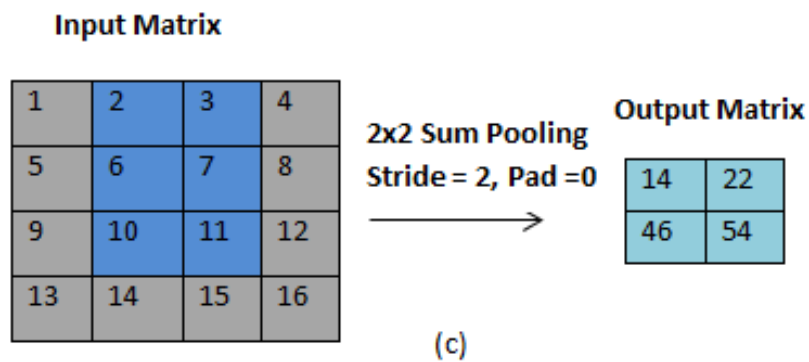
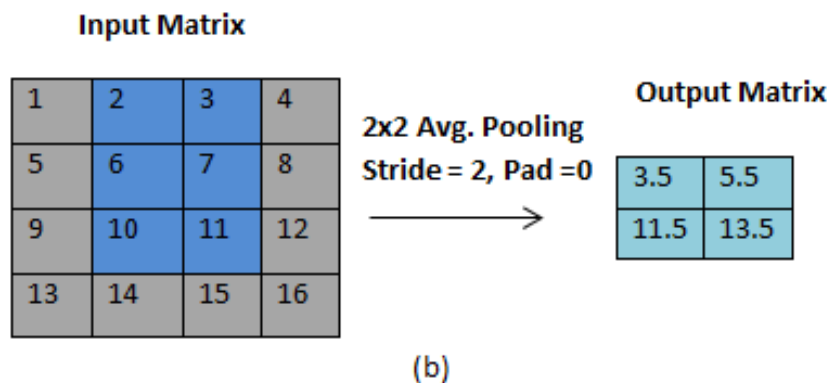
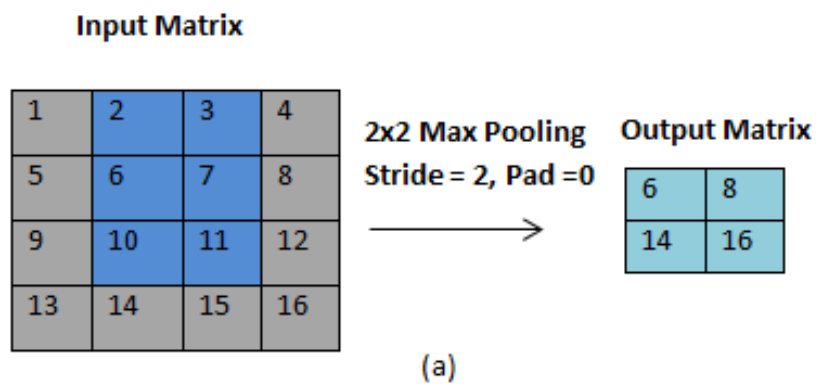


Figure 4.11: Pooling Layer: A  $2 \times 2$  pooling layer with a stride of 2 and pad of 0 is applied to a  $4 \times 4$  input matrix, resulting in a  $2 \times 2$  output matrix

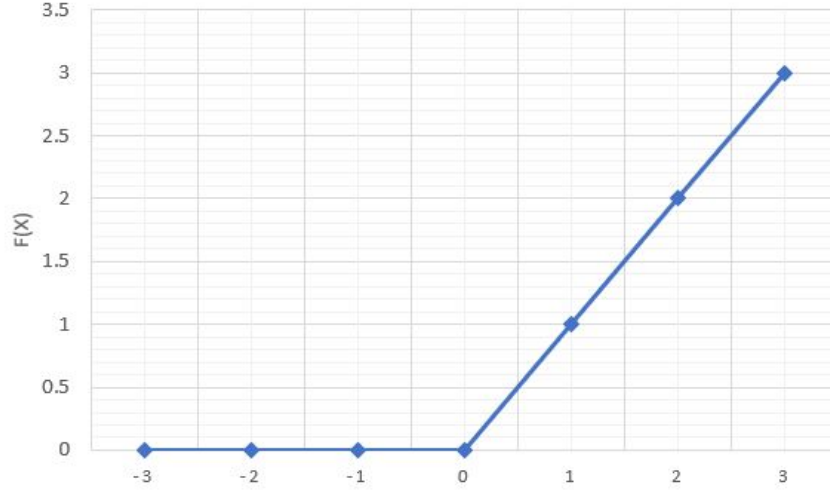


Figure 4.12: Rectified Linear Unit that zeros any negative inputs

$$\tilde{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (4.13)$$

$$\mu = \frac{1}{m} \sum_{i=1}^m x_i \quad (4.14)$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu)^2 \quad (4.15)$$

where  $Y(C)$  is the per channel output of BNorm,  $x_i$  is the  $i_{th}$  images single channel output from the previous Conv layer,  $\tilde{x}_i$  is the normalized version,  $\mu$  is the per channel mini batch mean,  $\sigma^2$  is the per channel mini batch variance and  $m$  is the mini batch size. The parameters  $\gamma$  and  $\beta$  are learned parameters, trained at the same time weights of the Conv layers are trained.

### Activation Layer

The activation layer removes all the negative components of the output feature function. Rectified Linear Units (ReLU) are often used to perform this computation. ReLU leaves the output same as the input if the input is greater than 0, and 0 if its less than or equal to 0. Figure 4.12 shows the function for ReLU . The equation for the ReLU is as follows:

$$f(y) = \max(0, y) \quad (4.16)$$

where  $f(y)$  is the output of the activation layer, and  $y$  is the the output from the previous layer.

This layer is placed in the CNN after every Conv layer and the pooling layer. There is also the Maxout activation function which takes the maximum value across the input and this facilitates optimization by dropout [83]. This ReLU layer replaces the sigmoid and the *tanh* activation functions used in the traditional neural network. The standard ReLU is used in this thesis because it is simple to implement and it also has a good performance.

### Fully Connected Layer

Fully connected layers in a neural network connects every neuron in one layer to every neuron in another layer. This uses the same principle as the Multi Layer Perceptron (MLP) used in the traditional neural network. Every output is the sum of every input multiplied by a corresponding weight. In CNNs they are often used at the end of the network, in the last one or two layers, when the data size has been significantly reduced by the Conv and pooling layers. The output layer is usually  $1 \times 1 \times c$  where  $c$  is the number of classes in the given classification task. Also, it is worthy to note that the fully connected layer used in this experiment uses features that were extracted on global level to perform the classification of the images.

### Learning Rate

Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. The lower the value, the slower the travel along the downward slope. The goal is to minimize the the error function as the weights is changing. The error function is the sum of all the errors from the mini batch:

$$L(W) = \frac{1}{m} \sum_{i=1}^m L_i(W) \quad (4.17)$$

where  $m$  is the mini batch size,  $L_i$  is the value of the loss function for the  $i^{th}$  mini batch and  $W$  are the network weights.

The gradient loss function is given as:  $\frac{\partial L(x_i, W_i)}{\partial W_i}$ . Therefore, the weight update is performed as

follows:

$$W_i = W_i - \frac{\alpha}{m} \sum_{i=1}^m \frac{\partial L(x_i, W_i)}{\partial W_i} \quad (4.18)$$

If the learning rate is too small, the gradient descent is slow. On the other hand, if the learning rate is too large, the gradient descent can overshoot the minimum, and may never converge or even start diverging. For the gradient descent, the weights are updated by following the direction of steepest descent, the negative gradient. The negative gradient is multiplied by the learning rate to scale. In this experiment, the decreasing logarithmic learning rate is used which is distributed between  $10^{-3}$  and  $10^{-4}$ .

#### 4.6.2 Error Function

In CNN, a back propagation of errors is an algorithm that uses gradient descent to update the networks error with respect to the weight. The gradient goes backward through the network with the gradient of the final layer of weights being calculated first, and the gradient of the first layer of weights being calculated last. This is repeated for many training instances or until the error reaches a desired minimum.

The error function calculates the margin between the output of the network and the desired output. The error function must be minimized. The most common error functions are Softmax log-loss and structured hinge-loss.

Softmax error finds the normalized probability of class  $X_c$  with respect to all other classes  $X_k$  and attempts to maximize the value of  $X_c$ . It normalizes the network predictions so that they can be interpreted as probabilities. The Softmax function is expressed as follows:

$$L_{softmax} = \log\left(\frac{e^{a_i}}{\sum_{k=1}^N e^{a_k}}\right) \quad (4.19)$$

The log-loss is related to the notation of entropy: since the network is predicting a probability distribution over labels for each input, the log-loss is equivalent to the cross entropy between the true label distribution and the network predictions. Entropy is a measure of the uncertainty in a given distribution. It tells how unpredictable the probability distribution is. Cross entropy

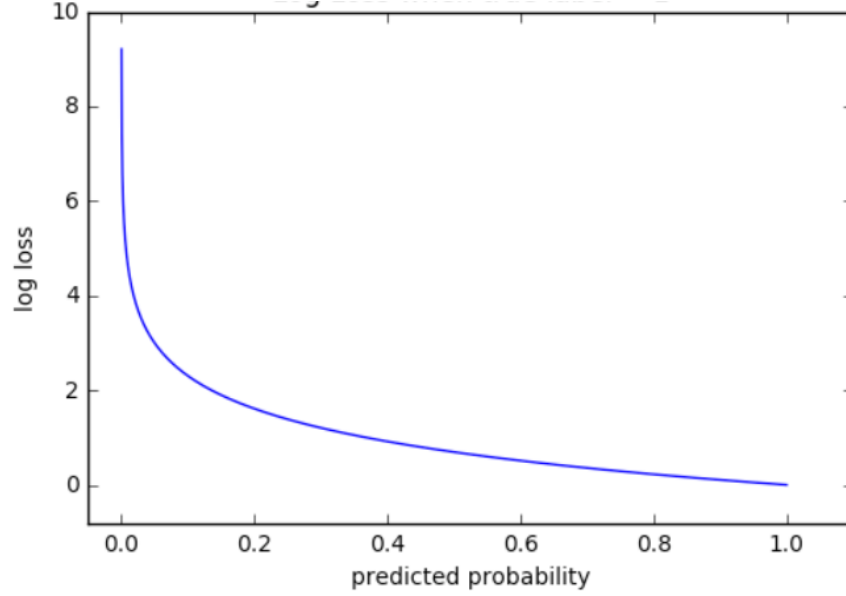


Figure 4.13: Log-loss when true label as explained in [84]

indicates the distance between what the model believes the output distribution should be, and what the original distribution is. As the predicted probability is approaching 1, the log-loss decreases, but as the predicted probability reduces close the zero, the log-loss increases very fast as shown in Figure 4.13.

The hinge-loss calculates the difference between the two closest outputs, and maximizes it. Every incorrectly classified data is assigned a loss value. It attempts to set the network output  $X_c$  to 1, and all other network outputs  $X_k$  to 0, by updating the weights as follows:

$$L_{shinge}(X, c) = \max \{0, 1 - (X_c - \max \{k! = c\} X_x)\} \quad (4.20)$$

### Mini Batch Size

The data used in training a neural network is split into small batches which is used to calculate error and update the weight. The size of the mini batch,  $m$ , determines how many images from the total training set will be combined in a batch. If the total size of the database is  $X$ , the number of training instances in each epoch will be  $\frac{X}{m}$ . Each epoch runs a different batch of the training data but use the same learning rate. The typical  $m$  value range between 16 and 256 and this depends on

the size of the data. The mini batch size increases the accuracy of the network and also reduces the training time. In the proposed CNN, we used random mini-batch size which varied through the training iteration but the average was around 132.

### Momentum

Momentum characterizes the rate of acceleration of gradients vectors in the right directions, thus leading to faster converging. This prevents the algorithm from falling into the local minima. The weight rule using momentum is as follows:

$$W_i = W_i - \frac{\alpha}{m} \sum_{i=1}^m \frac{\partial L(x_i, W_i)}{\partial W_i} + \rho \Delta W_i \quad (4.21)$$

where  $\rho$  is the momentum and  $\Delta W_i$  is the previous weight change. Here some information from the previous weight change is used to update the new weight. The momentum has to be greater or equal to zero and less than 1.

### Weight Decay

During training in deep learning, it is common to use weight decay, where after each update, the weights are scaled with a factor less than 1. This is done to gradually reduce the weight to avoid over-fitting. The weight decay is generally very small, in the range of  $10^{-3}$  to  $10^{-5}$ . The weight decay has a small impact on the accuracy in this system, and is set to  $10^{-5}$  in this study.

### Weight Initialization

Weight initialization gives the convolution layer an initial value. Using a random weight initialization with small standard deviation like 0.01 can lead to vanishing gradient [85]. Because neural networks are typically trained by minimizing a loss function with respect to the weights using gradient descent, the weight variance may lead to diminishing the output values. To solve the problem, the authors in [86] had to pre-train a model with 8 conv layers to initialize deeper models. Paper [85] suggested an updated weight initialization that restricts the size of the outputs by keeping the



variance of the input and the output of the Conv layer similar:

$$Var[y_i] = n_i Var[w_i x_i] \quad (4.22)$$

where  $n_i$  is the size of the convolution layer and  $y_i, w_i$  and  $x_i$  are the random variable that represents the input channel, the filter weights and the output. In this way, the updated weight initialization changes the standard deviation of the random value to:

$$\sigma_w = \sqrt{\frac{2}{n_i}} \quad (4.23)$$

where  $\sigma_w$  is the standard deviation of the weights, and  $n$  is the size of the  $i^{th}$  Conv layer. This initialization method avoid reducing or magnifying the magnitudes of input signals.

#### 4.6.3 Proposed Insobriety Detection System using CNN

In this thesis, we used the MATLAB in-built model as follows:

1. **Using a Pre-trained Network:** In this case, a network that was built and trained for a specific database such as network AlexNet is re-trained for a new data. The new data must be similar to the data it was trained, and the image size must be the same, too.. The input of the new database is fed into the network which will be used to classify the new data. Also, the output classes of the network and the new data must be the same.
2. **Fine-tuning of a Pre-trained network or Retraining a built model.** In the case of fine-tuning a pre-trained network, an already built network is used and modified to suit a completely new data. Some weights from the old data will be used but it will be fine-tuned so as to work well with the new data. Also, the output classes and the pre-processing techniques must not be the same. This technique works well for a network that requires a very large database. In the case of retraining a built CNN model, a network specific to the data size, number of samples and the output

classes is built. Every building block of the model is completely modified to suit the new data. This gives a relatively higher accuracy when compared to the other techniques.

The CNN structure proposed in thesis was built using the Matlab pre-built model. It was completely retrained and fine-tuned using our data. By doing this, it creates a network specific to the data size, number of samples, and output classes. Some of the difficulty encountered while completely retraining a network is that the training time is large, and there is also a difficulty in optimally selecting the parameters. We also used CNN for both the feature extraction and classification.

#### 4.6.4 Experiments

This section discusses the experiment that was performed using the system that was created. For training the system, the UPatras “Sober-Drunk” database was used for training, and additional collected data was used for testing. Leave-one-out cross-validation strategy was used for cross-validating the results.

To boost the accuracy of the system, the CNN parameters were optimized. The CNN parameters can greatly influence the accuracy of the system. Parameters were chosen based on the analysis of the previous works on the subject, as well as heuristically in the course of the experiment.

The proposed insobriety detection system is then compared with previous works. Although the results are often not directly comparable because of different testing sets and class separations, it can be used to infer which method has the highest accuracy.

The Sober-Drunk database collected by a team at the University of Patras, Greece was used for this experiment. The method of collection and labeling of this database has been described in section 4.5.3.

The experiments and the accuracy of the system is validated by using leave-one-subject-out validation, which in this case is equivalent to 40-fold cross validation, as there are 40 subjects.

This means that the subject used for testing were not used for training. In the first training set, all the frames from subject one is left out of the training set. The network is trained with the remaining 39 subjects. Then the accuracy is calculated by testing the network with all the frames from subject one. This is repeated for all the subjects, and the accuracy is calculated based on the number of the correctly classified images.

In the UPatras database, the number of drunk subject frames was greater than the number of sober ones. To verify that the system would work for both sober subjects, images from the sober subjects were collected in the Biometric Technologies Laboratory and used for testing.

### CNN Training

The proposed system was created using MATLAB R2018B pre built model (MatConvnet). All the 16,000 usable frames of 40 subjects from the University of Patras Drunk-Sober database were used. The images has a total of 4000 sober images and 12000 drunk images. 90% of the images was used for training and the remaining 10% was used for testing. Also, we followed leave-one subject-out cross-validation strategy which excludes the test subject from the training set. Based on this strategy, each network is trained on 39 subjects and tested on the remaining 1 subject, thus resulting in 40-iteration cross validation . We chose 40 fold cross-validation because we have 40 subjects. The aim was to exclude the subject used for the testing in our training set. The idea is that in real life scenario, subjects that will be tested on our system will most likely not be included in our training. For this experiment we used two different ROI - the forehead and the eyes.

For the images with forehead as ROI, the input of the network was normalized to a uniform size of  $45 \times 15$ . The first convolution layer in the CNN contains 20 convolution kernels with size of  $5 \times 5$ . The output of the convolution layer serves as the input to the first max pooling layer with  $2 \times 2$  filter and stride of 2. These steps are repeated, and the feature map is continuously reduced until it is passed to the fully connected layer as a feature vector for classification.

The final hyper-parameters used for this network are as follows: the mean filter size is  $5 \times 5$ , the average mini batch size of 132, the logarithmic learning rate is 0.0001, the momentum is 0.9,

the training/test ratio is 90/10, the number of epochs is 30 and the Softmax is used as the error function.

Using the images with eyes as the region of interest, we repeated the above steps but the size of the images were different. The images were normalized to a uniform size of  $25 \times 15$ . The final optimum hyper parameters are as follows: the mean filter of size  $3 \times 3$ , the average mini batch size of 132, the logarithmic learning rate with of 0.0001, the momentum of 0.9, the training/test ratio is 90/10, the number of epochs is 30 and the Softmax is used as the error function.

Figure 4.14 illustrated the rate the error of the training set goes down to 0. This is as a result of back-propagation to tune the weights which helps in better classification.

### Tuning the Parameters

This section describes how altering the parameters of the network impacted the accuracy of the network. The insights from other literature on CNN was used to select the optimal parameters that gave the highest accuracy of classification. Table 4.2 - Table 4.9 shows the results of experimentation with various values of the CNN parameter in order to select the optimal parameters. For each Table, one parameter is varied while the other parameters are kept constant. At the end, the optimal value of the varied parameter is selected. The parameters that have very little or no effects are not discussed.

Table 4.2: Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Number of epochs= 30; filter size =  $5 \times 5$ ; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Training/testing set = 90/10	91.50	95.50	94.50
Training/testing set = 80/20	78.98	90.50	87.62
Training/testing set = 50/50	97.00	72.50	78.63

It is also important to note that there is some randomness in the training of a CNN. The mini batches of data are randomly formed, and, therefore, are not the same between training instances. The effects of the parameter tuning is discussed as follows:

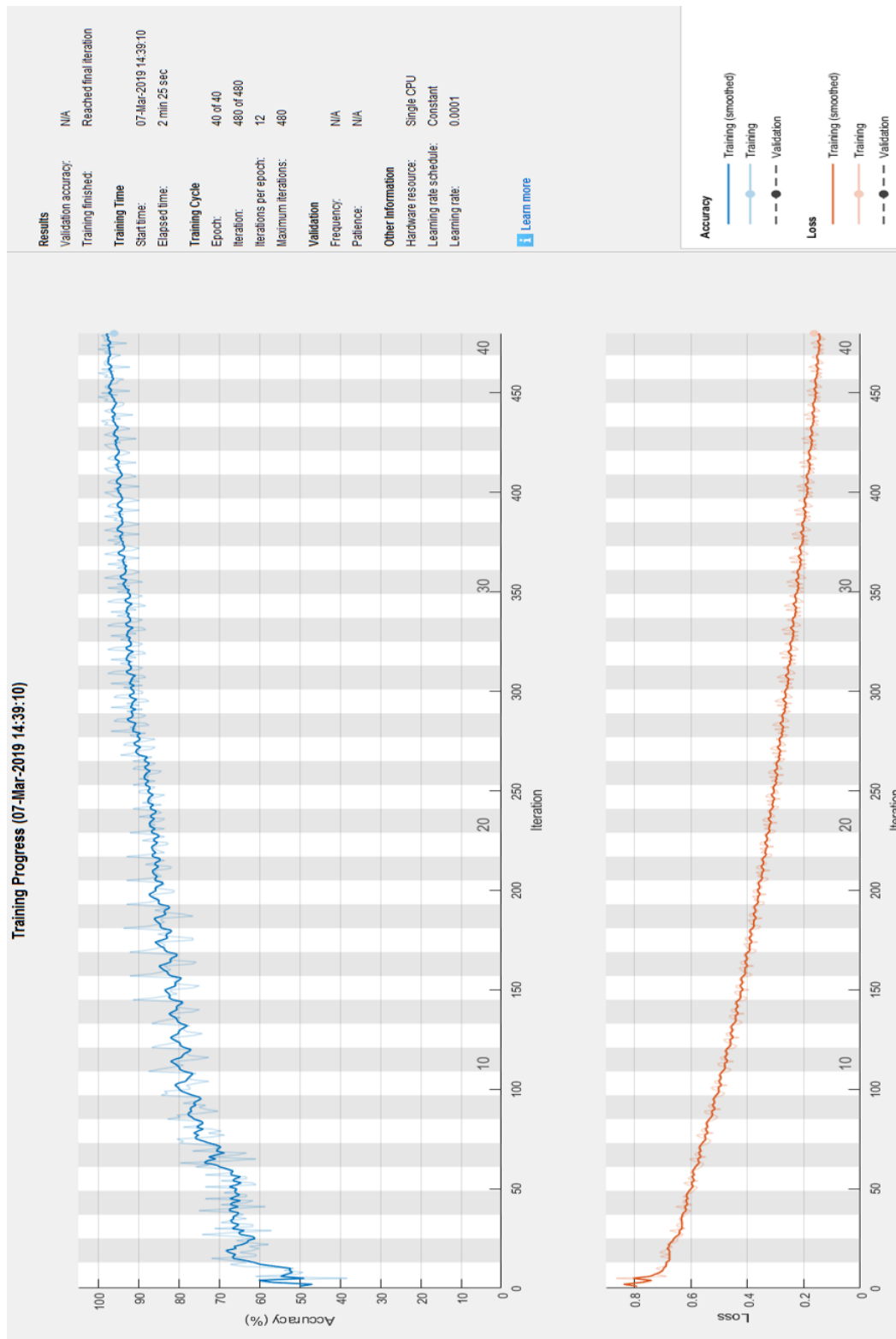


Figure 4.14: CNN Training in Matlab: The accuracy increases as the training goes through the epochs and updates the weight accordingly while the loss is going down to zero

Table 4.3: Tuning of hyper-parameters of the network built with forehead Images: 3 parameters are constant (Training/Testing Set = 90/10; filter size =  $5 \times 5$ ; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Number of Epochs = 40	95.50	94.92	95.06
Number of Epochs = 30	91.50	95.50	94.50
Number of Epochs = 20	92.50	87.50	88.75

Table 4.4: Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Training/Testing Set = 90/10; Number of Epochs = 30; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
$3 \times 3$ filter	87.50	95.17	93.25
$5 \times 5$ filter	91.50	95.50	94.50
$7 \times 7$ filter	92.75	77.33	81.19

- For the network using the forehead images, altering the percentage of data used for training and testing the system affected accuracy of classification. As we reduce the training data from 90% to 50% and increase the testing set data from 10% to 50%, the overall accuracy of the system dropped by about 15%. CNN learns better when trained with a larger size of image. This is the reason why the accuracy decreased as the training set data was reduced. Also, there was no need to augment the data because the images available was sufficient to properly train the network when 90% of the data was used for training.
- Modifying the learning rate also did have a considerable impact on the accuracy of the network. Learning rate controls how much the weight of the network is adjusted.

Table 4.5: Tuning of hyper-parameters of the network using forehead images: 3 parameters are constant (Training/Testing Set = 90/10; filter size =  $5 \times 5$ ; Number of Epochs = 30)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Learning Rate = 0.001	88.50	94.33	92.88
Learning Rate = 0.0001	91.50	95.50	94.50
Learning Rate = 0.00001	91.25	86.33	87.56

When the value is very small, the network will not converge fast and may get stuck. On the other hand if it is too large, the gradient descent can overshoot the minimum and it may never converge. When the learning rate was changed from the optimal value of 0.0001 to 0.001, the change in accuracy was about 2%, but when reduced to 0.00001, the accuracy reduced by 7%.

- Reducing the filter size had a moderate effect on the network whereas increasing the size affected the accuracy considerably. Using  $3 \times 3$  filter reduced the accuracy of the network by roughly 1% but when  $7 \times 7$  filter was used, the accuracy reduced significantly by 13.31%. It is also interesting to note that some subjects performed better with different filter size other than the optimal filter size of  $5 \times 5$  but we chose the  $5 \times 5$  filter because it gave the highest overall accuracy after the cross-validation.
- Changing the number of epochs in the network had a considerable effect on the accuracy. Ordinarily, increasing the number of epochs in the network should increase the accuracy of classification, but it also increases the training time. When the value was reduced to 20, the accuracy decreased by approximately 6%. Increasing the value to 40, increased the accuracy by a very small margin but it almost doubled the training time. Thus, the optimal number of epochs that gives an acceptable accuracy value with a considerable training time is 30.

Table 4.6: Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Number of epochs= 30; filter size =  $3 \times 3$ ; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Training/testing set = 90/10	93.25	96.67	95.81
Training/testing set = 80/20	88.38	91.90	91.02
Training/testing set = 50/50	94.25	86.25	88.25

The following observations were noted from the experiments on the network using the eye images:

Table 4.7: Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; filter size =  $3 \times 3$ ; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Number of Epochs = 40	95.25	96.17	95.97
Number of Epochs = 30	93.25	96.67	95.81
Number of Epochs = 20	89.75	91.92	91.38

Table 4.8: Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; Number of Epochs = 30; Learning Rate = 0.0001)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
$3 \times 3$ filter	93.25	96.67	95.81
$5 \times 5$ filter	88.75	96.00	94.19
$7 \times 7$ filter	90.75	91.17	91.81

- The optimum learning rate for this network is 0.0001. This value is neither too fast nor too slow for the weight adjustment of the network. When the value was increased to 0.001, the accuracy dropped by roughly 3% but when the value was reduced to 0.00001, the accuracy reduced by 5%.
- Adjusting the size of the CNN filter also affected the accuracy of classification. Using a filter size of  $5 \times 5$  yielded a classification accuracy 94.19% which is less than the optimum accuracy by 2% whereas a filter size of  $7 \times 7$  reduced the accuracy by 4%.
- Changing the number of epochs, which shows the number of times the learning

Table 4.9: Tuning of hyper-parameters of the network using the eye images: 3 parameters are constant (Training/Testing Set = 90/10; filter size =  $5 \times 5$ ; Number of Epochs = 30)

Parameters Altered	Sober Accuracy	Drunk Accuracy	Overall Accuracy
Learning Rate = 0.001	85.50	95.08	92.69
Learning Rate = 0.0001	93.25	96.67	95.81
Learning Rate = 0.00001	92.00	90.08	90.56



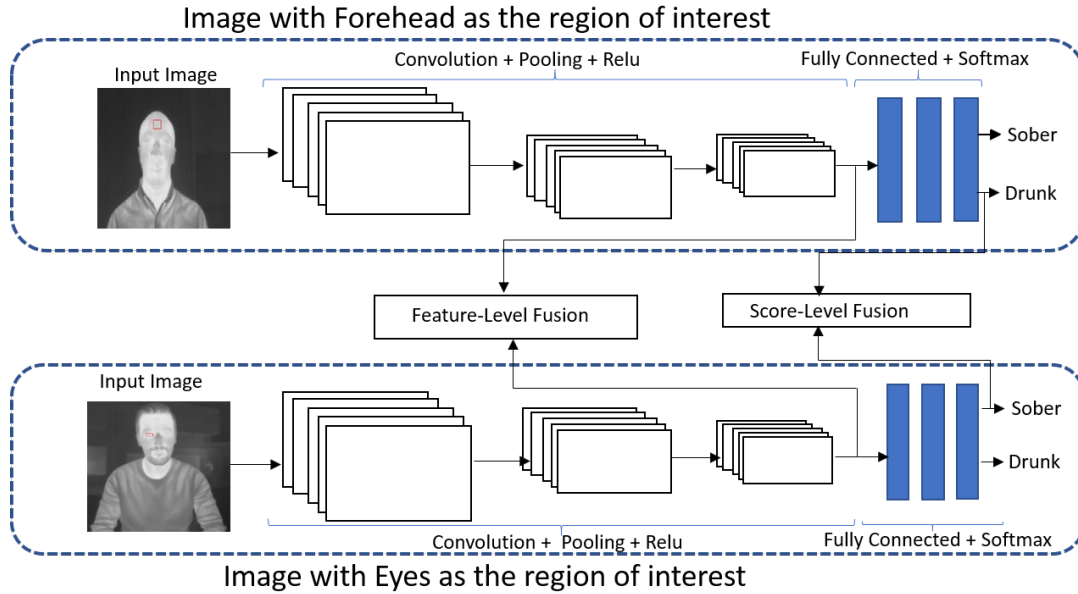


Figure 4.15: Architecture of the system showing the stages where fusion was implemented

algorithm will loop through the network, had a considerable effect on the accuracy. When the value was reduced to 20, the accuracy dropped by roughly 4.5%, but when it was increased the 40, the change in accuracy was less than 1%. Thus, 30 is still the optimum number of epoch for this network.

- Lastly, changing the training/testing ratio from 90/10 to 50/50 ratio reduced the accuracy by roughly 8%.

#### 4.6.5 Fusion

After the parameter tuning to obtain the optimum parameters for the forehead and eyes images, the next step is to fuse the two networks together to boost the accuracy of classification. To further improve the performance of the system, we performed fusion of the results of the two CNNs using the score-level and the feature-level approach. This is shown in Figure 4.15. The feature level fusion was performed using Discriminant Correlation Analysis [87]. For the score-level fusion, the scores for prediction was extracted, then the average and the maximum values was computed respectively. These values were used for the overall class prediction.

#### 4.6.6 Performance evaluation and comparison to other works

The confusion matrix of the classification is presented in Figure 4.16. This matrix shows the total correctly and incorrectly classified subjects for the 40 iterations of the 40-fold cross-validation. Equation 4.9 was used to calculate the accuracy of classification.

We compared our results with previous studies conducted on the same database. In [16], an approach called Local Difference Pattern was applied that detect regions of temperature variation on the face images of the sober and drunk subject, with the reported “identification success” of around 85%. An approach that combines both the detection of the temperature changes in thermal image of iris and sclera regions (using SURF, Speeded up Robust Features) as well as motion trajectories of drunk and sober persons proposed in [79] yielded 89.23%. A three-layer feed-forward Neural Network (NN) proposed in [80] was trained on the whole face as well as on the forehead of the sober and the drunk subjects. The experimental procedure was not common, since “a training with the data from one person and testing with those from the rest of the persons was applied”, with the average “success rate” being 93.1% and 84.2% using the thermal pattern of the forehead region of sober and drunk subjects (only five subjects were used for experiment). Finally, the “classification success rate” reported in [81] that used fusion of dissimilar features by means of a 2-layer neural network with one neuron in the second layer achieved 66.6% for 2 neurons in the first layer, 81.5% for 4 neurons in the first layer, and 96.2% with 8 neurons on the first layer (this result is shown in Table 4.10 ).

Our approach using the CNN in five different experiments (the bottom five row in Table 4.10) performed as follows:

- The first experiment was conducted using only images of the forehead region. The forehead region is easily accessible and a thermal image can be captured without invading the privacy of the subject. From the network built with this region, the overall accuracy of classification was 94.5%.

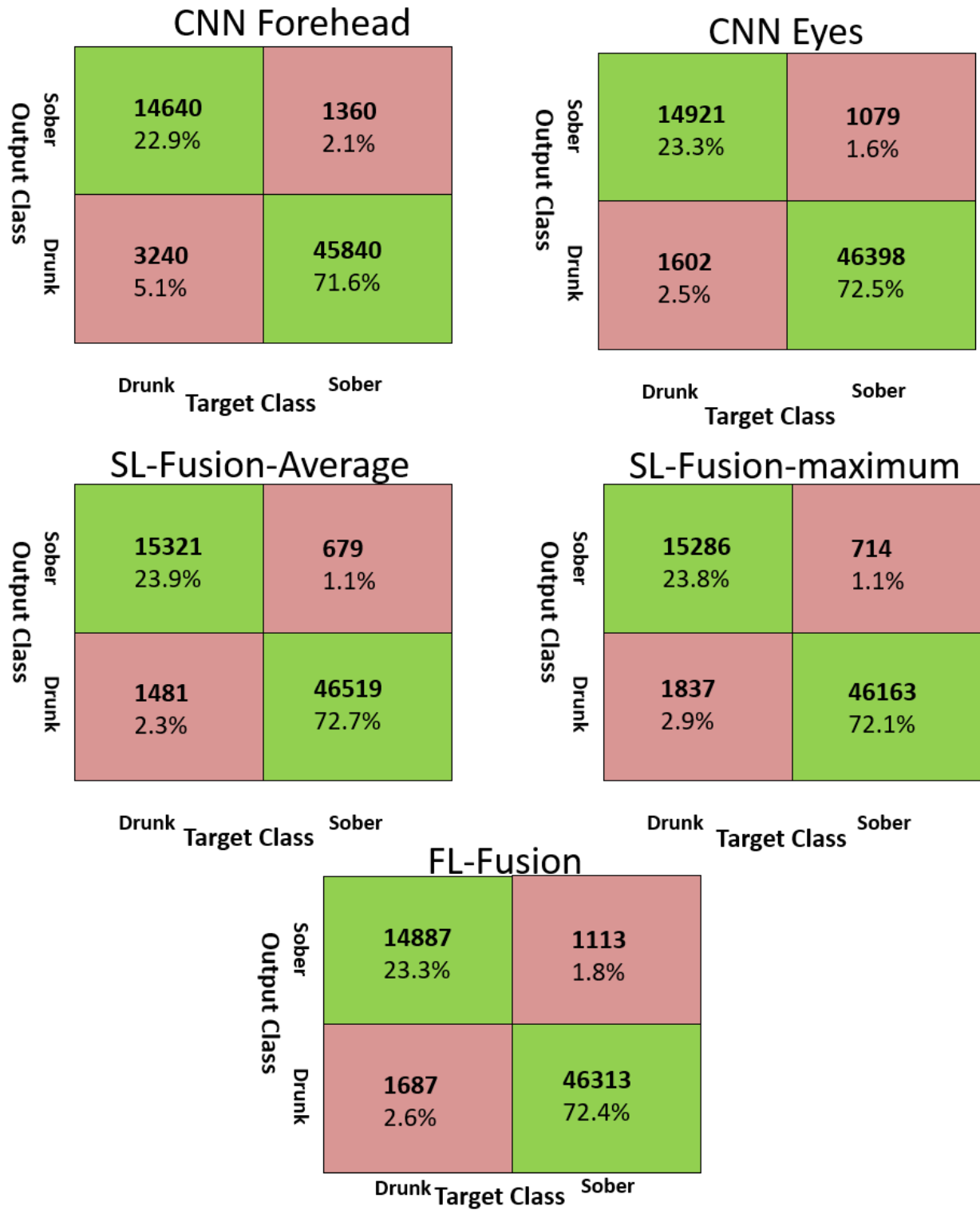


Figure 4.16: Confusion Matrix for each experiment for 40 iterations

- The second experiment was done using the eyes as the region of interest, This is more detailed and more invasive but the accuracy of classification was better than the one using the forehead region. The overall accuracy of classification was 95.81%
- To boost the accuracy of the first two experiments, the fusion strategy was applied. In this third experiment, the score from the network created for the forehead and eye regions were fused together. The higher value of the score was used in prediction (Score level-fusion-Maximum). This experiment gave a slightly higher accuracy value of 96.00%.
- The further enhance the accuracy, another experiment was conducted using the same score-level fusion strategy but in the case the average of the scores were used for prediction. The resulted to an accuracy of 96.63%.
- Lastly, a feature-level fusion was performed using the principle of Discriminant Correlation Analysis [87]. The fusion was done after the features were extracted and the merged features were used for classification. This experiment resulted in the classification accuracy of 95.63%.

Table 4.10 shows a comparison of the accuracy achieved in our study with that of other researchers.

Analyzing the results in Table 4.10 the CNN that used forehead region (CNN (Forehead)) and eyes region (CNN (Eyes)), without fusion, performed better than the first two reported approaches but worse than the third one, except for the accuracy in detecting the drunk subjects. The accuracy of classifying a drunk subjects was generally higher than classifying a sober person when using CNN. One of the reasons is that the number of images of drunk subjects used for training exceeded the number of images of sober subjects. Note that the CNNs have been proven to perform better in classification when a larger datasets are used to train the network, and the available database is of

Table 4.10: Comparison of the accuracy for different methods of sobriety classification

Method	Sober Accur.	Drunk Accur.	Overall Accur.
Local Diff. Pattern [14]	N/A	N/A	85%
SURF Features [79]	N/A	N/A	89.23%
Three-layer NN [80]	93.1%	84.2%	N/A
NN-Fusion Diss.Feat. [81]	N/A	N/A	96.2%
CNN (Forehead)	91.50%	95.50%	94.50%
CNN (Eyes)	93.25%	96.67%	95.81%
SL-fusion-Average	95.76%	96.91%	96.63%
SL-fusion-Maximum	95.50%	96.17%	96.00%
FL-fusion	93.04%	96.49%	95.63%

rather medium size. For the classification using fusion, the CNN showed a superior performance: the CNN with score-level fusion performed the best, with the average-based fusion (CNN-SL-fusion- Av) showing slightly better results than the maximum-based score-level fusion (CNN-SL-fusion-Max). The CNN with the feature-level fusion (CNN-FL-fusion) did not performed better than one with the network with score-level fusion. The average accuracy was the best for the CNN with the average feature-level fusion.

#### 4.6.7 Real-life Experiments

This section describes a real-life application of our insobriety detection system which is designed for the described application. The purpose of this study is to study the feasibility of applying an automatic insobriety detection algorithm in a real world scenario. The implementation was designed and tested in the Calgary Drop-in Center. A camera placed is placed in front of the volunteer subject who is seated or standing. The algorithm used is the proposed CNN from the previous sections, trained on the 40 subjects from the UPatras database. Ethics approval from University of Calgary's Conjoint Health Research Ethics Board (CHREB) was acquired for this study. The subjects recruited for this experiment were staff of the Calgary drop in center. The clients of the shelter were not part of the study. The recruitment was done by the director of the IT department and the staff were more than willing to be part of the study.

Consent forms were created for the participants and they were briefed on the implications of

participating in the study. One of the probable reasons why they were comfortable in participation is the fact that their identity cannot be easily found from an infrared image. The study is not invasive and the procedure was very fast.

Also, additional infrared images were taken at the Biometrics technologies lab at the University of Calgary. The infrared images were taken as the subjects were standing on front the camera. Only general statistics was used in publications, with no identifiable information. Once the study was completed, the facial images were destroyed.

## Setup

The Thermoteknix infrared camera was used for recording thermal videos of the subjects. The camera takes a continuous infrared videos of clients and it saves in the DVR after which it is copied in a flash drive. For data collection in the Biometric Technologies Laboratory, the subject was standing in front of the infrared camera placed on a 7 foot tripod as shown in Figure 4.17. Afterward the forehead temperature of the subjects was taken with Euat forehead thermometer, in order to have a ground truth data that we be compared against the temperature value estimated using the infrared camera.

An application was created to make the process of detecting thermal abnormalities easy for the user. This application was created with Matlab to test the feasibility.

The interface of the created application is shown in Figure 4.18.

In the application, the region of interest is selected using the 'select the region of interest' tab. This enables the user to select the region of interest that will be used a to extract the health status if the subject. Once this is selected, the 'get temperature' button is used to display the estimated temperature of the subject. The "status check" button is used to display the information whether the subject's temperature is normal or not, and the "sobriety" button is used to confirm whether the subject is sober or drunk.

With the information provided for the user, a decision is made as to what type of service the subject requires. After all the information is extracted, the application interface shall be as shown

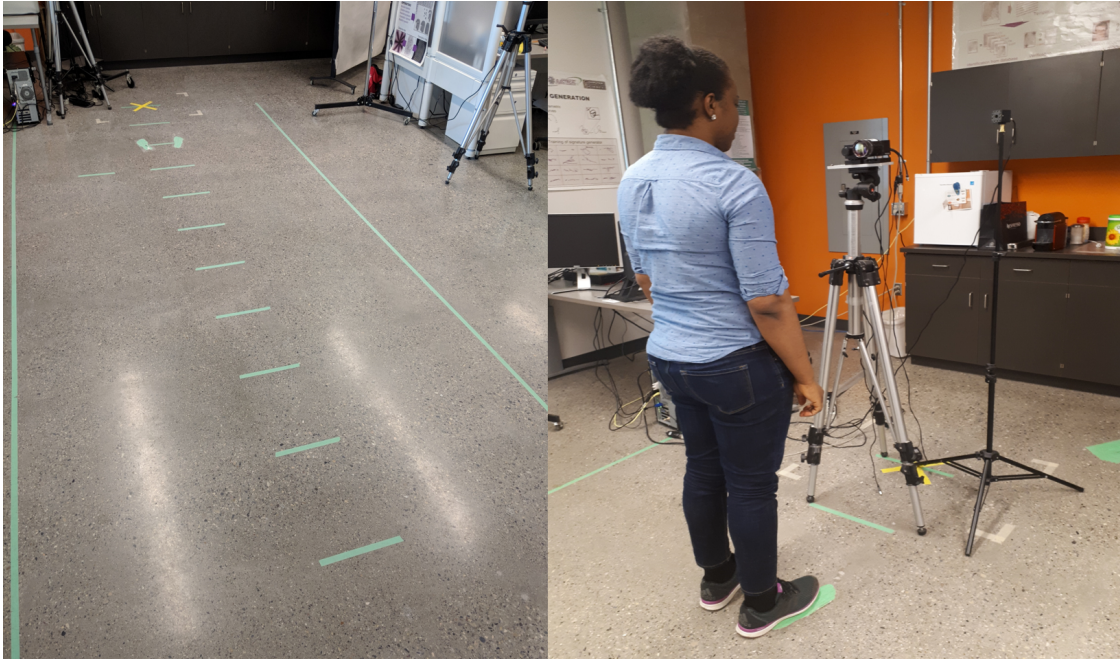


Figure 4.17: Example of a subject standing in front of the infrared camera as the infrared image recording was taken

in Figure 4.19

#### 4.6.8 Results

Of the 55 “sober” subjects recruited for this experiment, 51 were correctly classified as sober subjects while 4 were mis-classified as drunk subjects. The accuracy of the system is relatively high.

### 4.7 Conclusion

This chapter presents the details of implementing a system that can detect sobriety of a subject by using the thermal image of the face. Four experiments were performed, and the accuracy was evaluated for each.

- The first experiment was performed using the forehead as the region of interest. Feature extraction was performed using LDP, in order to determine how each pixel

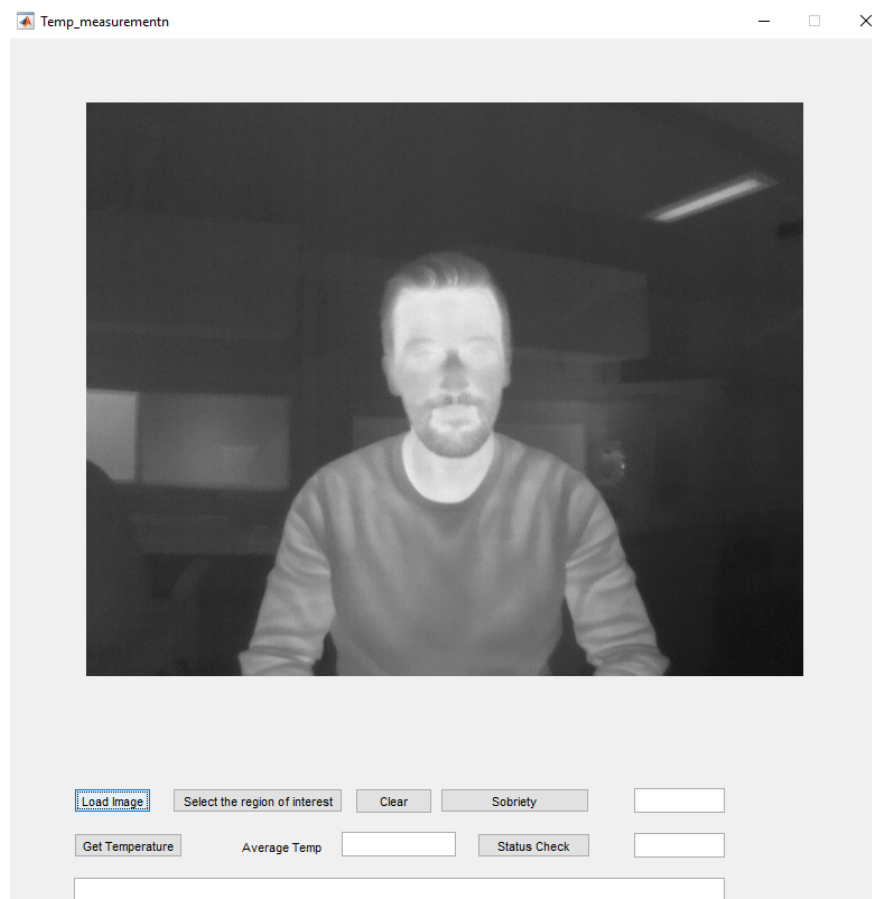


Figure 4.18: The Interface of the application after the thermal image is uploaded



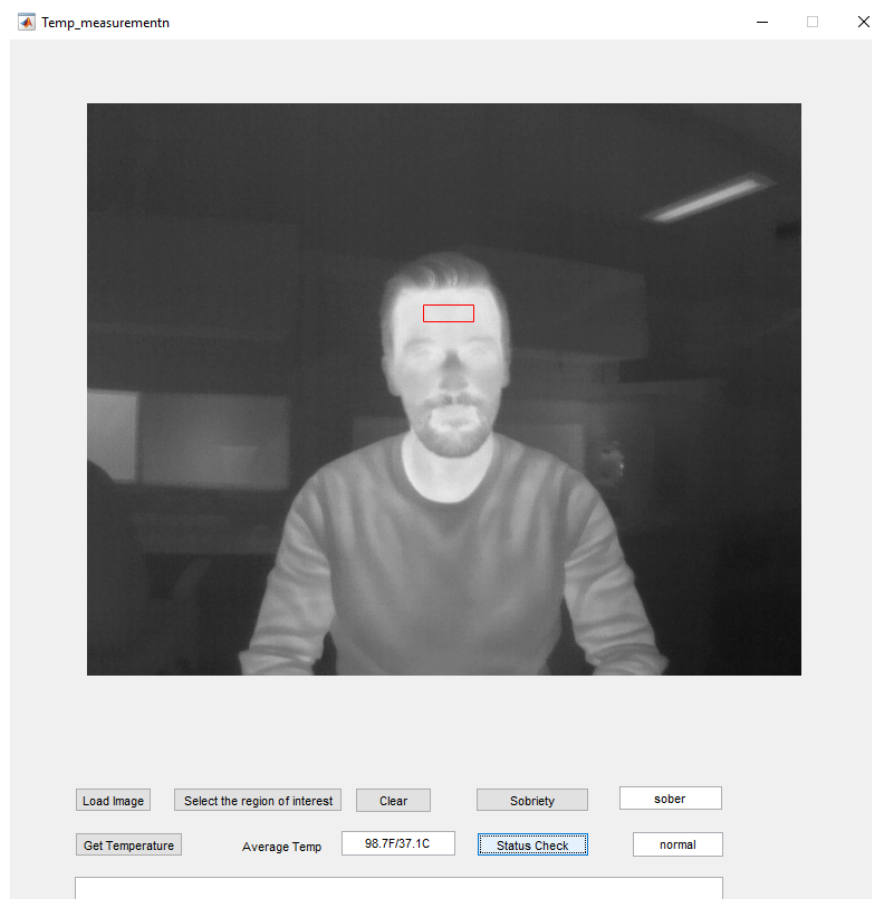


Figure 4.19: The Interface of the application with the necessary information for the user

intensity in the thermal image varies compare to the neighboring pixels. After the features were extracted, classification was done with SVM. This resulted in an overall accuracy of 81.13%.

- In the second experiment, CNN, a deep learning technique, was used for the same classification, with the purpose to improve the accuracy to an acceptable value. CNN was used for both feature extraction and classification. In this experiment, another region of interest was introduced which is the eyes. This experiment gave a better classification accuracy than from the first experiment. The network that used forehead images yielded an accuracy of 94.50%, whereas the network that used the eye image had an accuracy of 95.81%.
- The third experiment was to fuse the two results obtained in experiment 2. Two major fusion strategies were applied - Feature-level fusion which gave an accuracy of 95.63% and score level fusion which gave an accuracy of 96.00%.
- The final experiment was to build a Matlab GUI application that was used to test the feasibility of the experiment on real life subjects and database images. It was tested using the full valid database, allowing future researchers to easily compare the results. For the real-life experiment, sober subjects were the main recruits and the classification accuracy was relatively high: 51 out of the 55 subjects that participated in this experiment were correctly classified.



# Chapter 5

## Conclusion and Future work

This chapter presents the summary of the important results and novelties in this thesis, and its contribution to developing a concept of health monitoring system for emergency shelter.

### 5.1 Summary of Results

The main objective of this thesis was to develop a biometric-based approach that would help the Calgary DI Center to expand their access system. In this regard, the health status of the client is of utmost importance in addition to identifying the sobriety level of the clients. Fever is the most common symptom of most illnesses, so the focus of this research was on identifying subjects with abnormal (f.e. elevated) temperature as that would help in identifying client in medical distress. The application created in this study can automatically estimate the temperature of a subject, as well as their sobriety level.

The following are the main conclusions that can be drawn from the experiments conducted:

- In chapter 3, the proposed system for temperature extraction was introduced. Using the thermal image from the subject, the principle of polynomial curve fitting was used to detect the temperature value using thermal camera readings. The temperature reading from the system was compared to the value from a standard forehead thermometer. The results were analyzed using a Bland-Altman plot. This shows a strong correlation between the values estimated using the proposed system and the actual temperature of the subjects.
- The second experiment was to identify the sobriety level of a subject. This was done using two machine learning approaches - Support Vector Machine and Convolutional Neural Network. In the experiment using SVM, the feature extraction

was done using a technique known as Local Difference Patter (LDP). There was a clear distinction between these values for a sober and drunk subject. The classification using SVM gave an accuracy of about 85%.

In order to boost this accuracy, CNN approach was used using the face and the eyes as the region of interest. The accuracy using the eyes as the region of interest gave about 95.5%, whereas the one using the forehead region gave an accuracy of 94.5%.

To further increase the classification accuracy, the classifiers using the forehead and the eyes were combined at both feature level and the score level fusion. We implemented both average and max score level fusion. Using this fusion strategy improved the classification accuracy to 96.63%.

- Finally, an application was built to perform a real life experiment. The application was used for 55 subjects and 51 of them were classified correctly. This shows that a real life implementation of the proposed systems is feasible.

## 5.2 Suggestions for future work

As shown through out this study, thermal images can provide a reliable information that could be used to access the health and the sobriety status of a subject. The system for detecting alcohol intoxication using CNN has a significant improvement over the system built using SVM. However, there still room for improvement. One thing that could be considered in future research is the use of Generative Adversarial Network (GAN) to generate multiple synthetic thermal images to boost the training set of the system. Neural Networks work better on large databases and since there is limitation of thermal images, GAN might be helpful in solving the problem of unavailability of very large dataset. Also, other regions of interest could be considered in addition to the eyes and forehead that was used in this thesis. The cheek or the neck regions could be studied for possible effect on the blood vessels found there.

In addition, a system that uses real time infrared camera should be used for experiment to check if real time collection and analysis of images will yield a high accuracy than the one used in this thesis. It will also be helpful to these the extreme environmental conditions to check how feasible this solution is on subjects that has been exposed.

Lastly, automatic detection of the forehead will be very helpful and could make the system work independent of any human interference. To automatically detect forehead, the eyes, nose of the human face can be used. There has been a lot of research and breakthrough in detecting some portions of the face like the eyes, nose, mouth and other features. Knowing this, and also the fact that the forehead is right above the eyes and the exact region of interest can be centered on top of the nose region, the region of the forehead that will be used can be detected automatically

# Bibliography

- [1] M. Shinn, “How psychologists can help to end homelessness,” *Australian Psychological Society*, 2009.
- [2] <https://www.un.org/ruleoflaw/files/FactSheet21en.pdf>, accessed on 2019-04-15.
- [3] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. Nixon, “Soft biometrics and their application in person recognition at a distance,” *IEEE Trans. Inf. Forensic and Sec*, vol. 9, no. 3, pp. 464–475, 2014.
- [4] E. Ng, G. Kawb, and W. Chang, “Analysis of infrared thermal imager for mass blind fever screening,” *Microvascular Research*, p. 104–109, 2004.
- [5] K. MM and C. FY, “Airport sentinel surveillance and entry quarantine for dengue infections following a fever screening program in taiwan,” *BMC Infectious Diseases*, vol. 12, p. 182, May 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2168874.2168907>
- [6] Y. H. Tan, C. W. Teo, E. Ong, L. B. Tan, and M. J. Soo, “Development and deployment of infrared fever screening systems,” *Thermosense XXVI, Proc. SPIE*, vol. 5405, pp. 68–78, 2010.
- [7] G. Koukiou and V. Anastassopoulos, “Drunk person identification using thermal infrared images,” *16th International Conference on Digital Signal Processing*, 2009.
- [8] A. J. Collins and J. A. Cosh, “Temperature and biochemical studies of joint inflammation,” *Annals of the Rheumatic Diseases*, p. 386–392, 1970.
- [9] G. Varju, C. Pieper, J. Renner, and V. Kraus, “Assessment of hand osteoarthritis: correlation between thermographic and radiographic methods,” *Rheumatology (Oxford)*, 2004.
- [10] K. Ammer, “Cold challenge to provoke a vasospastic reaction in fingers determined by temperature measurements: a systematic review,” *Thermology International*, pp. 109–118, 2009.
- [11] R. Lawson, “Implications of surface temperatures in the diagnosis of breast cancer,” *Can Med Assoc J*, p. 309–310, 1956.

- [12] S. G.Kandlikarl, I. Perez-Raya, and P. A. et. al, “Infrared imaging technology for breast cancer detection – current status, protocols and new directions,” *International Journal of Heat and Mass Transfer*, vol. 108, pp. 2303–2320, 2017.
- [13] G. Koukiou and V. Anastassopoulos, “Facial blood vessels activity in drunk persons using thermal infrared,” *Proceedings of the 4th International Conference on Imaging for Crime Detection and Prevention*, 2011.
- [14] —, “Drunk person identification using local difference patterns,” *IEEE International Conference on Imaging Systems and Technology(IST)*, 2016.
- [15] —, “Face locations suitable drunk persons identification,” *International Workshop on Biometrics and Forensics (IWBF)*, 2013.
- [16] —, “Eye temperature distribution in drunk persons using thermal imagery,” *International Conference of the BIOSIG Special Interest Group (BIOSIG)*, 2013.
- [17] —, “Neural networks for identifying intoxicated persons,” *Forensic Science International*, vol. 252, pp. 69–76, 2015.
- [18] P. Viola and M. J. Jones, “Robust real-time face detection,” *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [19] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886, 2012.
- [20] I. Pavlidis, P. Buddharaju, C. Manohar, and P. Tsiamyrtzis, “Biometrics: Face recognition in thermal infrared,” *Medical Systems and Devices*, vol. 29, pp. 1–16, 2005.
- [21] W. Li, M. Li, Z. Su, and Z. Zhu, “A deep-learning approach to facial expression recognition with candid images,” 2015, pp. 279–282.
- [22] S. G.Kandlikarl, I. Perez-Raya, and P. A. et. al, “Thermal imaging as a biometrics approach to facial signature authentication,” *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 214–222, 2013.



- [23] Y. N. D. Chenna, P. Ghassemi, T. J. P. 1, and J. C. Q. Wang, "Free-form deformation approach for registration of visible and infrared facial images in fever screening," *Sensors*, p. 104– 109, 2018.
- [24] T. Cover and P.Hart, "Nearest neighbor pattern classification," *IEEE Transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [25] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.
- [26] E. Sami, "Support vector machines for classification and locating faults on transmission lines," *Applied Soft Computing*, vol. 12, pp. 1650–1658, 2012.
- [27] J. Nayak, B. Naik, and H. Behera, "A comprehensive survey on support vector machine in data mining tasks: Application and challenges," *International Journal of Database Theory and Application*, vol. 8, pp. 169–186, 2015.
- [28] T. Joachis, "Learning to classify text using support vector machines," *Kluwer Academic Publishers*, 2002.
- [29] K. Polat and S. Gunes, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing*, vol. 17, no. 4, pp. 694–701, 2007.
- [30] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural computation*, vol. 15, no. 7, pp. 1667–1689, 2003.
- [31] G. Valentini, "Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles," *Artificial Intelligence in Medicine*, vol. 26, no. 3, pp. 281–304, 2002.
- [32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffne, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

- [35] I. Song, H. J. Kim, and P. B. Jeon, “Deep learning for real-time robust facial expression recognition on a smartphone,” *Consumer Electronics (ICCE), 2014 IEEE International Conference*, pp. 564–567, 2014.
- [36] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, “Dexpression: Deep convolutional neural network for expression recognition,” *arXiv preprint arXiv:1509.05371*, 2015.
- [37] A. Lopes, E. de Aguiar, and T. Oliveira-Santos, “A facial expression recognition system using convolutional networks,” 2015, pp. 273–280.
- [38] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hu, “A convolutional neural network cascade for face detection,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5325–5334, 2015.
- [39] S. Alizadeh and A. Fazel, “Convolutional neural networks for facial expression recognition,” *arXiv preprint arXiv:1704.06756*, 2017.
- [40] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” *BMVC*, vol. 1, p. 6, 2015.
- [41] F. Schroff, D. Kalenichenko, and J. Philbin, “acenet: A unified embedding for face recognition and clustering,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.
- [42] K. Nguyen, C. Fookes, and S. Sridharan, “Improving deep convolutional neural networks with unsupervised feature learning,” *IEEE International Conference on Image Processing (ICIP)*, pp. 2270–2274, 2015.
- [43] K. He, X. Zhang, S. Ren, and J. Sun, “Deep learning for real-time robust facial expression recognition on a smartphone,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” *AAAI*, pp. 4278–4284, 2017.
- [45] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to biometrics*. Springer Science & Business Media, 2011.

- [46] M. He, S.-J. Horng, P. Fan, R.-S. Run, R.-J. Chen, J.-L. Lai, M. K. Khan, and K. O. Sentosa, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010.
- [47] M. Vatsa, R. Singh, and A. Noore, "Improving iris recognition performance using segmentation, quality enhancement, match score fusion and indexing," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 1021–1035, 2008.
- [48] Z. Wang, Q. Han, Q. Li, X. Niu, and C. Busch, "Complex common vector for multimodal biometric recognition," *Electronic Letters*, vol. 45, 2009.
- [49] K. Hollingsworth, T. Petrs, and Patrick J. Flynn, "Iris recognition using signal-level fusion of frames from video," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, p. 837–848, 2009.
- [50] S. Naveen and R. Moni, "A robust novel method for face recognition from 2d depth images using dwt and dft score fusion," *IEEE First International Conference on Computational Systems and Communications (ICCSC)*, pp. 1–6, 2014.
- [51] Y. H. Park, D. N. Tien, E. C. Li, S. M. Kim, and H. C. Kim, "A multimodal biometric recognition of touched fingerprint and finger-vein," *IEEE International Conference on Multimedia and Signal Processing*, pp. 247–250, 2011.
- [52] R. E. F., "Quantitative thermal imaging," *Clinical Phys. Physiol. Meas.*, vol. 11, pp. 87–95, 1990.
- [53] D. C. A., "Thermography and the possibilities for its applications in clinical and experimental dermatology," *Clinical Dermatology*, vol. 13, pp. 329–336, 1995.
- [54] L. Chan, G. Cheung, I. Lauder, and C. Kumana, "Screening for fever by remote sensing infrared thermographic camera," *Journal of travel medicine*, vol. 11, pp. 273–279, 2004.
- [55] C. Liu, R. Chang, and W. Chang, "Limitations of forehead infrared body temperature detection for fever screening for severe acute respiratory syndrome," *Infection Control*, vol. 25, pp. 1109–1111, 2004.

- [56] E. Ring, A. Jung, J. Zuber, P. Rutowski, B. Kalicki, and U. Bajwa, "Detecting fever in polish children by infrared thermography," *Proceedings of the 9th International Conference on Quantitative Infrared Thermography*, pp. 2–5, 2008.
- [57] A. Nguyen, N. Cohen, H. Lipman, C. Brown, N. Molinari, W. Jackson, and D. Fishbein, "Comparison of 3 infrared thermal detection systems and self-report for mass fever screening," *Emerging infectious diseases*, vol. 16, p. 1710, 2010.
- [58] P. A. Mackowiak, "Physiological rationale for suppression of fever," *Clinical infectious diseases*, pp. 185–189, 2000.
- [59] P. B. Persson, "Energie-und wärmehaushalt, thermoregulation," *Physiologie des Menschen*, pp. 906–927, 2010.
- [60] J. Güttler, C. Georgoulas, and T. Bock, "Contactless fever measurement based on thermal imagery analysis," *IEEE Sensors Applications Symposium (SAS)*, 2016.
- [61] P. Perera, M. Fernando, S. Meththananda, and R. Samaranayake, "Accuracy of measuring axillary temperature using mercury in glass thermometers in children under five years: A cross sectional observational study," *Health*, vol. 6, no. 16, 2014.
- [62] J. Lefrant, L. Muller, J. Coussaye, M. Benbabaali, C. Lebris, C. Zeitoun, C. Mari, G. Saissi, J. Ripart, and J. Eledjam, "Temperature measurement in intensive care patients: comparison of urinary bladder, oesophageal, rectal, axillary, and inguinal methods versus pulmonary artery core method," *Intensive care medicine*, 2003.
- [63] A. El-Radhi and S. Patel, "An evaluation of tympanic thermometry in a paediatric emergency department," *Emergency Medicine Journal*, vol. 23, pp. 40–41, 2006.
- [64] S. Smitz, T. Giagoultsis, W. Dewé, and A. Albert, "Comparison of rectal and infrared ear temperatures in older hospital inpatients," *Journal of the American Geriatrics Society*, vol. 48, pp. 63–66, 2000.
- [65] G. Gasim, I. Musa, M. Abdien, and I. Adam, "Accuracy of tympanic temperature measurement using an infrared tympanic membrane thermometer," *BMC research note*, 2013.

- [66] J. Kistemaker, E. Hartog, and H. Daanen, “Reliability of an infrared forehead skin thermometer for core temperature measurements,” *Journal of medical engineering and technology*, pp. 252–261, 2006.
- [67] L. Simonsen, M. Clarke, L. Schonberger, N. Arden, N. Cox, and K. Fukuda, “Pandemic versus epidemic influenza mortality: a pattern of changing age distribution,” *Journal of Infectious Diseases*, vol. 178, pp. 53–60, 1998.
- [68] H. Nishiura and K. Kamiya, “Fever screening during the influenza (h1n1-2009) pandemic at narita international airport,” *BMC infectious diseases*, vol. 11, p. 111, 2011.
- [69] P. Guest, *Numerical Methods of Curve Fitting*, 2012.
- [70] H. Motulsky and A. Christopoulos, *Fitting Models to Biological Data using Linear and Nonlinear Regression*, 2003.
- [71] D. G. Altman and J. M. Bland, “Measurement in medicine: the analysis of method comparison studies,” *The Statistician*, vol. 32, no. 3, pp. 307–317, 1983.
- [72] <http://www.physics.upatras.gr/sober/>, accessed on 2018-02-25.
- [73] M. Verma and B. Raman, “Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval,” in *Multimedia Tools Application*, 2017.
- [74] T. Ojala, M. Pietikainen, and D. Hardwood, “A comparative study of texture measures with classification based on feature distributions,” *Pattern Recognition*, pp. 51–59, 1996.
- [75] A. Ben-Hur and J. Weston, “A users guide to support vector machines,” *Data mining techniques for the life sciences*, pp. 223–239, 2010.
- [76] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [77] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, “A practical guide to support vector classification,” 2003.
- [78] <http://francescopochetti.com/support-vector-machines/>, accessed on 2019-01-25.

- [79] M. Bhuyan, P. S. S. Dhawle, and G. Koukiou, “Intoxicated person identification using thermal infrared images and gait,” *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 1–3, 2018.
- [80] K. Georgia and V. Anastassopoulos, “Neural networks for identifying drunk persons using thermal infrared imagery,” *Forensic Science International*, no. 252, pp. 69–76, 2015.
- [81] —, “Fusion using neural networks for intoxication identification,” *International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–5, 2018.
- [82] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1414.6806*, 2014.
- [83] I. Goodfellow, D. Warde-Farley, and M. Mirza, “Maxout networks,” *arXiv preprint arXiv:1302.4389*, 2013.
- [84] [http://wiki.fast.ai/index.php/Log\\_Loss](http://wiki.fast.ai/index.php/Log_Loss), accessed on 2019-02-25.
- [85] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- [86] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [87] M. Haghghat, M. Abdel-Mottaleb, and W. Alhalabi, “Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996, 2016.

# **Appendix A**

## **Copyright Permissions**

In order to conduct the research proposed in this thesis, it was necessary to obtain external biometric datasets. All utilized datasets were collected under explicit user agreement of academic use on behalf of the Biometric Technologies Lab, Department of Electrical & Computer Engineering, University of Calgary. All databases are acknowledged, cited, and have been used only for academic research purpose. The agreements are presented in the following sections:

### **A.1 UPatras Database**

This research utilized the Upatras dataset collected by the by the University of Patras Computer Vision Group. The UPCV dataset is a publicly available thermal dataset and free for research and scientific purposes. The dataset can be downloaded from <http://www.physics.upatras.gr/sober/>. The database has been downloaded on-behalf of Biometric Technologies Lab. In this thesis, the database is cited and credit is given to the owner of the dataset.

## **Appendix B**

### **Ethics Approval**