Computer Vision Based Methods for Detection and Measurement of Psychophysiological Indicators

Book based on dissertation

Irani, Ramin

Publication date:
2017

Document Version
Early version, also known as pre-print

Computer Vision Based Methods for Detection and Measurement of Psychophysiological Indicators

Irani, Ramin

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

COMPUTER VISION-BASED METHODS FOR DETECTION AND MEASUREMENT OF PSYCHOPHYSIOLOGICAL INDICATORS

BY
RAMIN IRANI

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

COMPUTER VISION-BASED METHODS FOR DETECTION AND MEASUREMENT OF PSYCHOPHYSIOLOGICAL INDICATORS

PH.D. DISSERTATION

by

Ramin Irani

Department of Architecture, Design and Media Technology
Aalborg University, Denmark

June 2017

Dissertation submitted: September 2017

PhD supervisor: Prof. Thomas B. Moeslund
Aalborg University, Denmark

PhD co-supervisor: Associate Prof. Kamal Nasrollahi
Aalborg University, Denmark

PhD committee: Associate Professor Lazaros Nalpantidis (Chairman)
Aalborg University, Denmark

Associate Professor Peter Ahrendt
Teknologisk Institut, Denmark

Assistant Professor Paulo Luís Serras Lobato Correia
Instituto Superior Técnico, Portugal

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-064-7

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Curriculum Vitæ

Ramin Irani received the BS degree in Electrical Engineering with emphasis on power system engineering from Azad University, Gonabad, Iran in 1996 and MS degree in Electrical Engineering with emphasis on Telecommunication from Azad University south Tehran, Iran in 2006. He also received his second MS degree from 'Blekinge Institute of Technology, Sweden' in Electrical Engineering with emphasis on Signal Processing in 2013. Ramin started his PhD study in Computer Vision at the Department of Architecture, Design and Media Technology, Aalborg University, Denmark, at April 2013.

His PhD study was aiming to developing computer vision methods for psychophysiological indications by focusing on the analysis of human facial videos. During the PhD study Ramin stayed 3 months at the Australian National University (ANU) as school visitor (Occupational Trainee) to conduct a research on stress recognition under supervision of Professor Tom Gedeon.

He has been involved in supervision of undergraduate students in topics of image processing and computer vision. His current research interests are visual analysis of people, biometrics, and machine learning.

Furthermore, Ramin has 8 years' experience in Tehran oil refining company. His main duties as electrical engineer during his career were supervising overhaul of electromotors and substations in different sections of the refinery, testing and measurement of resistance of earth, wells, tanks, and all refinery equipment's, troubleshooting of high voltage cables, using fault finding equipment, employee in charge of workmen for repairing and maintenance of high voltage and low voltage switches of electromotors.

ENGLISH SUMMARY

Recently, computer vision technologies have been used for analysis of human facial video in order to provide a remotely indicator of some crucial psychophysiological parameters such as fatigue, pain, stress and heartbeat rate. Available contact-based technologies are inconvenient for monitoring patients' physiological signals due to irritating skin and require huge amount of wires to collect and transmitting the signals. While contact-free computer vision techniques not only can be an easy and economical way to overcome this issue, they provide an automatic recognition of the patients' emotions like pain and stress. This thesis reports a series of works done on contact-free heartbeat estimation, muscle fatigue detection, pain recognition and stress recognition.

In measuring physiological parameters, two parameters are considered among many different physiological parameters: heartbeat rate and physical fatigue. Even though heartbeat rate estimation from video is available in the literature, this thesis proposes an improved method by using a new heartbeat footprint tracking approach from the face. The thesis also introduces a novel way of analyzing heartbeat traces from the facial video to provide visible heartbeat peaks in the signal. A method for physical fatigue time offset detection from facial video is also introduced.

One of the major contributions of the thesis, related to monitoring the patients, is recognizing level of pain and stress. The patients' pain must be continuously measured to evaluate treatment effectiveness. For objective measurement of the pain level, we proposed a new spatio-temporal technique based on energy changes of facial muscles due to pain. Obtained experimental results reveal that the proposed algorithms outperform state of the art algorithm [80]. Stress is another vital psychophysiological signal that is discussed in the last part of the thesis. The measurement of stress is important to assess conformability and health conditions of patients. Since the stress causes physiological changes and facial expression, we proposed a novel method based on thermal and RGB video data which is collected in Australian National University.

In addition, the thesis validates and tests a closed-loop tele-rehabilitation system based on functional electrical stimulation and computer vision analysis of facial expressions in stroke patients. Results from analysis of facial expressions show that present facial expression recognition systems are not reliable for recognizing patients' emotional states especially when they have difficulties with controlling their facial muscles.

Regarding future research, the authors believe that the approaches proposed in this thesis may be combined with other factors, such as vocal information and gesture that

usefully indicate patients' health status. Such a combination will provide a more reliable and accurate system for recognizing patients' emotional and physiological responses in the tele-rehabilitation process.

DANSK RESUME

Computer vision teknologier er for nyligt blevet anvendt til at analysere menneskelige ansigter ud fra videooptagelser for at tilvejebringe en ekstern indikator for en række afgørende psykofysiologiske parametre såsom træthed, smerte, stress og hjerterytme. Tilgængelige kontakt-baserede teknologier er uhensigtsmæssige til overvågning af patienters fysiologiske signaler pga. irriteret hud og mængden af ledninger, der er nødvendig for at indsamle og videregive signaler. Denne problematik kan løses nemt og økonomisk med kontakt-fri computer vision teknikker, og teknikkerne kan ydermere bidrage med en automatisk genkendelse af patienternes følelser i periodevise diagnoser såsom smerter og stress. Denne afhandling indeholder det samlede arbejde, der er udarbejdet om kontakt-fri påvisning af hjerterytme, muskeltræthed, smerte og stress.

Ved måling af fysiologiske parametre er to parametre udvalgt blandt adskillelige fysiologiske parametre: hjerteslagets rate og fysisk træthed. Selvom der i litteraturen forefindes metoder til estimering af hjerteslagets rate ud fra videooptagelser, foreslår denne afhandling en forbedret fremgangsmåde ved at anvende en ny tilgang, som inkluderer sporing af hjerteslagets fodaftryk i ansigtet. Afhandlingen introducerer også en ny metode til analyse af hjerteslagets spor i videooptagelser af ansigtet med det formål at fremskaffe synlige peaks af hjerteslaget i signalet. Ydermere introduceres en metode til estimering af den fysiske trætheds forskydning ud fra videooptagelser af ansigtet.

Et af de store bidrag fra afhandlingen relateret til overvågning af patienter er genkendelsen af smerte- og stressniveau. For at evaluere effektiviteten af behandlingerne skal patienternes smerteniveau måles kontinuerligt. For at opnå objektive målinger af smerteniveauet, har vi foreslået en ny spatio-temporal teknik baseret på energiændringer i ansigtets muskler grundet smerter. Forsøgsresultater påviser, at de foreslåede algoritmer overgår etablerede algoritmer. Stress er et andet vigtigt psykofysiologisk signal, som blev diskuteret i den sidste del. Måling af stress er essentielt for at opnå indsigt i patienternes tilpasning og helbredstilstand. Da stress forårsager fysiologiske ændringer på kroppen og ændrede ansigtsudtryk, foreslår vi en ny metode til estimering af stressniveauet baseret på videooptagelser fra termisk og RGB kamera. Data blev indsamlet på Australian National University.

Afhandlingen validerer og tester endvidere et closed-loop tele-rehabiliterings system baseret på funktionel elektrisk stimulation og computer vision til at analysere ansigtsudtryk hos patienter ramt af slagtilfælde. Resultater fra analysen af ansigtsudtryk indikerer, at nuværende systemer til genkendelse af ansigtsudtryk ikke er pålidelige til genkendelse af patienters følelser, især når patienterne har vanskeligt ved at kontrollere deres ansigtsmuskler.

Vedrørende fremtidig forskning, så er det forfatternes vurdering, at metoderne præsenteret i denne afhandling kan kombineres med andre faktorer, såsom stemme og gestusinformation, der kan bruges til at indikere patientens helbred. Sådan en kombination vil give et mere pålideligt og nøjagtigt system til at genkende patienternes emotionelle og fysiologiske reaktioner ved tele-rehabiliteringsprocessen.

TABLE OF CONTENTS

THESIS DETAILS

Thesis Title: Computer vision-based method for detection and measurement of psycho-physiological indicator

PhD Student: Ramin Irani

Supervisor: Prof. Thomas B. Moeslund, Aalborg University

Co-supervisor: Associate Prof. Kamal Nasrollahi, Aalborg University

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published (at the time of handing in the thesis) scientific paper which are listed below. Parts of the papers are used directly or indirectly in the extended summary of the thesis in the introduction. As part of the assessment, co-author statements to explicitly mentioning my contributions have been made available to the assessment committee and are also available at the Faculty.

The main body of this thesis consists of the following papers divided into three research themes presented in the thesis (the index number of the articles refers to the part and chapter of the thesis it is presented):

PART II: Estimation of physiological indicators

[1] Ramin Irani, Kamal Nasrollahi, Thomas B. Moeslund, "Improved Pulse Detection from Head Motions using DCT." in *9th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014, pp. 118-124.

[2] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Heartbeat Rate Measurement from Facial Video," *IEEE Intell. Syst.*, Dec. 2015.

[3] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Contactless Measurement of Muscles Fatigue by Tracking Facial Feature Points in A Video," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 1–5.

[4] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial Video based Detection of Physical Fatigue for Maximal Muscle Activity," *IET Comput. Vis.*, 2016.

[5] K. Nasrollahi, M. A. Haque, R. Irani, and T. B. Moeslund, “Contact-Free Heartbeat Signal for Human Identification and Forensics (submitted),” in *Biometrics in Forensic Sciences*, 2016, pp. 1–14.

PART III: Estimation of psychological indicatiors

[6] Ramin Irani, Kamal Nasrollahi, Thomas B. Moeslund, “Pain recognition using spatiotemporal oriented energy of facial muscles.” *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2015, pp. 80-87.

[7] Ramin Irani, Kamal Nasrollahi, Marc O. Simon, Ciprian A. Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H. Lundtoft, Thomas B. Moeslund, Tanja L. Pedersen, Maria-Louise Klitgaard, Laura Petrini, “Spatiotemporal analysis of RGB-D-T facial images for multimodal pain level recognition.” *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2015, pp. 88-95

[8] R. Irani, D. Simonsen, O. K. Andersen, K. Nasrollahi, and T. B. Moeslund, “Application of Automatic Energy-based Pain Recognition in Functional Electrical Stimulation,” *Internatinal J. Integr. Care*, vol. 15, no. 7, pp. 1–2, Oct. 2015.

[9] Ramin Irani, Kamal Nasrollahi, Thomas B. Moeslund, “Design of 4D Spatiotemporal Oriented Energy Filter for Kinect-based Pain Recognition (Technical Report).

[10] Simonsen, D., Irani, R., Nasrollahi, K., Hansen, J., Spaich, E., Moeslund, T., Andersen, O.S., “Validation and test of a closed-loop tele-rehabilitation system based on functional electrical stimulation and computer vision for analyzing facial expressions in stroke patients.” In: *Jensen, W., Andersen, O.K.S., Akay, M. (eds.) Replace, Repair, Restore, Relieve Bridging Clinical and Engineering Solutions in Neurorehabilitation SE - 103, Biosystems & Biorobotics*, vol. 7, pp. 741–750.

PART VI: Estimation of psychophysiological indicators

[11] Ramin Irani, Kamal Nasrollahi, Abhinav Dhall, Thomas B. Moeslund, and Tom Gedeon, “Thermal Super-Pixels for Bimodal Stress Recognition,” *sixth International Conference on Image Processing Theory*, 2016, pp. 1–14.

PREFACE

This thesis is submitted as a collection of papers in partial fulfillment of a PhD study in the area of Computer Vision at the Section of Media Technology, Aalborg University, Denmark. It is organized in four parts. The first part contains the framework of the thesis with a summary of the contributions. The rest of the parts contain layout revised articles published in different venues in connection to the research carried out during the PhD study.

The focus of this study is on analyzing human facial video to extract meaningful information for monitoring and recognizing some crucial health parameters e.g. heartbeat rate, fatigue, pain and stress. These parameters behave as indicators which imply status of patients' health and performance during rehabilitation exercise. The core contributions of the thesis are divided into three main topics: estimation of physiological, psychological and psychophysiological indicators. Ten articles and one technical report have been included in the thesis.

The work has been carried out from April 2013 to June 2017 as a part of the FTP financed project titled “Tele-rehabilitation after stroke Continued Functional Electrical Therapy (FET) in own home” The project aimed to have collaborations between health professionals, patients, private enterprises and research institutions. While writing this thesis I had collaboration with academicians from the other departments of Aalborg University, Denmark and Australian National University, Australia. I was employed as a PhD fellow with both research and teaching responsibilities during the time of PhD study.

I am grateful to my supervisors Prof. Thomas B. Moeslund and Associate professor Kamal Nasrollahi for support, encouragement and guidance through this research. The way of their help is always admirable for me and I never forget it. I also wish to thank my colleagues during my stay at Australian National University for warmly welcoming me and for their collaboration and support. My special thanks goes to all of my colleague in Visual Analysis of People laboratory (VAP), the staff and management of Aalborg University for providing the technical capacity and the state of the art facilities that enabled me to successfully complete my studies.

Have a nice reading!

PART I

OVERVIEW OF THE WORK

Chapter 1

Introduction

Ramin Irani, Kamal Nasrollahi, Thomas Moslund

Aalborg University

1.1. **Abstract**

Human facial video contains information regarding facial expressions, mental conditions, disease symptoms, and physiological parameters such as heartbeat rate, blood pressure, and respiratory rate. It also contains psychological signals (e.g. pain) which in some cases might be associated with significant physiological responses. A good example is stress which reveals both psychological and physiological responses on face in terms of facial expression and temperature changes. This dissertation focuses on how facial video analysis can be applied to these physiological and psychological signals known as psycho-physiological signs. This chapter presents a summary of the main themes and the results of research endeavors conducted during the doctoral degree program.

1.2. **Background**

The face is one of the most important parts of the body for nonverbal communication [1]. It reveals a person's age, identity, and emotions. The face even reveals physiological parameters, such as temperature, heartbeat rate and respiratory [2-6]. Therefore facial analysis leads to some crucial parameters and signs that can be utilized in many various applications like, patient diagnostics, Human Computer Interactions (HCI), security applications, physical and psychological therapy. In the last few decades, thanks to progress in the field of computer vision techniques, automatic face analysis systems have been widely studied and have received much attention from researchers in fields ranging from computer science to health care and psychology.

Automatic facial analysis by computer vision approaches consists of three main steps: face detection, facial feature extraction and classification (figure 1.1). The state-of-the-art algorithms in the context of the aforementioned applications for each of these blocks are reviewed in the following subsections:

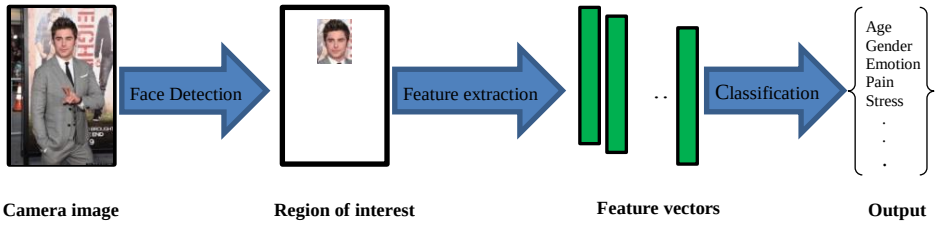


Figure 1.1: Pipeline of facial expression recognition. Source of the photo in the step Camera image: [7]

1.2.1. Face detection

In facial expression recognition, the process begins by detecting a face in a scene. Numerous techniques are developed to detect faces in an image. One of the most popular techniques is Viola and Jones algorithm [8], which is used for face detection in many facial expression recognition systems [9-13]. Besides, some researchers have used different methods, such as skin color-based face detection [14, 15]. Viola and Jones algorithm is based on Haar-like rectangular features. Computational efficiency is a benefit of this algorithm. In addition, features can be evaluated at any scale and location. This algorithm cannot handle rotated faces and faces of poor quality. Using skin color-based face detection for tracking faces has several advantages. It is highly robust to geometric variations of the face orientation, scale and occlusions. However, different colors of different faces and illumination variation can affect the performance of this algorithm.

To overcome the issues of the above traditional face detection, it is possible to determine face region directly by identifying geometrical structure of the faces, which is called face alignment. Face alignment aims to localize facial key points/ landmarks automatically and determines the shape of the face components such as eyes and nose (figure 1.2). It is essential to many facial analysis tasks e.g. expression recognition. Among the many different approaches for face alignment, supervised descent method (SDM) which solves a non-linear least square optimization problem [16] has emerged as one of the most popular state-of-the-art method. It is able to work in real time and provides a good and accurate estimate of facial landmarks. However, the accuracy suffers from the quality of the captured image sequences such as resolution, pose and brightness.

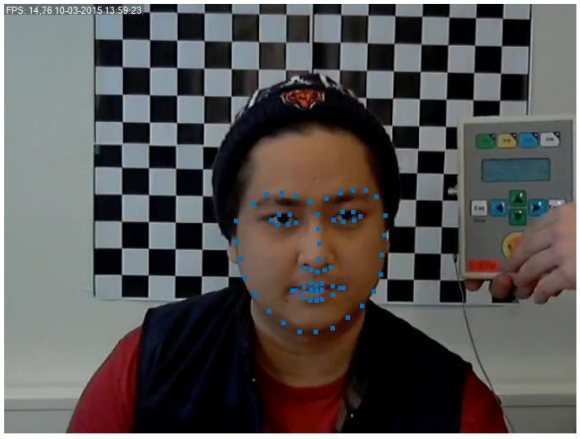


Figure 1.2: Face detection based on landmark localization

Finally, in the cases that available automatic face detection algorithm cannot be useful e.g. thermal images; we can apply template-based matching approach [17]. Template matching is a technique that finds the face region on an image or sequence of images (video frames) that has the highest correlation with the template. The template is provided manually which is the main drawback of this approach.

1.2.2. Facial feature extraction

Facial features are important to any classification process in facial analyzing systems. Utilizing inadequate features, can cause even the best classifier to accomplish an accurate recognition. The facial features are traditionally, classified into two major types [18]: Holistic-based representation (Appearance-based approach), Analytic-based representation (Geometric-based approach).

Geometric features present the shape and locations of facial components (like: mouth, eyes, brows, and nose) as well as the position of facial feature points (e.g. the corners of the eyes). This approach which is almost applied in all facial expression recognition relies on detecting sets of landmarks (fiducial points) e.g. [19 - 22] or connected face mesh e.g. [23, 24] in the first frame and then tracking them throughout the sequence [9]. Disadvantage of this approach is that it only considers the motion of a number of points. Therefore, much information in the skin texture is ignored. In contrast to geometric-based features, appearance-based features rely on deformation of skin texture such as wrinkles, bulges, and furrows and are good in providing global shape and texture.

Most of the present appearance-besed methods adopt Gabor-wavelet for recognizing facial expression [25-28]. Gabor filters are obtained by modulating a 2D sine wave with a Gaussian envelope. Zhang et al. [27] compared geometric-based features (the geometric positions of 34 fiducial points) and a set of multi-scale and multi-orientation Gabor filters coefficients at these points. Experimental results in [27] show that Gabor features describe facial deformation in details better than geometric positions. Tian [28] compared geometric-based features and Gabor-wavelet features with different image resolutions and her experiments show that Gabor-wavelet features work better for low resolution face images. There are also some literatures that applied the Gabor feature extractor for gender and age recognition [29-32]. The main drawback of this method is, convolving face images with a set of Gabor filters are computationally expensive and therefore it is inefficient in both time and memory due to the high redundancy of Gabor-wavelet features.

Local Binary Patterns (LBP) is another traditional well-known approach which is the most popular and successful approach in many facial analysis applications e.g. face recognition [33 - 35], facial expression analysis [36-39], demographic (gender, race, age, etc.) [40-41]. LPB is a powerful means of texture description, which labels the pixels of an image by thresholding a neighborhood of each pixel with the center value

and considering the results as a binary number. An advantage of LBP features is the simplicity of the LBP features. Comparing with large set of Gabor-wavelet coefficients, LBP allows very fast feature extraction without complex analysis in extracting. LBP features lie in a much lower dimensional space that reduces the memory space by an order of 17 [42], meaning that the LBP features are effective for facial analyzing systems, but the limitation of the LBP is that it cannot capture dominant features with large scale.

1.2.3. Classification

Many classification techniques have been applied to recognize facial expressions, including Support Vector Machines (SVM) [43-45], Neural Networks (NN) [27, 46], Bayesian Networks (BN) [47], k-Nearest Neighbor (kNN) [26, 48] and Hidden Markov Model (HMM) [9]. SVM is a powerful machine learning technique based on statistical learning theory that has been widely used for facial expression recognition. However, it is not proper for temporal modeling of facial expressions (dynamic texture).

In the case of temporal modeling, Hidden Markov Model (HMM) has proven to be a useful approach [9]. It is an effective tool which produces the output based on probability distribution over the sequences of input observations. The fact that the facial expressions have a unique temporal pattern has made HMM popular in the facial expression recognition community.

Over the last few years, researchers significantly advanced human facial analyzing with deep convolutional neural networks (CNNs). CNN is a kind of multilayer neural network, which has been designed for two dimensional data [49]. Fasel [50, 51] developed a system using CNN with receptive fields of different sizes. Then they applied it to face recognition and facial expression recognition. Osadchy et al. [52] presented a method based on CNN architecture that detects faces and estimates their pose in real time. In a very recent work, Levi et.al in [53] proposed an approach based on deep convolutional neural networks for both age and gender recognition and the test with newly audience benchmark in [54] proves that their proposed method outperforms the existing state of the art approaches.

1.3. Scope of the thesis

Computer vision-based facial analysis provides the ability to remotely and continuously monitor a patient's vital signs, including heartbeat and breathing rates, as well as recognize whether a patient feels pain or not. This contactless vision-based solution is cheaper than available contact-based measuring systems. It furthermore does not cause the irritation that results from skin's sensitivity to electrode connections. Tele-rehabilitation is one application example of automatic facial analysis systems. We know that in the rehabilitation process, visual information

provides important cues for a therapist supervising patients. Facial expressions due to pain or happiness accompanied with physiological parameters such as fatigue and/or breath rate are the most effective cues for recognizing patients' emotional states. Training at a distance without continuous supervision obviously makes it difficult for therapist to detect non-verbal social cues. During supervised rehabilitation the therapist adopts the exercise based on emotional feedback from the patients. For instance, when a patient feels pain, the therapist might stop the exercise or lower the intensity of the exercise. This individualized adaptation is hard to do when the exercise is performed at home without direct control of the therapist.

With regard to the above discussion, this field of study involves a vast variety of applications due to the diversity of information gathered from facial visual cues. In this thesis, the focus has been on three different topics within this field:

- Estimation of physiological indicators
- Estimation of psychological indicators
- Estimation of psycho-physiological indicators

Below we present these topics first in a general manner and then our concrete work and findings.

1.3.1. Estimation of physiological indicators

Measurement and monitoring of physiological parameters play an important role in different applications e.g. sport training, tele-rehabilitation and healthcare centers [55]. They indicate the state of patients' body function. Heartbeat rate, respiration rate, blood volume pulse and fatigue are some example of physiological parameters. Among the mentioned physiological signals, heartbeat rate is the most important one that provides information about the condition of cardiovascular system in applications like rehabilitation training program, and fitness assessment. For example, increasing or decreasing a patient's heartbeat rate beyond the norm in fitness assessment or rehabilitation training can indicate whether continuing the exercise is safe [56]. Traditional techniques such as pulse oximetry and electrocardiogram for measuring the heartbeat rate need the sensors to be connected to the patients' body (figure 1.3). This contact-based method may cause skin irritation and soreness. It also is uncomfortable especially, in the cases that sensors should be affix on the subjects' body during sleep and sport training. Installing huge amount of cables is another drawback of these systems.

Figure 1.3: An example of invasive connected-based sensors on patient's body that are not comfortable [57].

Recently, Ultra Wideband Radar, Microwave Doppler Radar and laser have been applied for contactless heartbeat rate measurement although all of these techniques require special and expensive hardware. Thereafter, computer vision-based sensors can be a solution. An interesting low cost and convenient method [58], which was proposed by Massachusetts Institute of Technology (MIT), measures the heartbeatrate by using a webcam. This study was driven with the fact that circulating the blood through vessels causes periodic subtle change to color skin. In this system (figure 1.4), ROI on the subject's face was automatically detected and tracked by a face tracker, and then like [59, 60] chromatic pixel values were split into RGB channels. Each channel separately averaged to gain raw RGB trace. All the traces then fed into an Independent Component Analysis (ICA) algorithm to recover three independent source signals from three color channels. For the sake of simplicity, authors always selected second component that typically includes a strong plethysmographic signal as desired source signal. Finally, hreatbeat rate (HR) and respiratory rate (RR) were obtained by filtering the selected component signal.

This system is effective, however in the case of head motion and noisy imaging conditions, the system cannot provide accurate results. To overcome this problem, Balakrishnan et al. proposed a motion-based contactless system for measuring HR [61]. Similar to color-based method; Balakrishnan's method was based on the fact that flow of blood through aorta due to pulsation of the heart muscles causes invisible motion on the head. In this approach, some feature points were automatically selected on ROI on the subject's facial video frame cheek by a method called Good Feature to Track (GFT). These feature points were tracked by a face tracker to generate some trajectories and then Principle Component Analysis (PCA) is applied to decompose trajectories into set of independent source signals.

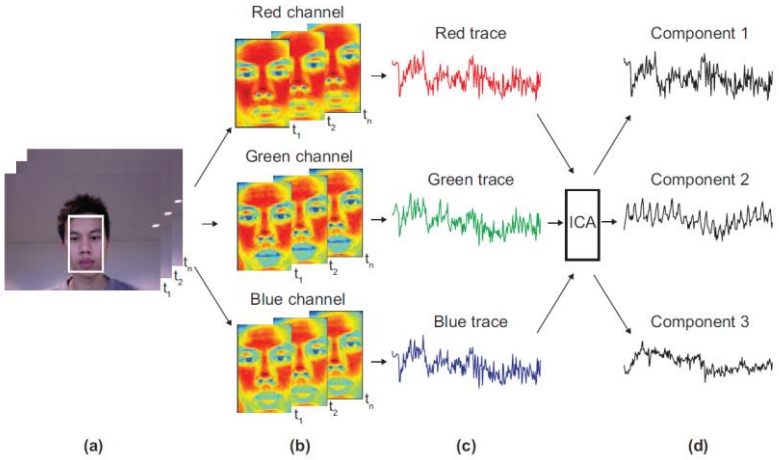


Figure 1.4: Cardiac pulse recovery methodology, a. The region of interest (ROI), b. The ROI is decomposed into the RGB channels c. ICA is applied on the normalized RGB traces to recover d. three independent source signals [58].

Selection of heartbeat signal was accomplished by using the percentage of total spectral power of the signal accounted for the frequency with the maximal power and its first harmonic. In contrast with [58], Balakrishnan's system was not only robust to noise sensitivity but also provide similar accuracy in results for grayscale video as well as color video. However, Balakrishnan's method was sensitive to facial expression change and head motion in video. Thus, we proposed an improvement of Balakrishnan's method in [56] by using the Discrete Cosine Transform (DCT) along with a moving average filter instead of the Fast Fourier Transform (FFT) of the previous method. This improved method (figure 1.5) provided better accuracy in HR measurement from video while having small expression and head motion changes [62]. Later on, we improved the results even further by combining the GFT feature points with facial landmarks extracted via supervised descent method (SDM). Combination of these two methods lets us to obtain stable trajectories that, in turn, allow for a better estimation of HR.

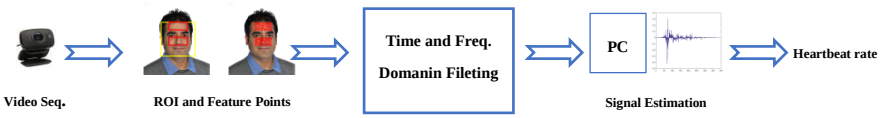


Figure 1.5: The block diagram of the improved system of one proposed in [56]. In this system, DCT algorithm applied to select the heartbeat rate component from output of PCA.

Another important physiological parameter is 'fatigue'. The term fatigue is usually used to describe overall feeling of tiredness or weakness. Fatigue may occur because of variety reasons. Fatigue can be mental or physical [63]. For instance, stress makes

people, mentally exhausted, but hard works or doing exercise for a long time makes people physically exhausted. Physical fatigue also known as muscle fatigue is a critical physiological indicator, in particular for athletes and therapists. Measuring the fatigue helps therapists to evaluate patients' progress. They monitor patients during exercise and make sure to keep the level of difficulty in a range that corresponding fatigue is not harmful to the patients.

Nowadays available technologies for measuring muscle fatigue are contact-based using devices, like, force gauge, EMG electrodes, or Mechanomyogram (MMG) sensors. Although measuring the fatigue by force gauge technique is simple. It requires devices such as hand grip dynamometer [64]. It hence, makes it impractical for some kind of exercises, for instance, those using dumbbells. EMG technique is largely used; however, complex implementation is the downside especially in case of automatic fatigue detection. Moreover, high sensitivity to noise adds more to its limitation. Furthermore, adhesive gel patches that are used along with the method might cause slight pain and skin irritation in some patients. MMG is another alternative method for non-invasive assessment of muscle fatigue, which is often used with EMG. Similar to EMG and other conventional fatigue detection techniques, such as accelerometer, goniometer and microphone [65], MMG sensors also require direct skin contact. Being expensive, bulky, and sensitive to noise are some restrictions of this method. Besides, they are not suitable for dynamic contraction. Our recent method proposed in [67, 68] relies on the notion that muscles start to shake as a result of tiredness triggered by an activity, and this shaking is reflected on the face. To the best of our knowledge the proposed method in this thesis is the only video-based non-invasive system for recognizing the muscle fatigue. The method is similar to the one introduced in [56], for detection of heartbeat rate from facial videos.

Part II of the thesis presents following papers [56,62,67,68 and 69]. The first chapter focuses on the estimation of the heartbeat rate using DCT transform. The results show that the proposed method is less sensitive to facial expression and muscle motion than the method proposed in [56]. The second chapter develops [61] with a combination of facial feature points proposed in [70, 71] and landmarks proposed in [72]. The estimated heartbeat rates determined by the proposed method are robust, after considering different light conditions and head positions. In chapter 3, an energy-based algorithm is proposed to indicate physical fatigue from maximal muscle activity and in chapter 4, similar to chapter 2, a combination of facial feature points and landmarks tracking is employed to improve the method of [67] in different light condition scenarios and head motions. The last chapter reviews the application of contact free measurement of heartbeat signals in human identification and forensic investigations. The outcome of the reviewed approach shows promising results for using HR as a soft biometric.

1.3.2. Estimation of psycological indicators

In social interaction, visual cues play a crucial role in exposing psychological information about the emotion and cognitive state [73]. One of this psychological information prominent in diagnostics and patient health care is pain. It is the most common reasons for seeking medical care with over 80% of patients complaining about some sorts of pain [74]. Pain is defined as "an unpleasant sensory and emotional experience associated with actual or potential damage or is described in terms of such damage" [75]. For example, when a person whacks his thumb with a hammer, sensory nerves around damaged cellules send the pain information to the brain. Perception of pain is formed by brain circuits and next the brain determines the emotion based on each painful experience after processing the pain information. Even though pain is produced by physical stimulus, the response of the brain is an emotional reaction. This emotion is often represented by changes in facial expression, and it makes patients susceptible to psychological consequences like anxiety and depression. Craig et al. in [76] evidenced that changes in facial appearance can be a very useful cue for recognizing the pain. Especially, this usefulness is highlighted in cases that patients are not able to make a verbal communication (e.g. children or patient after stroke).

Pain is one of the prime indicators in health assessment, and thus the ability of making a reliable evaluation of pain is of outmost importance for health related issues. The most common practice in pain assessment is to obtain information through direct communication with patients. Even though information can be readily accessed in this method, the reliability of information is undermined by some factors such as inconsistent metrics, reactivity to suggestions, efforts at impression management, and differences in pain conceptualization between physicians and sufferers. Furthermore, self-reporting becomes more complicated and inefficient when it comes to clients who are not capable of conducting a clear communication like children or patients with neurological impairment or breathing problems. It was stated in Atul Gawande's recent book [77] that patient's treatment is promoted as a result of the continuous monitoring of the pain level in certain intervals by medical staff. However, such an approach might be demanding, prone to mistakes, and stressful.

To address this difficulty, automatic pain recognition technology based on computer vision techniques for facial images was introduced, which has drawn great deal of attention over the recent years. Literature review shows that the number of articles with focus on automatic acquisition of pain level is limited. Studies in [78] are some examples within the area. In [79] a system capable of recognizing the level of pain intensity has been developed and introduced. It takes advantage of features like LBP, and utilizes different classes of classifiers such as PCA, SVM, and RVR to identify the pain level. The system generated interesting results; nevertheless, it has a main downside which is inability to read dynamics of the face. It has been observed during this thesis that pain is reflected on face through changes in some facial muscles and

the pattern of their motion. Such motions release energy whose level is directly related to the level of pain. This phenomenon underlies the basis of the present study. Attempts have been made to develop a system for pain recognition, which measures released energy level of facial muscles on a period of time.

To the best of our knowledge, the only system, which functions based on the similar principle, has been introduced by Hammal in [80]. It identifies four levels of pain intensity by using a combination of AAM and an energy based filter, long normal. Although the system employs the released energy concept, it performs on a frame-by-frame basis (video sequence). Our proposed system in [81] is empowered to capture released energy of the facial muscles in spatial as well as temporal domain. In order to achieve so, a special type of spatio-temporal filter is used. Such a filter proved to be successful in other applications like region tracking for extraction of information simultaneously in both spatial and temporal domains. The block diagram of this system illustrates in figure 1.6.

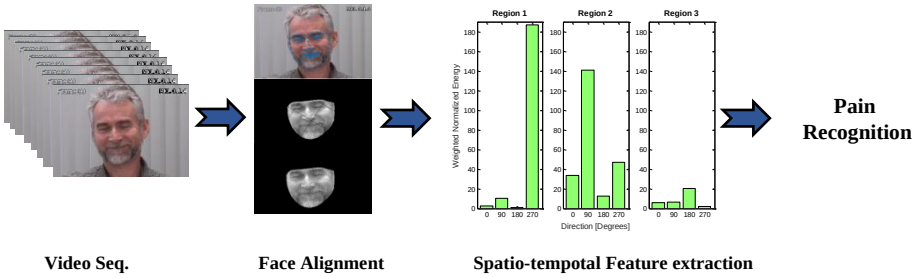


Figure 1.6: Block diagram of pain recognition, based on energy-based spatio-temporal feature extraction.

In [81], first faces are detected in each input video sequence. Thereafter, detected faces are aligned with a predetermined framework by active appearance model (AAM) using the provided landmarks, (The landmarks are included in the employed database). Registration to the framework results in disappearance of some parts of the face. This effect might be seen as formation of holes or lines in the registered face. To handle the problem, an inpanting algorithm was applied. Ultimately, the released energy of motion of facial muscles in aligned faces is detected and identified by 3D spatiotemporal filtering applied in x, y, and t dimensions.

The discussed approach was implemented in Chapter 7, in which we applied this method to the UNBC-MacMaster Shoulder Pain Expression Archive Database [82] for evaluating our model. Chapter 8 extends the approach proposed in Chapter 7 to multimodal dynamic pain recognition. Released energy is extracted from three RGB, depth and thermal inputs. In Chapter 9 we examine and validate the dynamic pain approach for automatic detection of pain due to electrical stimulation. It can be help-

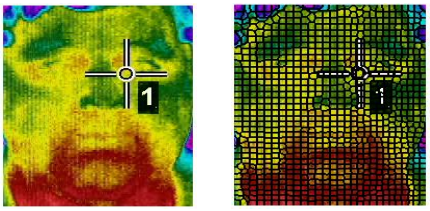
ful in tele-rehabilitation systems to adjust the intensity of electrical stimuli when patients feel painful. Chapter 10 is a technical report describing a novel 4D steerable and spatiotemporal filter. Application of proposed filter is in recognizing the pain from multimodal inputs proposed on Chapter 8. This filter can extract the features without requiring a 3D spatiotemporal filter to be applied individually for each modal. The last chapter is a joint work between Aalborg University's Health Department and us. The paper described the validation and test of a Microsoft Kinect based tele-rehabilitation system incorporating closed-loop controlled functional electrical stimulation for assisting training of hand function in stroke patients. Stroke patients often suffer from deficits in movement/motor control. They may regain motor control through intensive rehabilitation training, but a significant amount of their time is spent on self-training without therapeutic supervision. One way to ensure the quality of unsupervised self-training is to make use of a tele-rehabilitation system. In addition, we analyzed patients' facial expression during this work.

1.3.3. Estimation of psychophisiological indicators

Study in field of "psychophysiology" has become more popular among scientists. Many believe that psychological issues cause physiological symptoms and any changes on physiological components cause a psychological one [83]. In other word, we can define psychophysiology as an interaction between emotion and body function [84]. Stress that is a major problem of people in modern society is a good example of a psychophysiological response. For example, when we want to perform a task within a given period, while we do not have enough time, a set of physiological reactions, like, heartbeat and respiration rates increase, that indicate a stressful situation [85]. These physiological reactions are accompanied by some emotions that can appear as fear, anxiety and disgust on patients' face.

Traditional stress recognition systems are based on self-report and/or physiological changes measurement. These systems, which use invasive sensors, are not able to monitor the patients instantaneously and continuously. Utilizing contactless sensors such as RGB and thermal cameras can be a solution to overcome this problem. Considering the fact that stress is associated with emotional responses, it gives rise to some changes on facial appearace (facial expression) which have been used as a clue in stress detection by some researchers [86-88]. Yet, researchers tend to make use of physiological rather than emotional responses due to some uncertainties raised by using facial expression as the source of information [89-93]. Recently, contactless assessment of physiological signals has been possible by imaging techniques such as RGB Video Recorder and Thermal Imaging. Vision based systems usually employ either RGB or Thermal Imaging for stress recognition. In order to employ the opportunities of fusing the both techniques, recent literature [94] proposed a computational model which takes information from the both modalities and uses a descriptor called Histogram of Dynamic Thermal Patterns (HDTP) for processing. However, the accuracy of 65% could be achieved by such method. Subsequent integration of Genetic

Algorithm (GA) – Support Vector Machine (SVM) classifier improved the accuracy to 85% [94]. In our recent work [95] the attempt has been made to enhance the accuracy further by representing thermal images as a group of super-pixels. Super-pixel is defined as a group of pixels with similar characteristics and special information. In thermal images (figure 1.7.a) superpixels are a group of pixels with similar color (figure 1.7.b) which uniquely assigned to the pixels with nearly same temperature. This method is able to simultaneously group highly correlated pixels and accelerate processing time.



a b

Figure 1.7: a. Typical facial region and b. its corresponding super-pixels [95]

The block diagram of the proposed system is shown in figure 1.8. The test subjects are filmed by a RGB camera that is synchronized with a thermal camera in parallel. These two types of video streams go through three different steps: 1) Face region detection and quality assessment, 2) Feature extraction, and 3) Classification and fusion. For RBG images, Viola & Jones technique [96] was used for face detection then face regions with less correlation are removed using a face quality assessment algorithm. Finally, Local Binary Patterns (LBP) [97] is extracted from the remaining facial regions and is used as feature points. However, for detecting the face area in the thermal images, we use a template matcher for face region detection [17]. Then, we compute the linear spectral clustering super-pixel algorithm (LCS) [99], instead of directly computing a facial descriptor. Further, the mean values of the generated super-pixels are used as the facial features. Having extracted the facial features from two types of inputs, we use a support vector machine (SVM) classifier for producing classification scores for each type of input. These scores are finally fused at decision level to recognize the stress [95].

In the part IV, we proposed an automatic bimodal stress recognition system from facial images. These images captured in two different modalities by using RGB and thermal cameras. As we discussed above, we recognize the stress with considering both facial expression (as psychological cue) and thermal array (as physiological cue).

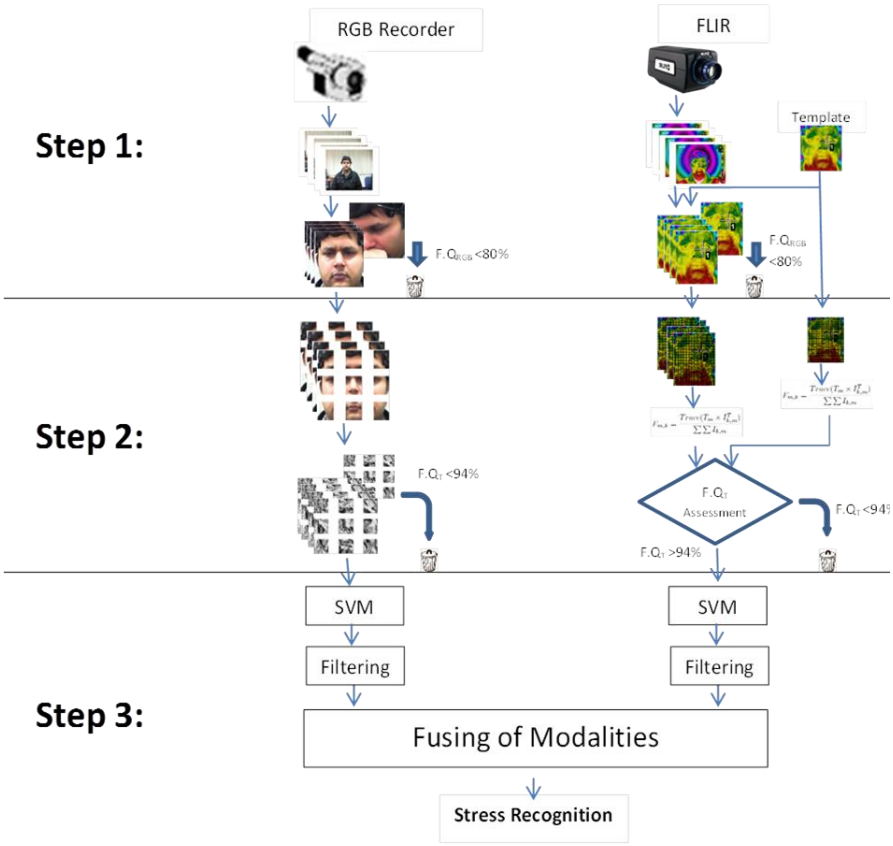


Figure 1.8. The block diagram of the biomodal stress recognition proposed in [88].

1.4. **Summary of the Contributions**

This study contains a collection of publications which cover three topics concerning the field of computer vision. Contributions to each topic are briefly, discussed below and Table 1.1 summarizes the main methods, devices and applications used in this study.

Topic 1. Estimation of physiological indicatiors:

- **Estimation of heartbeat signal from facial video and its application in forensics:** Although researchers at the Massachusetts Institute of Technology have already created estimates of heartbeat rates using facial video information, chapter two proposed an approach using the DCT algorithm to improve accuracy in the facial expression and head motion conditions. This

method is then developed in chapter three utilizing a facial expression log and SDM method. The applied algorithms' result is more stable and accurate than the proposed method in chapter 2 in different light conditions and head motions. The contact-free heartbeat rate algorithm for biometric recognition is explored in Chapter 6.

- **Estimation of physical fatigue from facial video:** To the best of our knowledge, we, in this thesis, have proposed the first system for contact-less measurement of fatigue from facial videos. In Chapter 4, we estimate the physical fatigue by measuring the energy released from the shaking of the face when fatigue is incurred. In Chapter 5, we improve upon the results obtained in Chapter 4 by applying the facial expression log and SDM algorithm used in Chapter 3.

Topic 2. Estimation of psychological indicators:

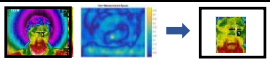
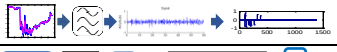
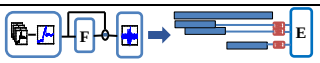
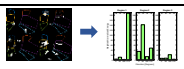
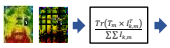
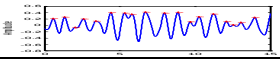
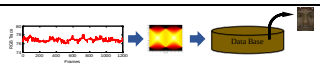
- **Detecting the pain based on the energy released due to facial muscle motion:** The papers published in the field of pain recognition are mostly based on geometric and appearance features similar to facial expression recognition. Chapter 7 proposes a novel spatio-temporal energy-based algorithm, which recognizes the pain based on released energy due to muscle motion. Chapter 8 develops this concept further for a multimodal database, which consists of RGB, Depth and Thermal videos. In chapter 9, we applied the algorithme in chapter 7 to estimate the subjects' pain level when they stimulate with electrical pulses. Chapter 10 presents a design of a 4D energy-based spatio-temporal filter which can measure facial muscles motion directly.

- **Computer vision analyzing facial expression in stroke patients in a test of functional electrical stimulation in Tele-rehabilitation system:** Our contribution in Chapter 11 is, analyzing accuracy of applying the state of the art facial expression algorithms on stroke patients in tele-rehabilitation training.

Topic 3. Estimation of psychophysiological indicators:

- **Biomodal stress recognition:** Chapter 12 proposes a bio-modal algorithm using super pixel for stress recognition. Since the stress effects on both physical and physiological changes on the body, this method fuses results obtained from RGB (physical muscle motion on the face) and Thermal (physiological parameters) frames to provide a system that outperform the state of the art work of [94].

Table 1.1: Contents of the thesis in relation to the challenges associated with of facial video analysis. Sources of some photos: [99, 100]

Stage	Contents	Figurative illustration	Chapters
Input data	*RGB Camera*		*2-7,9, 11-12*
	Thermal Camera		*8 and 12*
	Kinect ver. 2		*8 and 10*
Face region of interest	*Haar like features*		*2-6, and 12*
	Facial landmarks		*3,5,7-10*
	Template Matcher		*12*
	Face quality assessment		*3,5 and 12*
Feature processing and classifications	*Good Feature Tracking*		*2-6*
	Filtering and decomposition		*2,3 and 6*
	Energy estimation		*4 and 5*
	Energy-based Spatio-temporal filter		*7-10*
	Local Binary Pattern recognition		*12*
	Super pixel and statistical processing		*12*
	Making decision algoritm and Sigmoid function		*2-9, 12*
Applications	*Heartbeat rate estimation*		*2,3 and 6*
	Physical fatique indicator		*4 and 5*
	Identification recognition		*6*
	Pain level indicatior		*7-9 and 10*
	Stress recognition		*12*

1.5. Conclusions

In the pursuance of a doctoral degree from April 2013 – June 2017, I worked in the broad area of computer vision and contributed in major applications that are most useful in tele-rehabilitation and therapy. This research work tries to indicate some vital psychophysiological signs utilizing contactless facial feature measurement. These signs play an important role in controlling the tele-rehabilitation process automatically, based on patients' feedback to the exercise and environmental stimuli. In general, this feedback should be a psychophysiological response. However, in some cases like feeling pain the emotional response revealed in facial expression is more indicative than physiological one. Conversely, in some other situations the opposite might be true. For example when patients become tiered or have a disease, their physiological signs such as heartbeat rate, temperature and fatigue may be more notable that their facial expressions. In this case, the psychological behavior might provide less or unreliable information for therapists. Likewise, there are some other feedbacks such as stress for which both emotional and physiological behaviors can be informative and useful. In such cases, a bio-modal approach can be employed. Accordingly, the thesis is categorized in three main parts with respect to evaluation of physiological, psychological and psychophysiological parameters, which, involves estimation of pain, stress, muscle fatigue and heartbeat rate.

This thesis comprises the stories of facial video by addressing the questions such as 1) how to read and assess the pain from facial muscle motion, 2) how to extract heartbeat signal 3) how to detect physical fatigue, and 4) how to recognize stress. All the contributions together opened up a broad horizon for future research works.

Even though patients' face expresses crucial information about patients' health status, it is not the only clue. Vocal information, gesture along with body action are also useful factors for health status. As future work, authors believe that proposing a multimodal system in combination with proposed approaches in this thesis will enhance the reliability and accuracy of automatic monitoring of patient's status and rehabilitation.

1.6. References

[1] A. Mehrabian, "Communication without Words," Psychology Today, Vo.1.2, No.4, pp 53-56, 1968.

[2] A. Seal, S. Ganguly, D. Bhattacharjee, M. Nasipuri, and D. K. Basu, "Thermal Human Face Recognition Based on Haar Wavelet Transform and Series Matching Technique," Lecture Notes in Electrical Engineering Multimedia Processing, Communication and Computing Applications, pp. 155–167, 2013.

[3] H. Rahman, M. U. Ahmed and S. Begum. "Non-contact Physiological Parameters Extraction Using Camera." SpringerLink. Springer, Cham, 27 Oct. 2015.

[4] A. Procházka, M. Schätz, O. Vyšata, and M. Vališ. "Microsoft Kinect Visual and Depth Sensors for Breathing and Heart Rate Analysis," Sensors (Basel, Switzerland).U.S. National Library of Medicine, 28 June 2016.

[5] G. F. Lewis, R. G. Gatto, and S. W. Porges. "A Novel Method for Extracting Respiration Rate and Relative Tidal Volume from Infrared Thermography," Psychophysiology. U.S. National Library of Medicine, July 2011.

[6] J. Chen et al., "RealSense = real heart rate: Illumination invariant heart rate estimation from videos," *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Oulu, 2016, pp. 1-6.

[7] "Neighbors' LA Premiere - HQ/Favs, " Ehs-wildcats.livejournal.com, 2017. [Online]. Available: http://ehs-wildcats.livejournal.com/665196.html?thread =13722988. [Accessed: 07- April- 2017].

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, pp. I-511-I-518 vol.1.

[9] S. Koelstra, M. Pantic and I. Patras, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940-1954, Nov. 2010.

[10] C. A. Corneanu, M. O. Simón, J. F. Cohn and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548-1568, Aug. 1 2016.

[11] S. K. Kamarol, Amalina, M. H. Jaward, J. Parkkinen, and R. Parthiban. "Spatiotemporal Feature Extraction for Facial Expression Recognition," IET Image Processing. IET Digital Library, 01 July 2016.

[12] S. Agrawal and P. Khatri, "Facial Expression Detection Techniques: Based on Viola and Jones Algorithm and Principal Component Analysis," *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, Haryana, 2015, pp. 108-112.

[13] M. Owayjan, R. Achkar and M. Iskandar, "Face Detection with Expression Recognition using Artificial Neural Networks," *2016 3rd Middle East Conference on Biomedical Engineering (MECBME)*, Beirut, 2016, pp. 115-119.

[14] H. G. Hosseini and Z. Krechowec, "Facial expression analysis for estimating patient's emotional states in RPMS," *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, 2004, pp. 1517-1520.

[15] M. N. Bin Mansor, S. Yaacob, R. Nagarajan and M. Hariharan, "Patient monitoring in ICU under unstructured lighting condition," *2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA)*, Penang, 2010, pp. 608-611.

[16] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 532-539.

[17] Template matching using correlation coefficients. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/28590-template-matching-using-correlation-coefficients.

[18] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424-1445, Dec 2000.

[19] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433-449, April 2006.

[20] M. Pantic and L. J. M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," in *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1449-1461, June 2004.

[21] M. F. Valsta and M. Pantic. "Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics," SpringerLink. Springer, Berlin, Heidelberg, 20 Oct. 2007.

[22] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," in *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172-187, Jan. 2007.

[23] C. Hu, Y. Chang, R. Feris and M. Turk, "Manifold Based Analysis of Facial Expression," *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 81-81.

[24] S. Koelstra, M. Pantic and I. Patras, "A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940-1954, Nov. 2010.

[25] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169-1179, Jul 1988.

[26] M. J. Lyons, J. Budynek and S. Akamatsu, "Automatic classification of single facial images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357-1362, Dec 1999.

[27] Z. Zhang, M. Lyons, M. Schuster and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, 1998, pp. 454-459.

[28] Y. Tian, "Evaluation of Face Resolution for Expression Analysis," *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 82-82.

[29] J. Kim, D. Han, S. Sohn and J. Kim, "Facial age estimation via extended curvature Gabor filter," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 1165-1169.

[30] H. Lin, H. Lu and L. Zhang, "A New Automatic Recognition System of Gender, Age and Ethnicity," *6th World Congress on Intelligent Control and Automation*, Dalian, 2006, pp. 9988-9991.

[31] F. Gao and H. Ai, "Face Age Classification on Consumer Images with Gabor Feature and Fuzzy LDA Method." SpringerLink. Springer, Berlin, Heidelberg, June 2009.

[32] H. Ren, Z.-N. Li, "Gender Recognition Using Complexity-Aware Local Features," 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Aug. 2014, pp. 2389-2394,.

[33] T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[34] C. Chan, J. Kittler, and K. Messer, "Multi-scale local binary pattern histograms for face recognition," in Proc. Int. Conf. Biometrics (ICB), 2007, pp. 809-818.

[35] W. Zhang, S. Shan, H. Zhang, W. Gao, and X. Chen, "Multi-resolution Histograms of Local Variation Patterns (MHLVP) for robust face recognition," in Proc. Audio and Video-based Biometric Person Authentication (AVBPA), 2005, pp. 937-944.

[36] S. M. Lajevardi and Z. M. Hussain, "Local feature extraction methods for facial expression recognition," *17th European Signal Processing Conference*, Glasgow, 2009, pp. 60-64.

[37] S. Liao, W. Fan, A. C. S. Chung, and D. Y. Yeung, "Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2006, pp. 665 – 668.

[38] G. Zhao and M. Pietikäinen, "Experiments with facial expression recognition using spatiotemporal local binary patterns," in Proc. Int. Conf. Multimedia and Expo (ICME), 2007, pp. 1091-1094.

[39] Wei-Lun Chao, Jian-Jiun Ding, Jun-Zuo Liu , “Facial expressionrecognition based on improved local binary pattern and class-regularizedlocality preserving projection” in Signal Processing , vol. 117, pp. 1-10,2015.

[40] N. Sun, W. Zheng, C. Sun, C. Zou, and L, Zhao, “Gender classification based on boosting local binary pattern,” in Proc. Int. Symposium on Neural Networks (ISNN), 2006, pp. II: 194-201.

[41] Z. Yang and H. Ai, “Demographic classification with local binary patterns,” in Proc. Int. Conf. Biometrics (ICB), 2007, pp. 464-473.

[42] C. Shan, S. Gong and P. McOwan, ”Facial expression recognition based on statistical local features,” Universal short title catalogue book, Chapter 4, 2008.

[43] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image," *Face and Gesture 2011*, Santa Barbara, CA, 2011, pp. 866-871.

[44] T. Guha and R. K. Ward, "Learning Sparse Representations for Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576-1588, Aug. 2012.

[45] A. Ramirez Rivera, J. Rojas Castillo and O. Oksam Chae, "Local Directional Number Pattern for Face Analysis: Face and Expression Recognition," in *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1740-1752, May 2013.

[46] P. V. Saudagare and D. S. Chaudhari, “Facial Expression Recognition using Neural Network An Overview,” International Journal of Soft Computing and Engineering (IJSCE), vol. 2, Iss. 1, pp.: 224-227, March 2012.

[47] H. Dibeklioglu, A. A. Salah and T. Gevers, "A Statistical Method for 2-D Facial Landmarking," in *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 844-858, Feb. 2012.

[48] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman and T. J. Sejnowski, "Classifying facial actions," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct 1999.

[49] B. Zhang, C. Quan and F. Ren, "Study on CNN in the recognition of emotion in audio and images," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, 2016, pp. 1-5.

[50] B. Fasel. Multiscale facial expression recognition using convolutional neural-networks. In Proceedings of the Third Indian Conference on Computer Vision, Graphics and Image Processing, Ahmedabad, India, 2002.

[51] B. Fasel. Robust face analysis using convolutional neural networks. In Proceedings of the 16th International Conference on Pattern Recognition, volume 2, pages 40–43, Quebec, Canada, 2002.

[52] R. Osadchy, M. Miller, and Y. LeCun, “Synergistic Face Detection and Pose Estimation with Energy-Based Model,” Proc. Advances in Neural Information Processing Systems, pp. 1017-1024, 2004.

[53] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston, MA, 2015, pp. 34-42.

[54] E. Eidinger, R. Enbar and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," in *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170-2179, Dec. 2014.

[55] R. Irani, K. Nasrollahi, and T. B. Moeslund, “Contactless Measurement of Muscles Fatigue by Tracking Facial Feature Points in A Video,” in IEEE International Conference on Image Processing (ICIP), 2014, pp. 1–5.

[56] Ramin Irani, Kamal Nasrollahi, Thomas B. Moeslund, “Improved Pulse Detection from Head Motions using DCT.” in 9th International Conference on Computer Vision Theory and Applications (VISAPP), 2014, pp. 118-124.

[57] Pediatric Pulmonology and Sleep Medicine of Florida LLC - Procedures.(n.d.). Retrieved from http://swflkidlung.com/sleep_medicine_procedures

[58] M. Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” Opt. Expr., vol. 18, pp. 10762–10774, May 2010.

[59] C. Takano and Y. Ohta, “Heart rate measurement based on a time-lapse image,” Med. Eng. Phys., vol. 29, no. 8, pp. 853–857, Oct. 2007.

[60] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light,” Opt. Express, vol. 16, no. 26, pp. 21434–21445, Dec. 2008.

[61] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting Pulse from Head Motions in Video,” in Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2013, pp. 3430–3437.

[62] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, “Heartbeat Rate Measurement from Facial Video,” IEEE Intell. Syst., Dec. 2015.

[63] A. J. Dittner, S. C. Wessely, and R. G. Brown, “The assessment of fatigue: a practical guide for clinicians and researchers,” J. Psychosom. Res., vol. 56, no. 2, pp. 157–170, Feb. 2004.

[64] W. D. McArdle and F. I. Katch, Essential Exercise Physiology 4th, 4th revised international ed edition. Philadelphia: Lippincott Williams and Wilkins, 2010.

[65] C. Orizio, M. Gobbo, B. Diemont, F. Esposito, and A. Veicsteinas, “The surface mechanomyogram as a tool to describe the influence of fatigue on biceps brachii

motor unit activation strategy. Historical basis and novel evidence," Eur. J. Appl. Physiol., vol. 90, no. 3–4, pp. 326–336, Oct. 2003.

[66] M. B. I. Raez, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications," Biol. Proced. Online, vol. 8, pp. 11–35, Mar. 2006.

[67] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Contactless Measurement of Muscles Fatigue by Tracking Facial Feature Points in A Video," in IEEE International Conference on Image Processing (ICIP), 2014, pp. 1–5.

[68] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial Video based Detection of Physical Fatigue for Maximal Muscle Activity (accpeted)," IET Comput. Vis., 2016.

[69] K. Nasrollahi, M. A. Haque, R. Irani, and T. B. Moeslund, "Contact-Free Heartbeat Signal for Human Identification and Forensics (submitted)," in Biometrics in Forensic Sciences, 2016, pp. 1–14.

[70] J. Shi and C. Tomasi, "Good features to track," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994, pp. 593–600.

[71] J. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," Intel Corp. Microprocess. Res. Labs, 2000.

[72] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532–539.

[73] E.-M. Seidel, U. Habel, M. Kirschner, R. C. Gur, and B. Derntl, "The impact of facial emotional expressions on behavioral tendencies in females and males," Journal of experimental psychology. Human perception and performance, Vol 36, no. 2, pp. 500-507.Apr-2010.

[74] M. S. Walid, S. N. Donahue, D. M. Darmohray, H. J. L. A., and R. J. J. Sam, "The Fifth Vital Sign—What Does It Mean?," Pain Practice, Vol 8, no. 6, pp. 417-422-Jul-2008.

[75] "IASP Taxonomy - IASP," IASP Taxonomy - IASP. [Online]. Available: http://www.iasp-pain.org/Taxonomy. [Accessed: 05-Apr-2017].

[76] K. Craig, K. Prkachin, and R. Grunau. The facial expression of pain. Handbook of pain assessment, Guilford, New York, 2001.

[77] A. Gawande. The checklist manifesto: How to get things right? Metropolitan Books, New York, 2010

[78] P. Jinye;Y. Ruijing;F. Xiaoyi;W. Wenxing;P. Xianlin, "Survey on Facial Expression Recognition of Pain, " Journal of Data Acquisition and Processing,Jan.

2016, [Online]. Available: http://en.cnki.com.cn/Article_en/CJFDTotal-SJCJ201601004.htm. [Accessed: 05-Apr-2017].

[79] S. Kaltwang, O. Rudovic, and M. Pantic, “Continuous pain intensity estimation from facial expressions,” in International Symposium on Advances in Visual Computing, 2012, pp. 368–377.

[80] Z. Hammal and J. Cohn. “Automatic detection of pain intensity,” In Proceedings of the 14th ACM international conferenceon Multimodal interaction, 2012, pp. 22–26.

[81] R. Irani, K. Nasrollahi and T. B. Moeslund, “Pain recognition using spatiotemporal oriented energy of facial muscles,” IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 80-87.

[82] P. Lucey, J. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, “Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database,” Image and Vision Computing, vol. 30, no. 3, pp. 197–205, 2012.

[83] G. Stuart, *principles and practice of psychiatric nursing*, 10th ed. St. Louis: Mosby, 2014.

[84] C. L. Bethel, K. Salomon, R. R. Murphy and J. L. Burke, "Survey of Psychophysiology Measurements Applied to Human-Robot Interaction," RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, 2007, pp. 732-737.

[85] S. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. Schramek, “The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition,” Brain and Cognition, vol. 65, no. 3, pp. 209 – 237, 2007.

[86] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, “A real-time human stress monitoring system using dynamic bayesian network,” in Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on., 2005, pp. 70–70.

[87] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, “Optical computer recognition of facial expressions associated with stress induced by performance demands,” Aviation, space, and environmental medicine, vol. 76, no. Supplement 1, pp. B172–B182, 2005.

[88] H. Gao, A. Yuce, and J.-P. Thiran, “Detecting emotional stress from facial expressions for driving safety,” in Image Processing (ICIP), IEEE International Conference on., 2014, pp. 5961–5965.

[89] F. Bousefsaf, C. Maaoui, and A. Pruski, “Remote assessment of the heart rate variability to detect mental stress,” in Pervasive Computing Technologies for

Healthcare (PervasiveHealth), 2013 7th International Conference on., 2013, pp. 348–351.

[90] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She, “Detection of psychological stress using a hyperspectral imaging technique,” Affective Computing, IEEE Transactions on, vol. 5, no. 4, pp. 391–405, 2014.

[91] I. Pavlidis and J. Levine, “Thermal image analysis for polygraph testing,” Engineering in Medicine and Biology Magazine, IEEE, vol. 21, no. 6,pp. 56–64, 2002.

[92] I. Pavlidis, N. L. Eberhardt, and J. A. Levine, “Human behaviour: Seeing through the face of deception,” Nature, vol. 415, no. 6867, pp. 35–35, 2002.

[93] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, “Perinasal imaging of physiological stress and its affective potential,” Affective Computing, IEEE Transactions on, vol. 3, no. 3, pp. 366–378, 2012.

[94] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, “Thermal spatiotemporal data for stress recognition,” EURASIP Journal on Image and Video Processing, vol. 2014, no. 1, 2014.

[95] R. Irani, K. Nasrollahi, A. Dhall, T. B. Moeslund and T. Gedeon, "Thermal super-pixels for bimodal stress recognition," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, 2016, pp. 1-6.

[96] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-511-I-518.

[97] M. Pietik¨ainen, A. Hadid, G. Zhao, and T. Ahonen, “Local binary patterns for still images,” in Computer Vision Using Local Binary Patterns. Springer, 2011, pp. 13–47.

[98] Z. Li and J. Chen, "Superpixel segmentation using Linear Spectral Clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1356-1363.

[100] T. Kanade, J. F. Cohn and Yingli Tian, "Comprehensive database for facial expression analysis," Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, 2000, pp. 46-53.

[101] "5 Quotes To Help You Destress", Odyssey, 2016. [Online]. Available: https://www.theodysseyonline.com/5-quotes-to-help-you-de-stress. [Accessed: 7- Apr.- 2017].

PART II

ESTIMATION OF PHYSIOLOGICAL INDICATORS

Chapter 2

Improved Pulse Detection from Head Motions Using DCT

Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund

This paper has been published in

9th International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 2014, pp. 118-124

The layout has been revised.

2.1. **Abstract**

The heart pulsation sends out the blood throughout the body. The rate in which the heart performs this vital task, heartbeat rate, is of curial importance to the body. Therefore, measuring heartbeat rate, a.k.a. pulse detection, is very important in many applications, especially the medical ones. To measure it, physicians traditionally, either sense the pulsations of some blood vessels or install some sensors on the body. In either case, there is a need for a physical contact between the sensor and the body to obtain the heartbeat rate. This might not be always feasible, for example, for applications like remote patient monitoring. In such cases, contactless sensors, mostly based on computer vision techniques, are emerging as interesting alternatives. This paper proposes such a system, in which the heartbeats (pulses) are detected by subtle motions that appear on the face due to blood circulation. The proposed system has been tested in different facial expressions. The experimental results show that the proposed system is correct and robust and outperforms state-of-the-art.

2.2. **Introduction**

Heartbeat rate is obviously a vital sign of human body's activity and its measurement is of great importance in many applications, for instance, fitness assessment, training programs and medical diagnosis. For example, in fitness assessment during the exercise, heartbeat rate is used as a crucial sign that helps to assess the condition of cardiovascular system. Here it can be used also for ensuring the safety of the process. If the heartbeat rate goes beyond the normal range, continuing the exercise is not safe any longer.

Heartbeat rate is usually measured by devices that take samples of heartbeats and compute the beats per minute (bpm). Currently, one of the popular non-invasive and standard devices for measuring the heartbeat rate is electrocardiogram (ECG). They are very accurate, but expensive. These devices are electrode-based and therefore require wearing adhesive gel patches or chest straps that may cause skin irritation and slight pain. Commercial pulse oximetry sensor is another technique that is placed on specific parts of body like fingertips or earlobe.

Though the above mentioned devices are accurate, they are inconvenient as they need to have physical contact with patient's body. Therefore, developing contactless methods, which are based on the patient's physiological signals, have recently been considered as an interesting alternative for measuring heartbeat rate. This technology would also decrease the amount of cabling and clutter related to Intensive Care Unit (ICU) monitoring, long-term epilepsy monitoring, sleep studies, and any continues heartbeat rate measurement (Poh, 2010). These contactless methods that are usually based on computer vision techniques can be divided into two groups. In the first group, known as photoplethysmography (PPG) methods, usually a red, or an infrared light is transmitted on the patients (face or body) and the reflected light is sensed by

the system. The variations in the transmitted and the reflected lights are then used to measure heartbeat rate. Besides using dedicated light sources, the main drawbacks of PPG systems are that they are susceptible to motion artefact (Verkruysse, 2008, Humphreys, 2007, Takano, 2007, Hu, 2008, Wieringa, 2005).

In the second group of computer vision based methods there is no need for a dedicated light source. These methods assume that the periodic circulation of the blood by the heart to the rest of the body, including the head, generates some periodic subtle changes to the skin color of the face and also generates some subtle head motions. These motions are not usually visible to naked eyes but they can be viewed by techniques like, for example, Eulerian video magnification (Wu, 2012). These periodic changes to the skin colors and head motions are then utilized to measure heartbeat rate. For example, in (Poh, 2010) periodic changes in the skin color of the face has been used for this purpose. In this system (Poh, 2010) face image of the subject is first found, by a simple camera. Then, it is separated into its colour channels and each channel is tracked independently. For each of these tracked colour channels, a trajectory is found. Then, all the trajectories are fed to an Independent Component Analysis (ICA) algorithm. The output of ICA, presents independents sources that have caused changes to the skin colour of the face. Then, it is assumed that the most periodic output of ICA should be generated by the most periodic source that is present on the face, i.e., heartbeat. This system is effective, but it suffers from sensitivity to skin color and noise. It means, if the skin is not detected properly, or if the captured facial video is noisy, the system does not provide accurate results.

To overcome the sensitivity to noise and skin detection of system (Poh, 2010), very recently in (Balakrishnan, 2013) a motion-based contactless system for measuring heartbeat rate was introduced. As mentioned above, this method is based on the fact that periodic circulation of the blood from the heart to the body, including the head through the aorta and carotid arteries, causes the head to move in a cyclic motion (Wu, 2012). Similar to (Poh, 2010), this system also uses a simple camera for recording facial images of patients. Having detected the face, they extracted vertical component of head motion by tracking feature points, and generate some trajectories for each feature point. These trajectories are then filtered by a Butterworth filter to remove the irrelevant frequencies. Next on the contrary (Poh, 2010) they use Principle Component Analysis (PCA) (instead of ICA) to decompose the filtered trajectories into a set of source signals. Then, they use the same assumption as (Poh, 2010), that the most periodic signal is generated by the most periodic source of the motion that is present in the face, i.e., by heartbeat. To find the periodicity of the outputs of PCA, they apply Fast Fourier Transform (FFT) to the trajectories, and use the percentage of total spectral power of the signal accounted for by the frequency with the maximal power and its first harmonic (Balakrishnan, 2013).

This method gives reasonable results when the face is frontal and does not move. Our experiment shows that involuntary motion and facial expression causes dramatic

effect on the accuracy of this system. Furthermore, as mentioned above, this system is based on using the frequency with maximal power as the first harmonic of the estimated heartbeat rate. But, this assumption is not always true, especially when the facial expression is changing. The proposed system in this paper improves the system of (Balakrishnan, 2013) by replacing the FFT with a Discrete Cosine Transform (DCT). Furthermore, we show that involving a moving average filter before the Butterworth filter improves the results. It is shown that the proposed system outperforms the system of (Balakrishnan, 2013), significantly.

The rest of this paper is organized as follows: The clear problem statement and the contributions of the proposed system are given in the next section. Section 4 explains the employed methodology of the proposed system. The experimental results are reported in Section 5. Finally, the paper is concluded in Section 6.

2.3. **Problem Statement and Main Contribution**

The proposed system in this paper develops a vision-based contactless algorithm for heartbeat rate measurement using the assumption that periodic blood circulation by the heart to the head generates subtle periodic motion on the face. The proposed system is based on the very recent *work* of (Balakrishnan, 2013), but it advances this work by:

1) Replacing the FFT of the system of (Balakrishnan, 2013) by a DCT, and

2) Using a moving average filter before the Butterworth filter that is employed in (Balakrishnan, 2013).

The proposed modifications are simple, but are shown to be very effective. The results of the proposed system are:

1) More correct compared to the results of the system of (Balakrishnan, 2013) when they are compared to the ground truth data.

2) More robust than the results of the system of (Balakrishnan, 2013) when the face is moving or facial expression is changing.

2.4. **Methodology**

The block diagram of the proposed system is shown in figure 2.1. As it can be seen from this figure, the subject is continuously filmed by a Logitech webcam with a resolution of 640x480 pixels. Then, the subject's face is detected by Viola and Jones (Viola, 2001) face detector. From the detected faces, the regions of interest of our system, and consequently the feature points are extracted and tracked by the Lucas Kanade's algorithm (Bouguet, 2000). Then, a moving average filter and a band pass

filter are applied to the vertical component of the trajectories of each feature point to remove extraneous frequencies and involuntary head motions. Then, the filtered trajectories are fed to PCA to find the strongest independent components. Among these components, the most periodic one belongs to heartbeat. To find this most periodic one, we apply DCT to all the components obtained by PCA. Each of these sub-blocks is explained in the following subsections.

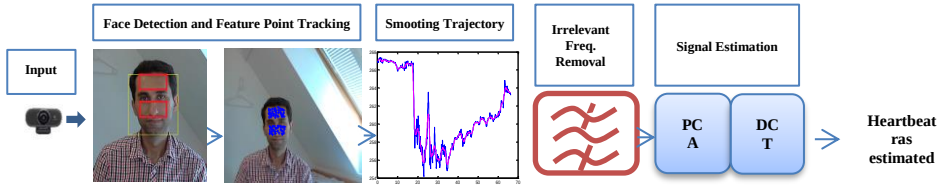


Figure 2.1: The block diagram of the proposed system.

2.4.1. Face Detection

Locating the face in the scene refers to identifying a region containing a human face. Viola and Jones algorithm (Viola, 2001) has been employed for this purpose which is based on Haar-like rectangular features that are extracted from integral images. This detector is fast and efficient, but it fails to detect rotated faces and those which are of poor quality. However, it works fine for the purposes of the proposed system.

The regions detected by the Viola and Jones detector cannot be directly used in our system, as it contains the areas of eyes and the mouth which are not good for the purposes of our system. Because these areas are the most changeable areas of the face, and they may change very much by any changes in facial expression, eye blinking, etc. Therefore, the trajectories obtained from these changeable will not reflect the motion caused by heartbeat. Instead, they reflect the motion caused by the changes in their own positions due to the changes in the facial expression. Tracking these sensitive regions therefore does not produce stable results. The most stable parts of the face, which are robust against changes in the facial expressions, are the forehead and the area around the nose. To keep these regions, we first keep 50% (experimentally obtained) of the width and 70% (experimentally obtained) of the height of the region that is detected the Viola and Jones's face detector. Then, in this refined region we remove the area of the eyes, by removing all the pixels that are located in the range of 25% to 45% (experimentally obtained) of the height of the refined region (figure 2.2).

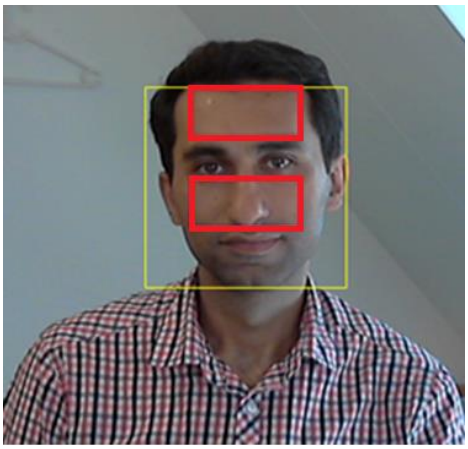


Figure 2: The yellow box is returned by the Viola and Jones face detector and the red boxes are those that are of the interest of the proposed system.

2.4.2. Feature Points selection

Having detected the regions of interest in the previous sub-block of the system, in this step they are fed to the Good Feature Tracking algorithm of (Shi, 1994) to select the feature points. This algorithm is based on finding and tracking the corners. To do so, it calculates the minimal eigenvalue of every point in our previously kept regions of the face and rejecting corners with minimal eigenvalues. Then, it goes through the strongest corners and removes those features that are too close the stronger features (Shi, 1994). To increase the efficiency of this system, it is suggested in (Balakrishnan, 2013) to divide the sub-regions obtained from Viola and Jones detector into smaller areas to achieve uniform selected regions. Therefore, we have adopted this idea here.

2.4.3. Trajectory Generation and Smoothing

To extract the motion trajectory signals from the selected feature points in the previous subsection, we have used Lucas Kanade's algorithm (Bouguet, 2000) to obtain x and y components of feature points inside our previously extracted regions of interest in each frame. Since the very tiny motions of the head, which are the basis for calculating the heartbeat rate in this work, are due to the blood circulation through aorta towards head (obviously in a vertical direction), we only consider the y components of the trajectories of the feature points in each frame.

The head motions are not only due to heartbeats (transferred to the head by aorta), but may appear for several reasons, for example, respiration, vestibular activity, facial expression, speaking and so on. To decrease the effects of the other sources, which cause quite large motions, a moving average filter is applied to the trajectories to smooth it (figure 2.3). This will be further explained in the experimental results.

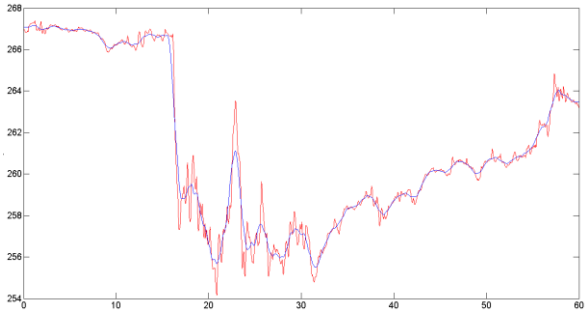


Figure 2.3: The effect of the employed moving average filter on the y components of the trajectory of one of the tracked feature points of one of the test subjects. The red and the blue signals are the original and the filtered signals, on the x axis of the above graph is the time and on the y axis is the y position of the tracked feature point over time.

Then, to remove the irrelevant frequencies (any frequency which might not be generated by the heartbeat) a pass band filter (an 8th order Butterworth filter) with cutoff frequency interval of [0.75 5] Hz has been applied to the obtained trajectory (Balakrishnan, 2013).

2.4.4. Signal Estimation

As mentioned above, the head motions are orginated from different sources and only the one casued by the blood circulation through aorta is reflecting the heartbeat rate. To separate the sources of head motions, we have applied a PCA algorithm to the obtained trajectories. PCA converts the given trajectories into a set of linearlly uncorrelated basis, i.e., the principal components.

Having separated the sources using PCA, the next step is to find the signal that has been generated by the heartbeat. Following (Balakrishnan, 2013) such a signal will be the most periodic signal. To quantify the signal periodicity we have utilized DCT as opposed to the system of (Balakrishnan, 2013) which have used FFT. Having applied DCT, we only keep those DCT components that carry the most significant power of the signal. To do so, we use the following algorithm:

- For the trajectory of the ith feature points, $i \in [1..N]$, S_i:
  - Calculate the DCT of the ith trajectory and obtain SC_i
  - Determine $\{K_j\}_i$ which is the set of indexes for $\{S_i(t)\}$ such that K_j is the index of the M first highest power components into SC_i which consists 50% of power of S_i,
    - $j \in [1..M_i]$ (M_i is number of components which carry 50%of total power of S_i)
  - Determine $\{Kh_l\}_i$ which is the set of the first 5 smallest index into $\{K_j\}_i$ for each S_i such that $2 \times Kh_l$ be found on SC_i
    - $l = 1:5$.

- The periodicity of the signal can be obtained by:

$$Q_i = norm[SC(Kh_l), SC(2 \times Kh_l)]/norm[SC_i]$$

- Si with largest Qi is the heartbeat rate signal, and the heartbeat rate can be obtained as: FFT(IDCT (min{Khi})) × 60 bpm

The effect of the above DCT-based algorithm for finding the heartbeat rate and its advantage over the FFT of (Balakrishnan, 2013) has been shown in the experimental results.

2.5. Experimental Results

The proposed approach has been implemented in Matlab R2013a. To be able to compare our system against state-of-the-art Balakrishnan et al.'s work (Balakrishnan, 2013) we have recorded the actual heartbeat rates of the test subjects by a Shimmer wireless ECG (Electrocardiogram) sensor. This sensor records and sends the ECG signals, to a remote computer as a data file. Figure2.4 (top) shows a typical data that has been captured by this sensor. The FFT of this signal is shown in figure 2.4 (bottom). It can be seen from this figure, that the FFT has 4 peaks on the frequencies 1.08, 2.14, 3.13, and 4.22. These show that most of the power of the recorded heartbeat signal is carried by these 4 component frequencies which seem to be approximately integer multiple components of f_0=1.08 Hz, as a fundamental frequency or first harmonic. Therefore, we can conclude that period of the heartbeat signal per minute is 1.08x60 = 64.8. The numbers of pulses on Figure 4 (bottom) prove this.

Having shown that the ECG signals obtained by the employed sensor are indeed periodic (Figure 4), we now first explain the testing scenarios in which our data have been recoreded. Then, we show the effects of the modifications that we have applied to the system of (Balakrishnan, 2013). Next, we give the details of the comparison of our system against the Balakrishnan et al.'s work (Balakrishnan, 2013).

2.5.1. Testing Scenarios

Five test subjects were asked to participate in testing the systems from which 32 different videos were recorded. These videos are recorded by a Logitech webcam at a frame rate of 30 fps in different facial expressions and head poses. These are the situations in which the videos have been recorded in:

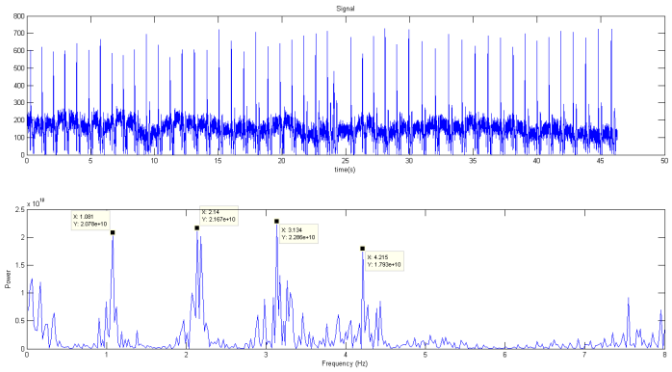


Figure 2.4: Recorded ECG signal and its FFT corresponding signals which shows the periodicity of the ECG signal.

- Subjects look directly into the camera without changing their facial expressions (This is the same imaging condition as the system of (Balakrishnan, 2013)).
- Subjects turn around their faces from left (-180°) to right (+180°) and look at seven different targets that are located at the same distance from each other.
- Subjects show smiling/laughing expression.
- Subjects repeat a given sentence.
- Subjects show angry expression.

The duration of each video is around 60 seconds.

2.5.2. The Moving Average Filter and DCT

Before obtaining the periodicity of the selected source signal (figure 2.1 block diagram), the only difference between our system and the work of (Balakrishnan, 2013) is that we have introduced a moving average filter. This does not have much effect when the face is standing still, and is facing the camera. But, as soon as the subject is changing his/her head pose and/or facial expressions are changing, there will be so many occlusions in the tracking of the feature points, that without using a moving average filter the results will be erroneous. Comparing figure 2.5 (top) to figure 2.5 (bottom) shows that including this moving average filter causes the employed PCA to pick a much smoother signal as the strongest component compared to the case where such a filter has not been included (Balakrishnan, 2013). This will gives us better results, for estimating the heartbeat rates, in the final step of the system.

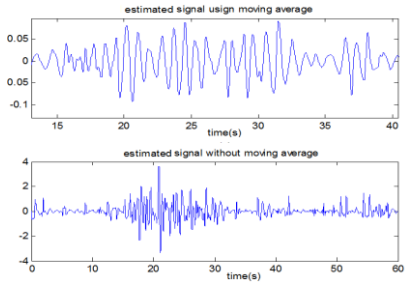


Figure 2.5: Comparing the estimated heartbeat rate signal when the moving average filter is used (top) and when it is not used (bottom).

Besides introducing the moving averge filter for smoothing the estimated signal, in our system we have used DCT to estimate the periodicity of the estimated signal. The effect of this decision and comparing it with the FFT of (Balakrishnan, 2013) is shown in figure 2.6. In this figure (top and middle parts) a signal and its FFT representation are shown. The maximum power of this FFT (3.603) gives a heartbeat rate of 3.603x60 = 216.18, while the actual heartbeat rate in this case is 60 bpm which can be estimated much better using the first harmonic (1.001x60=60.06). Therefore, the total spectral power of the signal and then using the maximal power and its first harmonic as have been used in (Balakrishnan, 2013) does not always produce the desired results. Instead, by using DCT in Figure 6 (bottom) it can be seen that a much better result will be obtained, if the component number 20 is selected as the component which carries the power of pulse frequency. Feeding this value of this component in the algorithm of section 3.4 results in an estimated beat rate of 60.88 bpm, which is very close to the actual value.

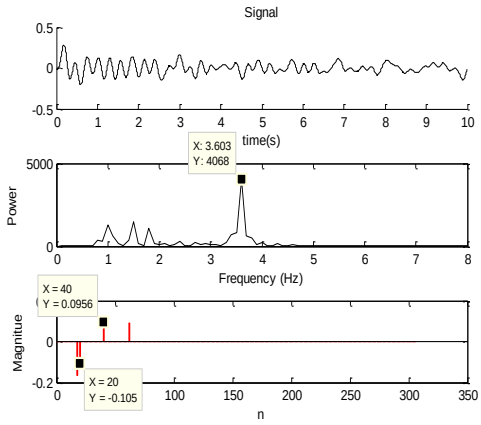


Figure 2.6: Extracting the beat rate of the signal (top) using the algorithm of (Balakrishnan, 2013) (middle) and our employed DCT (bottom).

2.5.3. Detailed Experiments

The proposed system has been compared against the state-of-the-art work of (Balakrishnan, 2013) using the testing data that was recorded in the previously explained testing scenarios. The results of comparing these systems (the proposed system and the work of (Balakrishnan, 2013) against the ground truth data obtained by the Shimmer ECG sensor for the case which the testing subjects are looking directly into the camera are shown in Table 2.1. In this table, (a) is the subject number, (b) is the ground truth data read by a Shimmer ECG device, (c) is the heartbeat rate estimated by the system of (Balakrishnan, 2013), (d) is the error of the method of (Balakrishnan, 2013), (e) is the heartbeat rate estimated by our proposed method, (f) is the error of our proposed method. It can be seen that the error of our system is generally better than that of (Balakrishnan, 2013).

Table 2.1 The proposed system against system of (Balakrishnan, 2013), please see the text for descriptions of the headings

a	b	c	d	e	f
S1-1	61.71	63.06	1.35	62.1	0.39
S1-2	66.67	67.04	0.37	67.03	0.36
S2-1	60.00	216.83	156.8	61.88	1.88
S2-2	59.00	61.06	2.06	59.10	0.1
S2-3	54.00	53.03	0.97	54.11	0.11
S3-1	66.65	69.05	2.40	67.63	0.98
S4-1	84.06	86.06	2.00	83.90	0.16
S5-1	47.62	48.03	0.41	46.17	1.45

The size of the window employed for the moving average filter in the previous experiment is set to one. It means, no moving average is applied to the data obtained from the previous test. Because, the signal is already smooth. But, when it comes to the case where facial expressions and/or head pose are changing, the effect of the moving average becomes more visible. Table 2.2 shows the results of the proposed system against the work of (Balakrishnan, 2013) and the ground truth. The descriptions of the headings (a)-(f) are the same as those for Table 1. The size of the moving average window changes between 40-80 samples, for different testing scenarios. It can be seen from this table that the proposed system is more robust than the work of (Balakrishnan, 2013) in most of the cases, when the facial expression and/or head pose are changing.

Table 2.2: Comparing the results of the proposed system against the state-of-the-art work of (Balakrishnan, 2013) when the facial expressions and/or head pose are changing. Please see the text for the descriptions of the headings.

		The test subject is smiling!				The test subject is speaking!			
a	**b**	**c**	**d**	**e**	**f**	**c**	**d**	**e**	**f**
S1-1	66.67	49.051	17,61	58.22	8.45	81.28	14.61	67.31	0.64
S2-1	59	48.02	10,98	59.35	0.35	75.09	16.09	58.41	0.59
S2-2	54.00	56.04	2.04	54.99	0.99	79.05	25.05	52.66	1.34
S3-1	66.65	48.04	18,61	56.77	9.88	51.03	15,62	66.50	0.15
S4-1	84.06	148.09	64.03	61.73	22.33	61.04	23.02	63.78	20.28
S5-1	47.62	47.03	0.59	48.44	0.82	46.72	0.90	48.14	0.52

		The test subject is angry!				The head pose is changing!			
a	**b**	**c**	**d**	**e**	**f**	**c**	**d**	**e**	**f**
S1-1	66.67	50.05	16.62	59.23	7.44	49.03	17.64	60.47	6.2
S2-1	59	63.04	4.04	59.86	0.86	48.02	10.98	59.69	0.69
S2-2	54.00	49.04	4.96	63.96	9.96	50.04	3.96	53.04	0.96
S3-1	66.65	49.05	17.60	59.92	6.73	50.05	16.6	58.87	7.78
S4-1	84.06	63.03	21.03	47.76	36.3	146.10	62.04	57.87	26.19
S5-1	47.62	50.05	2.43	59.23	11.61	51.05	3.43	45.90	1.72

2.6. Conclutions

Motivated by the fact that in many applications like, e.g., remote patient monitoring, there is not a possibility for installing a device on the body of the patients, this paper has proposed a contactless heartbeat rate measurement using computer vision techniques. The system finds some robust feature points inside the facial areas of the users and tracks them over time to generate some trajectories of the feature points. These trajectories are then smoothed by a moving average filter. Then, the irrelevant frequencies are removed from the trajectories. All of these refined trajectories are then fed to a PCA algorithm to find the strongest independent component. This component is assumed to be the estimated heart beat signal. To find the periodicity of this estimated signal a DCT-based algorithm has been used. Experimental results on several

video sequences show that the estimated heartbeat rates in different facial expressions and head poses are very close to the ground truth. Furthermore, it is shown that the proposed system outperforms state-of-the-art.

2.7. **References**

[1] Balakrishnan, G., Durand, F. and Guttag, J., 2013, Detecting Pulse from Head Motions in Video. Computer Vision and Pattern Recognition, Proceedings CVPR '13, 2013 IEEE Computer Society Conference on.

[2] Bouguet, J., 2000, 'Pyramidal implementation of the Lucas Kanade feature tracker', Intel Corporation, Microprocessor Research Labs, Technical Report.

[3] Hu, S., Zheng, J., Chouliaras, V. and Summers, R., 2008, 'Feasibility of imaging photoplethysmography', in Proceedings of IEEE Conference on Biomedical Engineering and Informatics, pp. 72–75.

[4] Humphreys, K., Ward, T. and Markham, C., 2007, 'Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry', Rev. Sci. Instrum. , vol. 78, Issue 4, pp. 853-857.

[5] Poh, M., McDuff, D. and Picard, R., 2010, 'Non-contact, automated cardiac pulse measurements using video imaging and blind source separation', Optic express, Optics Express, Vol. 18, Issue 10, pp.10762-10774.

[6] Shi, J., Tomasi, C., 'Good features to track,' 1994, Computer Vision and Pattern Recognition, Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, pp.593-600.

[7] Takano, C. and Ohta, Y.,2007, 'Heart rate measurement based on a time-lapse image', Med. Eng. Phys. Vol. 29, Issue 8, pp. 853-857.

[8] Verkruysse, W., Svaasand, L. and Nelson, J., 2008, 'Remote plethysmographic imaging using ambient light', Opt. Express, Vol. 16, Issue 26, pp. 21434-21445.

[9] Viola, P. and Jones, M.,2001, 'Rapid object detection using a boosted cascade of simple features', Computer Vision and Pattern Recognition, , Proceedings CVPR '01., 2001 IEEE Computer Society Conference on, pp. 511-518.

[10] Wieringa, F., Mastik, F. and van der Steen, A., 2005, 'Contactless multiple wavelength photoplethysmographic imaging: a first step toward "SpO2 camera" technology', Ann. Biomed. Eng. Vol. 33, Issue 8, pp.1034-1041.

[11] Wu, H., Rubinstein, M., Shih, E., Guttag, J., Durand, F. and Freeman, W., 2012 'Eulerian Video Magnification for Revealing Subtle Changes in the World', ACM Transactions on Graphics, Proceedings SIGGRAPH 2012, Vol. 31, Issue 4, pp. 65-72.

Chapter 3

Heartbeat Rate Measurement from Facial Video

Mohammad Ahsanul Haque, Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund

This paper has been published in

IEEE Intelligent Systems, vol. 31, no. 3, pp. 40 - 48, 2016

The layout has been revised.

3.1. **Abstract**

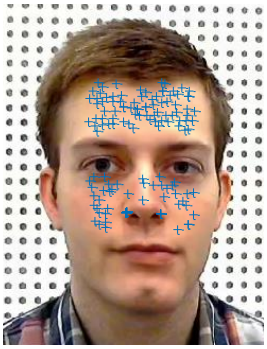
Heartbeat Rate (HR) reveals a person's health condition. This chapter presents an effective system for measuring HR from facial videos acquired in a more realistic environment than the testing environment of current systems. The proposed method utilizes a facial feature point tracking method by combining a 'Good feature to track' and a 'Supervised descent method' in order to overcome the limitations of currently available facial video based HR measuring systems. Such limitations include, e.g., unrealistic restriction of the subject's movement and artificial lighting during data capture. A face quality assessment system is also incorporated to automatically discard low quality faces that occur in a realistic video sequence to reduce erroneous results. The proposed method is comprehensively tested on the publicly available MAHNOB-HCI database and our local dataset, which are collected in realistic scenarios. Experimental results show that the proposed system outperforms existing video based systems for HR measurement.

3.2. **Introduction**

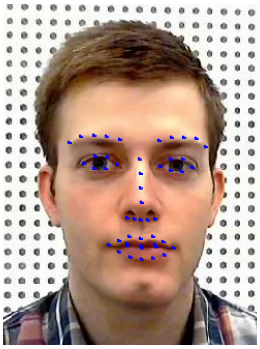
Heartbeat Rate (HR) is an important physiological parameter that provides information about the condition of the human body's cardiovascular system in applications like medical diagnosis, rehabilitation training programs, and fitness assessments [1]. Increasing or decreasing a patient's HR beyond the norm in a fitness assessment or rehabilitation training, for example, can show how the exercise affects the trainee, and indicates whether continuing the exercise is safe.

HR is typically measured by an Electrocardiogram (ECG) through placing sensors on the body. A recent study was driven by the fact that blood circulation causes periodic subtle changes to facial skin color [2]. This fact was utilized in [3]–[7] for HR estimation and [8]–[10] for applications of heartbeat signal from facial video. These facial color-based methods, however, are not effective when taking into account the sensitivity to color noise and changes in illumination during tracking. Thus, Balakrishnan et al. proposed a system for measuring HR based on the fact that the flow of blood through the aorta causes invisible motion in the head (which can be observed by Ballistocardiography) due to pulsation of the heart muscles [11]. An improvement of this method was proposed in [12]. These motion-based methods of [11], [12] extract facial feature points from forehead and cheek (as shown in Figure 3.1.a by a method called Good Feature to Track (GFT). They then employ the Kanade-Lucas-Tomasi (KLT) feature tracker from [13] to generate the motion trajectories of feature points and some signal processing methods to estimate cyclic head motion frequency as the subject's HR. These calculations are based on the assumption that the head is static (or close to) during facial video capture. This means that there is neither internal facial motion nor external movement of the head during the data acquisition phase. We denote internal motion as facial expression and external motion as head pose. In real life scenarios there are, of course, both internal and external head motion. Current

methods, therefore, fail due to an inability to detect and track the feature points in the presence of internal and external motion as well as low texture in the facial region. Moreover, real-life scenarios challenge current methods due to low facial quality in video because of motion blur, bad posing, and poor lighting conditions [14]. These low quality facial frames induce noise in the motion trajectories obtained for measuring the HR.



a



b

Figure 3.1: Different facial feature tracking methods: a. facial feature points extracted by the good feature to track method and b. facial landmarks obtained by the supervised descent method. While GFT extracts a large number of points, SDM merely uses 49 predefined points to track.

The proposed system addresses the aforementioned shortcomings and advances the current automatic systems for reliable measuring of HR. We introduce a Face Quality Assessment (FQA) method that prunes the captured video data so that low quality face frames cannot contribute to erroneous results [15], [16]. We then extract GFT feature points (Figure 3.1.a) of [11] but combine them with facial landmarks (Figure 3.1.b), extracted by the Supervised Descent Method (SDM) of [17]. A combination of these two methods for vibration signal generation allows us to obtain stable trajectories that, in turn, allow a better estimation of HR. The experiments are conducted on a publicly available database and on a local database collected at the lab and a commercial fitness center. The experimental results show that our system outperforms state-of-the-art systems for HR measurement. The chapter's contributions are as follows:

i. We identify the limitations of the GFT-based tracking used in previous methods for HR measurement in realistic videos that have facial expression changes and voluntary head motions, and propose a solution using SDM-based tracking.
ii. We provide evidence for the necessity of combining the trajectories from the GFT and the SDM, instead of using the trajectories from either the GFT or the SDM.

iii. We introduce the notion of FQA in the HR measurement context and demonstrate empirical evidence for its effectiveness.

The rest of the chapter is organized as follows. Section 3 provides the theoretical basis for the proposed method, which is then described in section 4. Section 5 presents the experimental results, and the paper's conclusions are provided in section 6.

3.3. **Theory**

This section describes the basics of GFT- and SDM-based facial point tracking, explains the limitations of the GFT-based tracking, and proposes a solution via a combination of GFT- and SDM-based tracking.

Tracking facial feature points to detect head motion in consecutive facial video frames was accomplished in [11], [12] using GFT-based method. The GFT-based method uses an affine motion model to express changes in the level of intensity in the face. Tracking a window of size $w_x \times w_y$ in frame I to frame J is defined on a point velocity parameter $\boldsymbol{\delta} = [\delta_x \ \delta_y]^T$ for minimizing a residual function f_{GFT} that is defined by:

$$f_{GFT}(\boldsymbol{\delta}) = \sum_{x=p_x}^{p_x+w_x} \sum_{y=p_y}^{p_y+w_y} \big(I(\boldsymbol{x}) - J(\boldsymbol{x}+\boldsymbol{\delta})\big)^2 \tag{1}$$

where $(I(\mathbf{x}) - J(\mathbf{x}+\boldsymbol{\delta}))$ stands for $(I(x,y) - J(x+\delta_x, y+\delta_y))$, and $\mathbf{p} = [p_x, p_y]^T$ is a point to track from the first frame to the second frame. According to observations made in [18], the quality of the estimate by this tracker depends on three factors: the size of the window, the texture of the image frame, and the amount of motion between frames. Thus, in the presence of voluntary head motion (both external and internal) and low texture in facial videos, the GFT-based tracking exhibits the following problems:

i. Low texture in the tracking window: In general, not all parts of a video frame contain complete motion information because of an aperture problem. This difficulty can be overcome by tracking feature points in corners or regions with high spatial frequency content. However, GFT-based systems for HR utilized the feature points from the forehead and cheek that have low spatial frequency content.
ii. Losing track in a long video sequence: The GFT-based method applies a threshold to the cost function $f_{GFT}(\boldsymbol{\delta})$ in order to declare a point 'lost' if the cost function is higher than the threshold. While tracking a point over many frames of a video, as done in [11], [12], the point may drift throughout the extended sequences and may be prematurely declared 'lost.'
iii. Window size: When the window size (i.e. $w_x \times w_y$ in (1)) is small a deformation matrix to find the track is harder to estimate because the variations of

motion within it are smaller and therefore less reliable. On the other hand, a bigger window is more likely to straddle a depth discontinuity in subsequent frames.

iv. Large optical flow vectors in consecutive video frames: When there is voluntary motion or expression change in a face the optical flow or face velocity in consecutive video frames is very high and GFT-based method misses the track due to occlusion [13].

Instead of tracking feature points by GFT-based method, facial landmarks can be tracked by employing a face alignment system. The Active Appearance Model (AAM) fitting [19] and its derivatives [20] are some of the early solutions for face alignment. A fast and highly accurate AAM fitting approach that was proposed recently in [17] is SDM. The SDM uses a set of manually aligned faces as training samples to learn a mean face shape. This mean shape is then used as an initial point for an iterative minimization of a non-linear least square function towards the best estimates of the positions of the landmarks in facial test images. The minimization function can be defined as a function over Δx:

$$f_{SDM}(x_0 + \Delta x) = \left\| g\big(d(x_0 + \Delta x)\big) - \theta_* \right\|_2^2 \tag{2}$$

where x_0 is the initial configuration of the landmarks in a facial image, $d(x)$ indexes the landmarks configuration (x) in the image, g is a nonlinear feature extractor, $\theta_* = g(d(x_*))$, and x_* is the configuration of the true landmarks. In the training images Δx and θ_* are known. By utilizing these known parameters the SDM iteratively learns a sequence of generic descent directions, $\{\partial_n\}$, and a sequence of bias terms, $\{\beta_n\}$, to set the direction towards the true landmarks configuration x_* in the minimization process, which are further applied in the alignment of unlabelled faces [17]. The evaluation of the descent directions and bias terms is accomplished by:

$$x_n = x_{n-1} + \partial_{n-1}\sigma(x_{n-1}) + \beta_{n-1} \tag{3}$$

where $\sigma(x_{n-1}) = g(d(x_{n-1}))$ is the feature vector extracted at the previous landmark location x_{n-1}, x_n is the new location, and ∂_{n-1} and β_{n-1} are defined as:

$$\partial_{n-1} = -2 \times \boldsymbol{H}^{-1}(x_{n-1}) \times \boldsymbol{J}^T(x_{n-1}) \times g(d(x_{n-1})) \tag{4}$$

$$\beta_{n-1} = -2 \times \boldsymbol{H}^{-1}(x_{n-1}) \times \boldsymbol{J}^T(x_{n-1}) \times g(d(x_*)) \tag{5}$$

where $\mathbf{H}(x_{n-1})$ and $\mathbf{J}(x_{n-1})$ are, respectively, the Hessian and Jacobian matrices of the function g evaluated at (x_{n-1}). The succession of x_n converges to x_* for all images in the training set.

The SDM is free from the problems of the GFT-based tracking approach for the following reasons:

i. Low texture in the tracking window: The 49 facial landmarks of SDM are taken from face patches around eye, lip, and nose edges and corners (as shown in Figure 3.1.b, which have high spatial frequency due to the existence of edges and corners as discussed in [18].
ii. Losing track in a long video sequence: The SDM does not use any reference points in tracking. Instead, it detects each point around the edges and corners in the facial region of each video frame by using supervised descent directions and bias terms as shown in (3), (4) and (5). Thus, the problems of point drifting or dropping a point too early do not occur.
iii. Window size: The SDM does not define the facial landmarks by using the window based 'neighborhood sense' and, thus, does not use any window-based point tracking system. Instead, the SDM utilizes the 'neighborhood sense' on a pixel-by-pixel basis along with the descent detections and bias terms.
iv. Large optical flow vectors in consecutive video frames: As mentioned in [13], occlusion can occur by large optical flow vectors in consecutive video frames. As a video with human motion satisfies temporal stability constraint [21], increasing the search space can be a solution. SDM uses supervised descent direction and bias terms that allow searching selectively in a wider space with high computational efficiency.

Though GFT-based method fails to preserve enough information to measure the HR when the video has facial expression change or head motion, it uses a larger number of facial feature points (e.g., more than 150) to track than SDM (only 49 points). This matter causes the GFT-based method to generate a better trajectory than SDM when there is no voluntary motion. On the other hand, SDM does not miss or erroneously track the landmarks in the presence of voluntary facial motions. In order to exploit the advantages of the both methods, a combination of GFT- and SDM-based tracking outcome can be used, which is explained in the methodology section. Thus, merely using GFT or SDM to extract facial points in cases where subjects may have both voluntary motion and non-motion periods does not produce competent results.

3.4. **The Proposed Method**

A block diagram of the proposed method is shown in Figure 3.2. The steps are explained below.

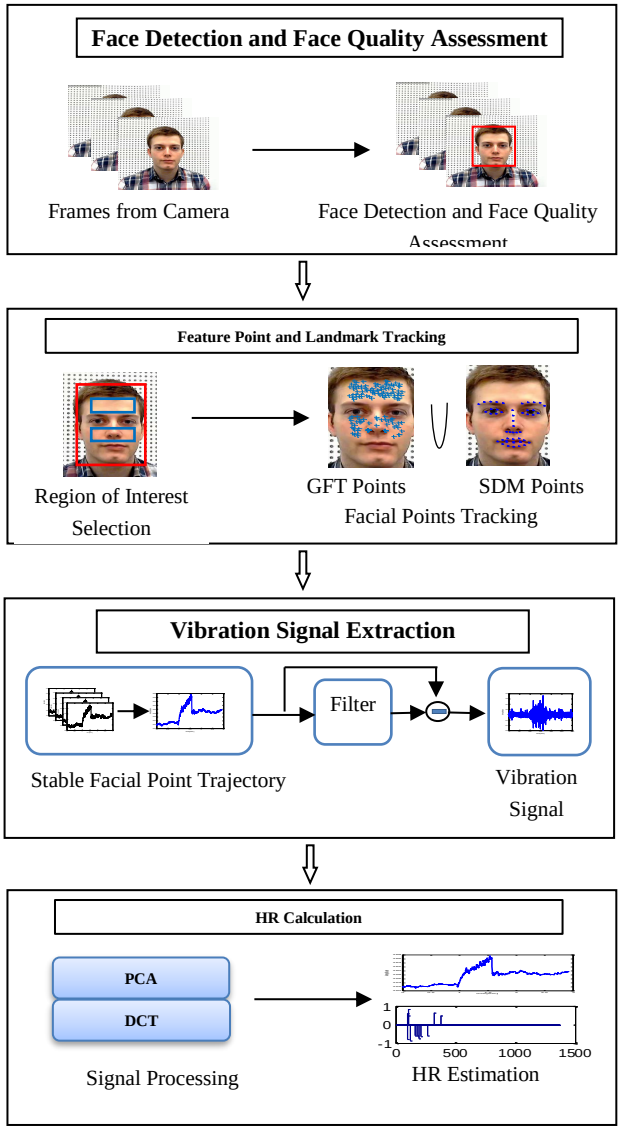


Figure 3.2: The block diagram of the proposed system. We acquire the facial video, track the intended facial points, extract the vibration signal associated with heartbeat, and estimate the HR.

3.4.1. Face Detection and Face Quality Assessment

The first step of the proposed motion-based system is face detection from facial video acquired by a webcam. We employed the Haar-like features of Viola and Jones to

extract the facial region from the video frames [22]. However, facial videos captured in real-life scenarios can exhibit low face quality due to the problems of pose variation, varying levels of brightness, and motion blur. A low quality face produces erroneous results in facial feature points or landmarks tracking. To solve this problem, a FQA module is employed by following [16], [23]. The module calculates four scores for four quality metrics: resolution, brightness, sharpness, and out-of-plan face rotation (pose). The quality scores are compared with thresholds (following [23], with values 150x150, 0.80, 0.8, and 0.20, for resolution, brightness, sharpness, and pose, respectively) to check whether the face needs to be discarded. If a face is discarded, we concatenate the trajectory segments to remove discontinuity by following [5]. As we measure the average HR over a long video sequence (e.g. 30 secs to 60 secs) discarding few frames (e.g., less than 5% of the total frames) does not greatly affect the regular characteristic of the trajectories but removes the most erroneous segments coming from low quality faces.

3.4.2. Feature Points and Landmarks Tracking

Tracking facial feature points and generating trajectory keep record of head motion in facial video due to heartbeat. Our objective with trajectory extraction and signal processing is to find the cyclic trajectories of tracked points by removing the non-cyclic components from the trajectories. Since GFT-based tracking has some limitations, as we discussed in the previous section, having voluntary head motion and facial expression change in a video produces one of two problems: i) completely missing the track of feature points and ii) erroneous tracking. We observed more than 80% loss of feature points by the system in such cases. In contrast, the SDM does not miss or erroneously track the landmarks in the presence of voluntary facial motions or expression change as long as the face is qualified by the FQA. Thus, the system can find enough trajectories to measure the HR. However, the GFT uses a large number of facial points to track when compared to SDM, which uses only 49 points. This causes the GFT to preserve more motion information than SDM when there is no voluntary motion. Hence, merely using GFT or SDM to extract facial points in cases where subjects may have both voluntary motion and non-motion periods does not produce competent results. We therefore propose to combine the trajectories of GFT and SDM. In order to generate combined trajectories, the face is passed to the GFT-based tracker to generate trajectories from facial feature points and then appended with the SDM trajectories. Let the trajectories be expressed by location time-series $S_{t,n}(x, y)$, where (x, y) is the location of a tracked point n in the video frame t.

3.4.3. Vibration Signal Extraction

The trajectories from the previous step are usually noisy due to, e.g., voluntary head motion, facial expression, and/or vestibular activity. We reduce the effect of such noises by employing filters to the vertical component of the trajectories of each feature point. An 8th order Butterworth band pass filter with cutoff frequency of *[0.75-*

5.0] Hz (human HR lies within this range [11]) is used along with a moving average filter defined below:

$$S_n(t) = \frac{1}{w}\sum_{i=-\frac{w}{2}}^{\frac{w}{2}-1} S_n(t+i)\,, where\ \frac{w}{2} < t < T - \frac{w}{2} \tag{6}$$

where w is the length of the moving average window (length is 300 in our experiment) and T is the total number of frames in the video. These filtered trajectories are then passed to the HR measurement module.

3.4.4. Heartbeat Rate (HR) Measurement

As head motions can originate from different sources and only those caused by blood circulation through the aorta reflect the heartbeat rate, we apply a Principal Component Analysis (PCA) algorithm to the filtered trajectories (S) to separate the sources of head motion. PCA transforms S to a new coordinate system through calculating the orthogonal components P by using a load matrix L as follows:

$$P = S\,.\,L \tag{7}$$

where L is a $T \times T$ matrix with columns obtained from the eigenvectors of $S^T S$. Among these components, the most periodic one belongs to heartbeat as obtained in [11]. We apply Discrete Cosine Transform (DCT) to all the components (P) to find the most periodic one by following [12]. We then employ Fast Fourier Transform (FFT) on the inverse-DCT of the component and select the first harmonic to obtain the HR.

3.5. Experimental Environments and Datasets

This section describes the experimental environment, evaluates the performance of the proposed system, and compares the performance with the state-of-the-art methods.

3.5.1. Experimental Environment

The proposed method was implemented using a combination of Matlab (SDM) and C++ (GFT with KLT) environments. We used three databases to generate results: a local database for demonstrating the effect of FQA, a local database for HR measurement, and the publicly available MAHNOB-HCI database [24]. For the first database, we collected 6 datasets of 174 videos from 7 subjects to conduct an experiment to report the effectiveness of employing FQA in the proposed system. We put four webcams (Logitech C310) at 1, 2, 3, and 4 meter(s) distances to acquire facial video with four different face resolution of the same subject. The room's lighting condition

was changed from bright to dark and vice versa for the brightness experiment. Subjects were requested to have around 60 degrees out-of-plan pose variation for the pose experiment. The second database contained 64 video clips by defining three scenarios to constitute our own experimental database for HR measurement experiment, which consists of about 110,000 video frames of about 3,500 seconds. These datasets were captured in two different setups: a) an experimental setup in a laboratory, and b) a real-life setup in a commercial fitness center. The scenarios were:

i. **Scenario 1 (normal):** Subjects exposed their face in front of the cameras without any facial expression or voluntary head motion (about 60 seconds).

ii. **Scenario 2 (internal head motion):** Subjects made facial expressions (smiling/laughing, talking, and angry) in front of the cameras (about 40 seconds).

iii. **Scenario 3 (external head motion):** Subjects made voluntary head motion in different directions in front of the cameras (about 40 seconds).

The third database was the publicly available MAHNOB-HCI database, which has 491 sessions of videos longer than 30 seconds and to which subjects consent attribute 'YES'. Among these sessions, data for subjects '12' and '26' were missing. We collected the rest of the sessions as a dataset for our experiment, which are hereafter called MAHNOB-HCI_Data. Following [5], we use 30 seconds (frame 306 to 2135) from each video for HR measurement and the corresponding ECG signal for the ground truth.

Table 3.1 summarizes all the datasets we used in our experiment.

3.5.2. Performance Evaluation

The proposed method used a combination of the SDM- and GFT-based approaches for trajectory generation from the facial points. Figure 3.3 shows the calculated average trajectories of tracked points in two experimental videos. We included the trajectories obtained from GFT [13], [18] and SDM [16], [17] for facial videos with voluntary head motion. We also included some example video frames depicting face motion. As observed from the figure, the GFT and SDM provide similar trajectories when there is little head motion (video1, Figure 3.3.b, and c). When the voluntary head motion is sizable (beginning of video2, Figure 3.3.e and f), GFT-based method fails to track the point accurately and thus produces an erroneous trajectory because of large optical flow. However, SDM provides stable trajectory in this case, as it does not suffer from large optical flow. We also observe that the SDM trajectories provide more sensible amplitude than the GFT trajectories, which in turn contributes to clear separation of heartbeat from the noise.

Table 3.1: Dataset names, definitions and sizes

No	Name	Definition	Number of data
1.	Lab_HR_Norm_Data	Video data for HR measurement collected for lab scenario 1.	10
2.	Lab_HR_Expr_Data	Video data for HR measurement collected for lab scenario 2.	9
3.	Lab_HR_Motion_Data	Video data for HR measurement collected for lab scenario 3.	10
4.	FC_HR_Norm_Data	Video data for HR measurement collected for fitness center scenario 1.	9
5.	FC_HR_Expr _Data	Video data for HR measurement collected for fitness center scenario 2.	13
6.	FC_HR_Motion_Data	Video data for HR measurement collected for fitness center scenario 3.	13
7.	MAHNOB-HCI_Data	Video data for HR measurement from [24]	451
8.	Res1, Res2, Res3, Res4	Video data acquired from 1, 2, 3 and 4 meter(s) distances, respectively, for FQA experiment	29x4
9.	Bright_FQA	Video data acquired while lighting changes for FQA experiment	29
10.	Pose_FQA	Video data acquired while pose variation occurs for FQA experiment	29

Unlike [11], the proposed method utilizes a moving average filter before employing PCA on the trajectory obtained from the tracked facial points and landmarks. The effect of this moving average filter is shown in Figure 3.4.a. The moving average filter reduces noise and softens extreme peaks in voluntary head motion and provides a smoother signal to PCA in the HR detection process.

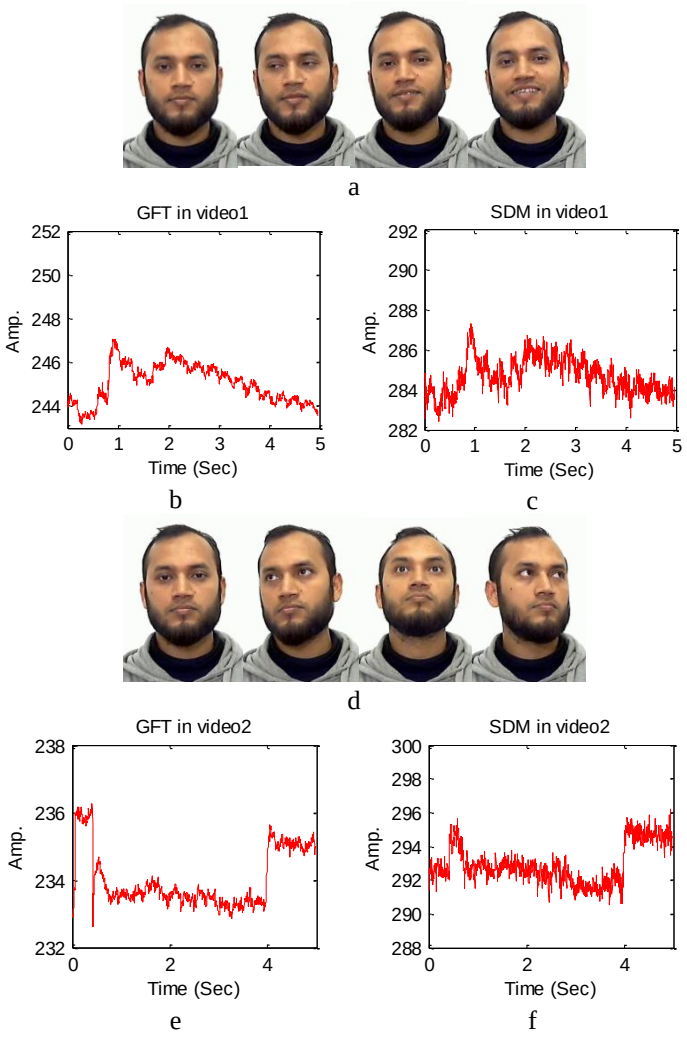
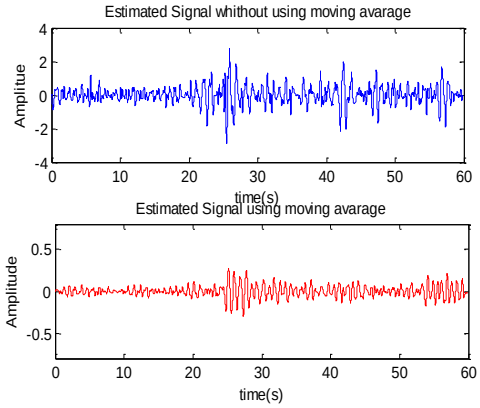


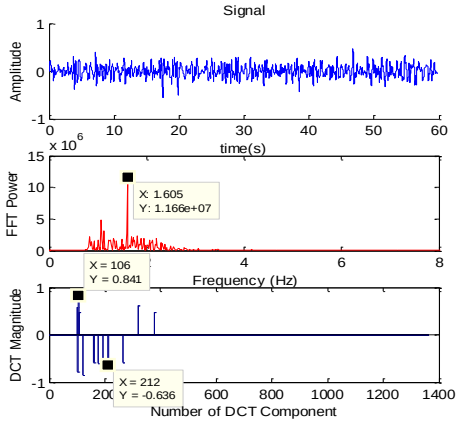
Figure 3.3: Example frames depict small motion (in a) and large motion (in d) from a video, and trajectories of tracking points extracted by GFT [18] (in b and e) and SDM [17] (in c and f) from 5 seconds of two experimental video sequences with small motion (video1) and large motion at the beginning and end (video2).

The proposed method utilizes DCT instead of FFT of [11] in order to calculate the periodicity of the cyclic head motion signal. Figure 3.4.b shows a trajectory of head motion from an experimental video and its FFT and DCT representations after pre-processing. In the figure we see that the maximum power of FFT is at frequency bin 1.605. This, in turn, gives HR 1.605x60=96.30, whereas the actual HR obtained from ECG was 52.04bpm. Thus, the method in [11] that used FFT in the HR estimation does not always produce good results. On the other hand, using DCT by following

[12] yields a result of 52.35 bpm from the selected DCT component X=106. This is very close to the actual HR.



a



b

Figure 3.4: The effect of a. the moving average filter on the trajectory of facial points to get a smoother signal by noise and extreme peaks reduction and b. the difference between extracting the periodicity (heartbeat rate) of a cyclic head motion signal by using fast Fourier transform (FFT) power and discrete cosine transform (DCT) magnitude.

Furthermore, we conducted an experiment to demonstrate the effect of employing FQA in the proposed system. The experiment had three sections for three quality metrics: resolution, brightness, and out-of-plan pose. The results of HR measurement on six datasets collected for FQA experiment are shown in Table 3.2. From the results, it is clear that when resolution decreases the accuracy of the system decreases

accordingly. Thus, FQA for face resolution is necessary to ensure a good size face in the system. The results also show that the brightness variation and the pose variation have influence on the HR measurement. We observe that when frames of low quality, in terms of brightness and pose, are discarded the accuracy of HR measurement increases.

Table 3.2: Analysing the effect of the FQA in HR measurement

Exp. Name	**Dataset**	**Average percentage (%) of error in HR measurement**
Resolution	Res1	10.65
	Res2	11.74
	Res3	18.86
	Res4	37.35
Brightness	Bright_FQA before FQA	18.77
	Bright_FQA after FQA	17.62
Pose variation	Pose_FQA before FQA	17.53
	Pose_FQA after FQA	14.01

3.5.3. Performance Comparison

We have compared the performance of the proposed method against state-of-the-art methods from [3], [5], [6], [11], [12] on the experimental datasets listed in

Table 3.1. Table 3.3 lists the accuracy of HR measurement results of the proposed method in comparison with the motion-based state of the art methods [11], [12] on our local database. We have measured the accuracy in terms of percentage of measurement error. The lower the error generated by a method, the higher the accuracy of that method. From the results we observe that the proposed method showed consistent performance, although the data acquisition scenarios were different for different datasets. By using both GFT and SDM trajectories, the proposed method gets more trajectories to estimate the HR pattern in the case of HR_Norm_Data and accurate trajectories due to non-missing facial points in the cases of HR_Expr_Data and HR_Motion_Data. On the other hand, the previous methods suffer from fewer trajectories and/or erroneous trajectories from the data acquired in challenging scenarios, e.g. Balakrishnan's method showed an up to 25.07% error in HR estimation from

videos having facial expression change. The proposed method outperforms the previous methods in both environments (lab and in a fitness center) of data acquisition, including all three scenarios.

Table 3.3: Performance comparison between the proposed method and the state-of-the-art methods of HR measurement on our local databases

Dataset name	Average percentage (%) of error in HR measurement		
	Balakrishnan et al. [11]	Irani et al. [12]	The proposed method
Lab_HR_Norm_Data	7.76	7.68	2.80
Lab_HR_Expr_Data	13.86	9.00	4.98
Lab_HR_Motion_Data	16.84	5.59	3.61
FC_HR_Norm_Data	8.07	10.75	5.11
FC_HR_Expr_Data	25.07	10.16	6.23
FC_HR_Motion_Data	23.90	15.16	7.01

Table 3.4: Performance comparison between the proposed method and the state-of-the-art methods of HR measurement on MAHNOB-HCI database

Method	RMSE (bpm)	Mean error rate (%)
Poh et al. [3]	*25.90*	*25.00*
Kwon et al. [7]	*25.10*	*23.60*
Balakrishnan et al. [11]	*21.00*	*20.07*
Poh et al. [6]	*13.60*	*13.20*
Li et al. [5]	*7.62*	*6.87*
Irani et al. [12]	*5.03*	*6.61*
The proposed method	*3.85*	*4.65*

Table 3.4 shows the performance comparison of HR measurement by our proposed method and state-of-the-art methods (both color-based and motion-based) on MAHNOB-HCI_Data. We calculate the Root Mean Square Error (RMSE) in beat-per-minute (bpm) and mean error rate in percentage to compare the results. From the results we can observe that Li's [5], Irani's [12], and the proposed method showed

considerably higher results than the other methods because they take into consideration the presence of voluntary head motion in the video. However, unlike Li's color-based method, Irani's method and the proposed method are motion-based. Thus, changing the illumination condition in MAHNOB-HCI_Data does not greatly affect the motion-based methods, as indicated by the results. Finally, we observe that the proposed method outperforms all these state-of-the-art methods in the accuracy of HR measurement.

3.6. **Conclusions**

This chapter proposes a system for measuring HR from facial videos acquired in more realistic scenarios than the scenarios of previous systems. The previous methods work well only when there is neither voluntary motion of the face nor change of expression and when the lighting conditions help keeping sufficient texture in the forehead and cheek. The proposed method overcomes these problems by using an alternative facial landmarks tracking system (the SDM-based system) along with the previous feature points tracking system (the GFT-based system) and provides competent results. The performance of the proposed system for HR measurement is highly accurate and reliable not only in a laboratory setting with no-motion, no-expression cases in artificial light in the face, as considered in [11], [12], but also in challenging real-life environments. However, the proposed system is not adapted yet to the real-time application for HR measurement due to dependency on temporal stability of the facial point trajectory.

3.7. **References**

[1] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, Distributed Computing and Monitoring Technologies for Older Patients, 1st ed. Springer International Publishing, 2015.

[2] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian Video Magnification for Revealing Subtle Changes in the World," ACM Trans Graph, vol. 31, no. 4, pp. 65:1–65:8, Jul. 2012.

[3] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," Opt. Express, vol. 18, no. 10, pp. 10762–10774, May 2010.

[4] H. Monkaresi, R.. Calvo, and H. Yan, "A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam," IEEE J. Biomed. Health Inform., vol. 18, no. 4, pp. 1153–1160, Jul. 2014.

[5] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote Heart Rate Measurement From Face Videos Under Realistic Situations," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 4321–4328.

[6] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam,” IEEE Trans. Biomed. Eng., vol. 58, no. 1, pp. 7–11, Jan. 2011.

[7] S. Kwon, H. Kim, and K. S. Park, “Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone,” in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2012, pp. 2174–2177.

[8] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Heartbeat Signal from Facial Video for Biometric Recognition,” in Image Analysis, R. R. Paulsen and K. S. Pedersen, Eds. Springer International Publishing, 2015, pp. 165–174.

[9] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Can contact-free measurement of heartbeat signal be used in forensics?,” in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 2015, pp. 1–5.

[10] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Efficient contactless heartbeat rate measurement for health monitoring,” Internatinal J. Integr. Care, vol. 15, no. 7, pp. 1–2, Oct. 2015.

[11] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting Pulse from Head Motions in Video,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3430–3437.

[12] R. Irani, K. Nasrollahi, and T. B. Moeslund, “Improved Pulse Detection from Head Motions Using DCT,” in 9th International Conference on Computer Vision Theory and Applications (VISAPP), 2014, pp. 1–8.

[13] J. Bouguet, “Pyramidal implementation of the Lucas Kanade feature tracker,” Intel Corp. Microprocess. Res. Labs, 2000.

[14] A. D. Bagdanov, A. Del Bimbo, F. Dini, G. Lisanti, and I. Masi, “Posterity Logging of Face Imagery for Video Surveillance,” IEEE Multimed., vol. 19, no. 4, pp. 48–59, Oct. 2012.

[15] K. Nasrollahi and T. B. Moeslund, “Extracting a Good Quality Frontal Face Image From a Low Resolution Video Sequence,” IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 10, pp. 1353–1362, Oct. 2011.

[16] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Quality-Aware Estimation of Facial Landmarks in Video Sequences,” in IEEE Winter Conference on Applications of Computer Vision (WACV), 2015, pp. 1–8.

[17] X. Xiong and F. De la Torre, “Supervised Descent Method and Its Applications to Face Alignment,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532–539.

[18] J. Shi and C. Tomasi, “Good features to track,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994, pp. 593–600.

[19] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.

[20] A. U. Batur and M. H. Hayes, "Adaptive active appearance models," IEEE Trans. Image Process., vol. 14, no. 11, pp. 1707–1721, Nov. 2005.

[21] Y. Feng, J. Xiao, Y. Zhuang, X. Yang, J. J. Zhang, and R. Song, "Exploiting temporal stability and low rank structure for motion capture data refinement," Inf. Sci., vol. 277, pp. 777–793, Sep. 2014.

[22] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," Int J Comput Vis., vol. 57, no. 2, pp. 137–154, May 2004.

[23] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera," in 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2013, pp. 443–448.

[24] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 42–55, Jan. 2012.

Chapter 4

Contactless Measurement of Muscles Fatigue by Tracking Facial Feature Points in a Video

Ramin Irani, Kamal Nasrollahi and Thomas B. Moeslund

This paper has been published in

IEEE International Conference on Image Processing (ICIP) - Paris, France, 2014, pp. 4181-4185.

The layout has been revised.

4.1. **Abstract**

Physical exercise may result in muscle tiredness which is known as muscle fatigue. This occurs when the muscles cannot exert normal force, or when more than normal effort is required. Fatigue is a vital sign, for example, for therapists to assess their patient's progress or to change their exercises when the level of the fatigue might be dangerous for the patients. The current technology for measuring tiredness, like Electromyography (EMG), requires installing some sensors on the body. In some applications, like remote patient monitoring, this however might not be possible. To deal with such cases, in this paper we present a contactless method based on computer vision techniques to measure tiredness by detecting, tracking, and analyzing some facial feature points during the exercise. Experimental results on several test subjects and comparing them against ground truth data show that the proposed system can properly find the temporal point of tiredness of the muscles when the test subjects are doing physical exercises.

4.2. **Introduction**

Fatigue is defined as feeling weakness. It is referred to as tiredness and exhaustion. It causes temporary inability in maintaining optimal cognitive or muscle performance. Fatigue can be mental or physical [1, 2]. In mental fatigue, patients cannot concentrate on a problem or cannot perform their daily activities as easy as they used to. But in physical fatigue, person's muscle feels weakness. Muscle fatigue impairs the normal performance capacity of the muscles as it takes more energy than normal case to achieve a desired performance. For instance, when you lift a very heavy weight or you hold your muscles in one position for a long time (called isometric contraction [3]), muscles get tired.

Physical or muscles fatigue is an important sign, for instance, for therapists for taking care of patient's progress. Based on such monitoring of patients, therapists can change the exercise, make it is easier or even stop it when the level of fatigue goes beyond a level that might be harmful for the patient. Nowadays, measuring muscle fatigue is usually done by a direct contact between the muscles and a sensor. Such sensors can be a force gauge, EMG electrodes, Mechanomyogram (MMG) sensors. Measuring the fatigue using a force gauge is very easy. But, it requires some devices like a hand grip dynamometer [4]. It, hence, is impossible to measure the fatigue using this method for an exercise using, for example, dumbbells. The EMG method uses electrodes to detect electrical current when muscles are contracted [5]. EMG can record signals from muscles, which can be accomplished using two approaches know as invasive (needle electrode-based) and non-invasive (skin surface electrode-based). The non-invasive one which is also known as surface EMG (sEMG) is popular for collecting signals from muscle fatigues [6]. This technique has been used very often [7-11], though it is complex to implement, particularly in automatic fatigue detection.

EMG signal is very sensitive to noise which generally should be filtered. It also requires wearing adhesive gel patches that may cause skin irritation and slight pain. The (MMG) is another non-invasive method for assessing of muscle fatigue which is often used with EMG technique. EMG records electrical signals, but, MMG captures mechanical signals generated from muscle contraction. Similar to EMG and other fatigue detecting techniques, the sensors applied in MMG, such as accelerometer, goniometer and microphone [12-14] require direct skin contact. MMG cannot be used for dynamic contraction. Moreover, they are expensive and physically balkier. Furthermore, MMG, similar to EMG, is sensitive to noise [5].

To overcome the problems of contact-based methods of muscle fatigue measurement, in this paper, we develop a contactless computer vision technique. The proposed method is based on the work of [15-17] which show that heartbeat rate can be measured from facial images. The point here is that blood circulation to the head makes some periodic movements on the face which are not visible to the naked eyes, but can be revealed by video magnification [18] to measure the heartbeat rate. We have extended the same concept to muscle fatigue measurement. We utilize this fact that the energy that is released from shaking of muscles (due to tiredness) results in shaking of the face which might not always be detected by naked eyes, but can be well discovered by computer vision techniques similar to those of [15, 16]. This actually makes good sense because any motion or any contraction during muscles activity happens by a group of motor units (including motor neuron and the skeletal muscle fibers). When a muscle is fatigued, some of the motor units drop out of service and leading to muscle's shaking status which consequently results in shacking of the face. To the best of our knowledge, there is not any similar previous works on detecting muscle fatigue using computer vision techniques.

The rest of this paper is organized as follows: Section 3 presents the details of the proposed approach for detecting muscle fatigue. Experimental results and performance evaluation of the proposed system are discussed in Section 4. Finally, conclusions are drawn in Section 5.

4.3. **The Proposed System**

As mentioned earlier, muscles start shaking after they get tired due to an activity. This shaking gets reflected on the face. This is exactly the purpose of this paper to detect this shaking by analyzing facial image and tracking specific facial features for measuring the muscles fatigue. The block diagram of the proposed system is shown in figure 4.1. First a camera (a Logitech webcam) is continuously filming the subject with a resolution of 640x480 pixels. Then, the subject's face is detected by Viola and Jones [19] face detector. Then, we extract and track some of the facial feature points. Thereafter, we extract muscle fatigue-related vibration signal of the head by removing large head motion using a moving average from the trajectories of the chosen

facial features [15]. Afterwards, the extracted vibrating signal is segmented and filtered using a pass band filter to calculate the released energy from the vibrating signal. Before filtering, to enhance the coherency between the blocks (segmented sequences) and decreasing windowing effects, 75% overlapping with Hamming window is utilized. Then, we calculate the power spectral density to obtain the energy that is released due to shaking of the face to finally index the fatigue. These are explained in the following subsections.

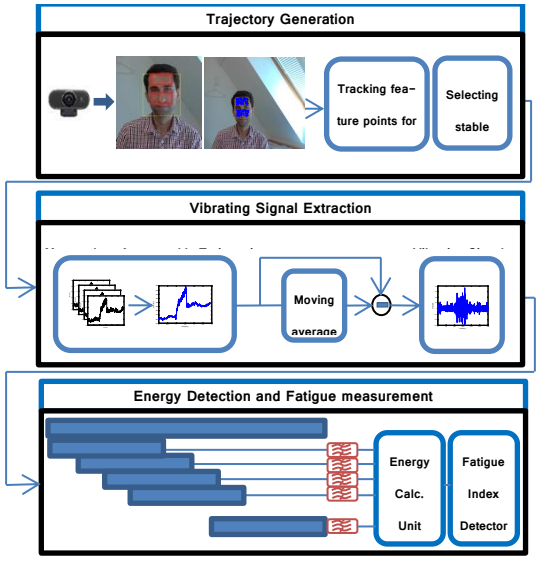


Figure 4.1: The block diagram of the proposed system

4.3.1. Trajectory generation

From the detected faces by Viola and Jones face detector [19] we extract the facial regions of interest using the method of [15]. These regions contain stable facial feature points which are those points that are not sensitive to changes in facial expressions. These facial features points are chosen and then tracked over time by [20] to generate facial feature trajectories. Then, we only keep those trajectories which their displacements in any two consecutive frames are not larger than a predefined threshold.

4.3.2. Muscle fatigue-related vibrating signal extraction

The chosen trajectories in the previous step are used to extract muscle fatigue-related vibrating signal. The trajectories are usually noisy due to, for example, error in feature tracking and any unwanted muscle motion like facial expression. To reduce the effect of such noises we use a mean filter. To do so, we use:

$$T(n) = \frac{1}{M}\sum_{m=1}^{M}(y_m(n) - \bar{y}_m) \quad (1)$$

where $T(n)$ is the shifted mean filtered trajectory, $y_m(n)$ is the n_{th} frame of the trajectory m, M is the number of the trajectories, N is the number of the frames in each trajectory and $\bar{y}_m$ is the mean value of the trajectory m, which is given by:

$$\bar{y}_m = \frac{1}{N}\sum_{n=1}^{N} y_m(n) \quad (2)$$

Then, the vibrating signal $V_s(n)$ which carries the shaking information of is obtained by:

$$V_s(n) = T(n) - \frac{1}{R}.\sum_{r=0}^{R-1} T(n-r) \quad (3)$$

where R is the number of points involved in the averaging (here we use the experimentally obtained value of 35).

4.3.3. Energy measurement and fatigue detection

To measure the released energy of the muscles we need to segment the trajectories (with length t_{sec} to small time blocks with Δt_{sec} length). Segmenting the trajectories help us to measure the fatigue in the steps of time. After windowing, each block is filtered by a pass band ideal filter with cut off frequency interval of [3-5] Hz. Figure 4.2 shows the power of the filtered vibrating signal with cut off frequency interval of [3-5] Hz. We observe that the power of the signal rises up when the fatigue happens (interval of [127–185]).

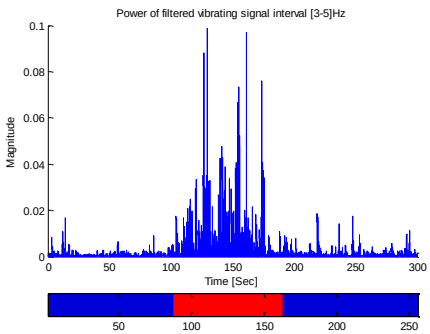


Figure 4.2: Power of the trajectory during fatigue test: The Blue regions are the resting time and the red region shows the fatigue due to exercise.

After filtering, the energy of i_{th} block, E_i, is calculated as:

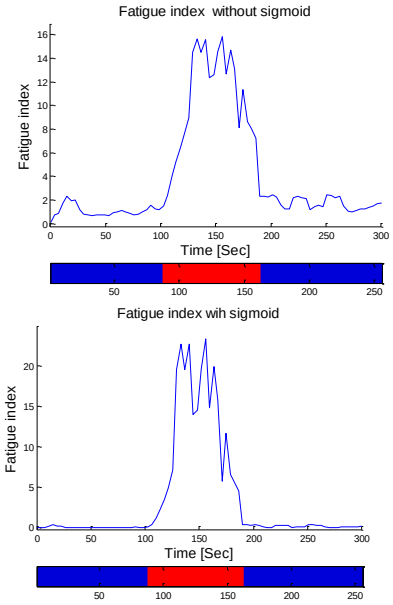
$$Ei = \sum_{j=1}^{M} |Y_{ij}|^2 \quad (4)$$

in which, E_i is the calculated energy of the i_{th} block, Y_{ij} is the Fast Fourier Transform (FFT) of the trajectories and M the is length of Y. Finally, fatigue occurrence is found by:

$$F_i = k \frac{E_i}{\frac{1}{N}\sum_{j=1}^{N} E_j} tanh(\gamma (\frac{E_i}{\frac{1}{N}\sum_{j=1}^{N} E_j} - 1)) \quad (5)$$

in which, F_i is the fatigue index, E_i is the calculated energy of the i_{th} block, N is the number of the initial blocks in the normal case (before starting the fatigue), K is the amplitude factor, and γ is a slope factor.

It can be seen that in Eq. (5) a bipolar sigmoid (tangent hyperbolic) has been applied to E_i (the calculated energy). This actually suppresses the fake peaks that appear in the results because of the facial expression and the volunteer motions. Figure 4.3 illustrates the effect of the sigmoid function on the output results. Experimentally, we got reasonable results with k = 10 and γ = 0,01.



a b

Figure 4.3: Comparing the output without using sigmoid a. and with using sigmoid b. The blue regions are the resting time and the red region is the fatigue due to exercise.

4.4. **Experimental REsults**

The proposed system has been implemented in Matlab R2013a. The test subjects participating in evaluation of the system were filmed by two webcams: one was filming the frontal views of the face and the other one filming the full body of the test subjects, for manual verification of the synchronized shaking of arms and faces of the test subjects during the experiments.

There were 20 persons involved in the testing, 14 in one testing scenario and six in another testing scenario. The two testing scenarios evaluated the system in two different fatigue detection exercises known as maximal muscle activity and submaximal muscles activity [21]. These tests which are explained in the following scenarios are usually used in detecting muscle fatigue [4].

4.4.1. Testing scenario 1 (maximal muscle activity)

In this scenario we have considered the proposed algorithm's accuracy in maximal muscle activity. For validating our results, we utilized a hand grip dynamometer to produce ground truth data by analyzing the recorded data. We asked the test subjects to squeeze the device as much as they can while they were looking at the webcam which was filming their faces. Next, the data obtained by the proposed system was compared against the one recorded by the dynamometer. Table 1 compares the duration of the fatigue detected by the proposed system and the dynamometer.

Figure 4.4.a graphically depicts the amount of the energy which is released in each time block during for one of the test subjects. Vertical axis of this figure shows the released energy in the frequency domain. According to the figure, the regions with blue color correspond to the durations that subjects are resting or doing exercise without fatigue. During the fatigue, depending on the level of subject's fatigue, the color is respectively changing to light blue, yellow, orange, and finally red. This figure actually shows the distribution of the Energy Spectral Density (ESD) in the frequency domain in the interval of [3-5] Hz, where the colors on the map depict the locations of strong variation components of $V_s(n)$.

Using this information, not only we can approximate the fatigue, but we can see which components carry most of the shaking energy due to the fatigue. However, for detecting the fatigue it is better to use a line graph like the one shown in figure 4.4.b wherein the boundaries between the fatigue and the rest areas are clearly visible. Fatigue in this figure happens when the fatigue index sharply goes beyond the threshold. It seems that the ideal value of the threshold is zero, but based on our experience it

should be a bit larger than zero to remove fake peaks due to unwanted motions. However, selecting larger thresholds decrease the accuracy of the fatigue duration. Figure 4.4.c shows the recorded data using the dynamometer. The part of the graph with a falling force indicates the fatigue region. It can be seen from these figures and also from table 4.1 that there is a good agreement between the results of the proposed system and the ground truth.

Table 4.1: Comparing the fatigue durations obtained by the proposed system against those obtained by the dynamometer in Testing Scenario 1.

Subjects	Ground truth by the dynamometer (sec.)	Fatigue duration by the proposed system (Sec.)
Subject 1	103.7 – 170.0	100.1 – 130.8
Subject 2	106.2 – 164.3	106.4 – 167.2
Subject 3	116.1 – 199.2	115.5 – 188.6
Subject 4	100.2 – 159.2	102.6 – 163.5
Subject 5	100.9 – 196.2	100.0 – 188.5
Subject 6	100.2 – 169.0	100.1 – 157.9
Subject 7	102.3 – 151.6	103.9 – 130.8
Subject 8	102.1 – 206.1	103.9 – 215.5
Subject 9	101.8 – 211.0	103.9 – 211.6
Subject 10	101.5 – 207.0	100.2 – 215.7
Subject 11	107.8 – 224.6	107.7 – 223.2
Subject 12	101.6 – 184.2	98.84 – 178.7
Subject 13	102.0 – 180.0	103.9 – 180.9

Subject 14	97.81 - 189.0	98.76 - 189.9

4.4.2. Testing scenario 2 (submaximal muscles activity)

The purpose of this test is to examine the proposed approach for another type of isometric exercise, the submaximal muscles activity. In this type of activity, the muscles in contrary to using the dynamometer are not required to exert the maxi-mum force, but they get tired by continuing an exercise. We implemented this sce-nario by holding a dumbbell.

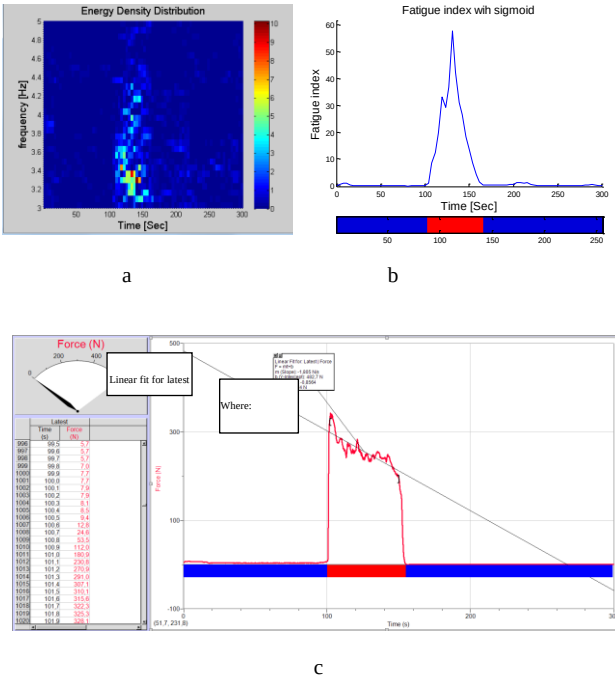


Figure 4.4: Testing scenario 1: a. Fatigue time spectral map, b. the same information by a line graph, and c. the ground truth data measured obtained by the dynamometer for one of the test subjects. The blue region is the resting time and the red region shows the fatigue due to exercise.

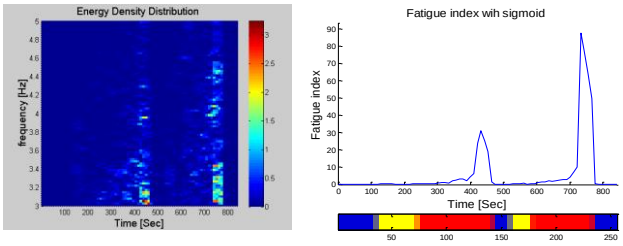
Six test subjects participating in this testing scenario were asked to look at the webcam for a while without any motion or expression. They were then asked to left a 5KG dumbbell slowly without any fast motion or reaction on their face, and hold the dumbbell as long as possible such that they feel a continuous pain on their shoulders (fatigue), then they rest for around one or two minutes (depending on their tiredness). Finally, they were asked to repeat the lifting weight again, similar to the first time. Table 4.2 shows the fatigue duration for each participant. Figures 4.5.a and

4.5.b show the output of the proposed system by the time spectral map and the line graph for one of the test subjects.

Similar to figure 4.4.b it can be seen in figure 4.5.b that there is a clear difference between the resting and the fatigue regions. However, it can be seen that when the exercise starts in the testing scenario 1, the fatigue index increases sharply, while it rises smoothly in the testing scenario 2. The difference is due to the different types of exercises in the two scenarios. In the testing scenario 1, we used maximal muscle activity while in the testing scenario 2 we used submaximal activity, which does not require maximum power to lift the dumbbell at the beginning.

Table 4.2: Fatigue duration for lifting the dumbbell at the first and the second attempts.

Subjects	First Attempt			Second Attempt		
	Lift up (Sec.)	Lift down (Sec.)	Fatigue Duration (Sec.)	Lift up (Sec.)	Lift Down (Sec.)	Fatigue Duration (Sec.)
Subject 1	120	220	200.4 - 223.5	300	480	339.1 - 493.1
Subject 2	130	310	167.5 - 293.1	430	550	418.8 - 544.4
Subject 3	125	380	146.9 - 384.2	496	742	508.5 - 757.1
Subject 4	123	379	172.4 - 381.2	476	618	535.4 - 599.0
Subject 5	125	280	230.1 - 283.8	392	502	452.5 - 506.2
Subject 6	128	420	226.5 - 431.4	520	720	539.2 - 733.4



a b

Figure 4.5: Testing scenario 2: a. Fatigue time spectral map, b. the same information by a line graph for test subject 5. Blue, yellow and red regions indicate resting time, exercise without fatigue and exercise with fatigue, respectively.

4.5. Conclusion

Muscle fatigue is nowadays measured by some sensors that need to be in direct contact with muscles. The proposed system in this paper consists in a novel contactless muscle fatigue measurement algorithm by detecting and tracking facial features. The proposed system has been tested on 20 test subjects in two different testing scenarios. Comparing the results of the proposed system against the results obtained by contact-based sensors shows that our system finds the fatigue indexes (thresholds between resting and fatigue areas) properly.

4.6. References

[1] A. J. Dittner, S. C. Wessely and R .G. Brown," The assessment of fatigue: A practical guide for clinicians and researchers." J. Psychosom. Res., vol. 56, pp. 157 – 170, 2004.

[2] T. Chalder, G. Berelowitz, T. Pawlikowska, L. Watts, S. Wessely, D. Wright and E. P. Wallacel, "Development of a fatigue scale." J. Psychosom. Res., Vol. 37, pp. 147-153, 1993.

[3] B. Maton "Human motor unit activity during the onset of muscle fatigue in submaximal isometric isotonic contractions." Eur J Appl Physiol, vol. 46, pp. 271-281, 1981.

[4] W. D. McArdle, F. I. Katch and V. L. Katch, Essentials of Exercise Physiology, 4th ed. Baltimore: Lippincott Williams & Wilkins, 2011.

[5] M. B. I. Reaz, M. S. Hussain and F. Mohd-Yasin, "Techniques of EMG Signal Snalysis: Detection, Processing, Classification and Applications." Biol. Proced. Online, vol. 8, pp. 11–35, 2006.

[6] M. R. Al-Mulla, F. Sepulveda, M. Colley, "A review of non-invasive techniques to detect and predict localised muscle fatigue." Sensors, vol. 11, pp. 3545–3594, 2011.

[7] M. González-Izal, A. Malanda, I. Navarro-Amézqueta, E. M. Gorostiaga, F. Mallor, J. Ibañez and M. Izquierdo, "EMG spectral indices and muscle power fatigue during dynamic contractions." J. Electromyograph. Kinesiol., vol. 20, pp. 233–240,2010.

[8] N. S. Stoykov, M. Lowery, and A. Kuiken, "A finite-element analysis of the effect of muscle insulation and shielding on the surface EMG signal." IEEE Trans. Biomed. Eng., vol. 52, pp. 117–121, 2005.

[9] K. Masuda, T. Masuda, T. Sadoyama, M. Inaki and S. Katsuta, "Changes in surface EMG parameters during static and dynamic fatiguing contractions." J. Electromyograph. Kinesiol., vol.9, pp. 39–46, 1999.

[10] D. Farina, P. Madeleine, T. Graven-Nielsen, R. Merletti and L. Arendt-Nielsen, "Standardizing surface electromyogram recordings for assessment of activity and fatigue in the human upper trapezius muscle." Eur. J. Appl. Physiol., vol. 86, pp. 469 –478, 2002.

[11] G. R. Naik, D. K. Kumar, K. Wheeler and S. P. Arjunan, "Estimation of muscle fatigue during cyclic contractions using source separation techniques," In Proc. of the 2009 digital image computing: Techniques and applications (DICTA 2009), Melbourne, 2009, pp. 217–222.

[12] C. Orizio, M. Gobbo, B. Diemont, F. Esposito and A. Veicsteinas, "The surface mechanomyogram as a tool to describe the influence of fatigue on biceps brachii motor unit activation strategy. Historical basis and novel evidence." Eur. J. Appl. Physiol., vol. 90, pp. 326–336. 2003.

[13] M. Stokes and M. Blythe, Muscle Sounds in Physiology, Sports Science, and Clinical Investigation: Applications and History of Mechanomyography, Oxford: Medintel, 2001.

[14] C. Orizio, "Muscle sound: Bases for the introduction of a mechanomyographic signal in muscle studies." Crit. Rev. Biomed. Eng., vol. 21, pp. 201–243, 1993.

[15] R. Irani, K. Nasrollahi, B. Moeslund, "Improved Pulse Detection from Head Motions Using DCT," in 9th International Conference on Computer Vision Theory and Applications (VISAPP) , Lisbon, 2014.

[16] G. Balakrishnan F. Durand, and J. Guttag, "Detecting Pulse from Head Motions in Video." In IEEE Computer Vision and Pattern Recognition (CVPR), Portland, 2013, pp. 3430 – 3437.

[17] M. Poh, D. J. McDuff and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." Optics Express, vol. 18, pp. 10762–10774, 2010.

[18] H. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world." ACM Trans. Graph., vol. 31, pp. 65:1-65:8, 2012.

[19] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features," in IEEE Computer Vision and Pattern Recognition (CVPR), Kauai, 2001, pp. 511–518.

[20] J. Y. Bouguet, "Pyramidal Implementation of the Lucas-Kanade Feature Tracker," Tech. Rep., Intel Corporation, Microprocessor Research Labs, 2000.

[21] R. M. Enoka, J. Duchateau, "Muscle fatigue: what, why and how it influences muscle function." J. Physiol., vol. 586, pp. 11 –23, 2008.

Chapter 5

Facial Video Based Detection of Physical Fatigue for Maximal Muscle Activity

Mohammad Ahsanul Haque, Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund

This paper has been published in

IET Computer Vision, vol. 10, no. 4, pp. 323 – 329, 2016.

The layout has been revised.

5.1. **Abstract**

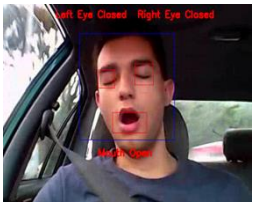
Physical fatigue reveals the health condition of a person at for example health check-up, fitness assessment or rehabilitation training. This chapter presents an efficient noncontact system for detecting non-localized physical fatigue from maximal muscle activity using facial videos acquired in a realistic environment with natural lighting where subjects were allowed to voluntarily move their head, change their facial expression, and vary their pose. The proposed method utilizes a facial feature point tracking method by combining a 'Good feature to track' and a 'Supervised descent method' to address the challenges originates from realistic scenario. A face quality assessment system was also incorporated in the proposed system to reduce erroneous results by discarding low quality faces that occurred in a video sequence due to problems in realistic lighting, head motion and pose variation. Experimental results show that the proposed system outperforms video based existing system for physical fatigue detection.

5.2. **Introduction**

Fatigue is an important physiological parameter that usually describes the overall feeling of tiredness or weakness in human body. Fatigue may be either mental or physical or both [1]. Mental fatigue is a state of cortical deactivation due to prolonged periods of cognitive activity that reduces mental performance. On the other hand, physical fatigue refers to the declination of the ability of muscles to generate force. Stress, for example, makes people mentally exhausted, while hard work or extended physical exercise can exhaust people physically. Though mental fatigue is related to cognitive activity, it can occur during a physical activity that comprises neurological phenomenon, for example directed attention as found in the area of intelligent transportation systems [2]. Unlike mental fatigue that is related to cognitive performance, physical fatigue specifically refers to muscles' inability to force optimally due to inadequate rest during a muscle activity [1]. Physical fatigue occurs from two types of activities: submaximal muscle activity (e.g. using a cycle ergometer or motor driven treadmill) and maximal muscle activity (e.g. pressing a dynamometer or lifting a load with great force) [3]. This kind of fatigue is a significant physiological parameter, especially for athletes or therapists. For example, by monitoring the occurrence of a patient's fatigue during physical exercise in rehabilitation scenarios, a therapist can change the exercise, make it easier or even stop it if necessary. Estimating the fatigue time offsets can also provide information in posterity health analysis [1].

A number of video-based non-invasive methods for fatigue detection and quantification have been proposed in the literature. The methods utilized some features and clues, as shown in Figure 5.1.a, automatically extracted from a subject's facial video and discriminate between fatigue and non-fatigue classes automatically. For example, the works in [4] use eye blink rate and duration of eye closure for detection of fatigue occurrence due to sleep deprivation or directed attention. In addition to these

features head pose and yawning behavior in facial video is used in [5]. A review of facial video based fatigue detection methods can be found in [2]. However, all of these methods address the detection of mental fatigue occurred from prolonged directed attention activity or sleep deprivation, more specifically known as driver fatigue, instead of physical fatigue occurred from maximal or sub-maximal muscle activity, as shown in Figure 5.1.b, c. Most of the available technologies for detecting physical fatigue occurrence (in terms of fatigue time offsets and/or fatigue level) use contact-based sensors such as force gauge, Electromyogram (EMG), and Mechanomyogram (MMG). Force gauge is minimally invasive, but it requires a device like a hand grip dynamometer [6]. EMG uses electrodes which requires wearing adhesive gel patches [7]. MMG is based on an accelerometer or goniometer that requires direct skin contact and is sensitive to noise [8].



a. Driver's mental fatigue experiment (image taken from [5])

b. Physical fatigue experiment using dumbbell

c. Physical fatigue experiment using hand dynamometer

Figure 5.1: Analyzing facial video for different fatigue scenarios.

To the best of our knowledge, the only video-based non-invasive system for non-localized (i.e., not restricted to a particular muscle) physical fatigue detection in a maximal muscle activity scenario (as shown in Figure 5.1.c) is the one introduced in [9] which uses head-motion (shaking) behavior due to fatigue in video captured by a simple webcam. It takes into account the fact that muscles start shaking when fatiguing contraction occurs in order to send extra sensation signal to the brain to get enough force in a muscle activity and this shaking is reflected in the face. Inspired by [10] for heartbeat detection from Ballistocardiogram, in [9] some feature points on the ROI (forehead and cheek) of the subject's face in a video are selected and tracked to generate trajectories of the facial feature points, and to calculate the energy of the vibration signal, which is used for measuring the onset and offset of fatigue occurrence in a non-localized notion. Though both physical fatigue and mental fatigue can occur simultaneously during a physical activity, the physiological mechanisms are not same. While mental fatigue represents the temporary reduction of cognitive performance, physical fatigue represent temporary reduction of force induced in muscle to accomplish a physical activity [2]. Unlike driver mental fatigue, physical fatigue for maximal muscle activity does not necessarily require a prolonged period. Thus,

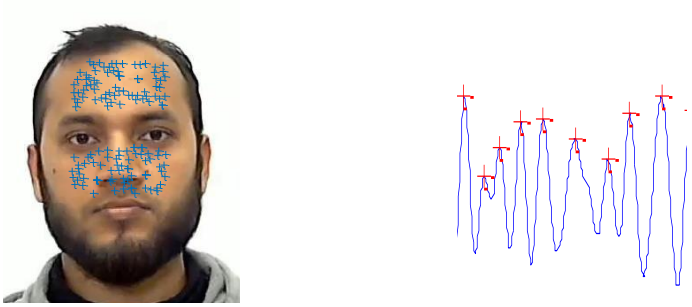
the visual clues found in the case of driver fatigue cannot be found in the case of physical fatigue from maximal muscle contraction. Changes in facial features in these two different types of fatigue are very different: in the driver mental fatigue eye blinking, yawning, varying head pose and degree of mouth openness are used (as shown in Figure 5.1.a), while in the non-localized maximal muscle contraction based physical fatigue *head motion behavior from shaking* is used. Consequently, physical fatigue occurred from maximal muscle activity cannot be detected or quantized by the computer vision methods used for detecting driver mental fatigue.

The previous facial video-based method in [9] for non-localized physical fatigue detection extracts some facial feature points as shown in Figure 5.2.a. Depending upon imaging scenario, the number of feature points and their position can vary. The method then employs signal processing techniques to detect head motion trajectories from feature points in the video frames and estimates energy escalation to detect fatiguing contraction. However, the work in [9] assumes that there is neither internal facial motion, nor external movement or rotation of the head during the data acquisition phase. We denote internal motion as facial expression and external motion as head pose. In real life scenarios there are, of course, both internal and external head motion. The current method, therefore, fails due to an inability to detect and track the feature points in the presence of internal and external motion, and low texture in the facial region. Moreover, real-life scenarios challenge current methods due to low facial quality in video because of motion blur, bad posing, and poor lighting conditions [11]. The proposed system in this chapter extends [9] by addressing the abovementioned shortcomings and thereby allows for automatic and more reliable detection of fatigue time offsets from facial video captured by a simple webcam. To address the shortcomings, we introduce a Face Quality Assessment (FQA) method that prunes the captured video data so that low quality face frames cannot contribute to erroneous results [12], [13]. Following [10], [14], we track feature points (Figure 5.2(a)) through a method Good Feature to Track (GFT) with Kanady-Lucas-Tomasi (KLT) tracker, and then combine these trajectories with 49 facial landmark trajectories (Figure 5.2(b)), tracked by a Supervised Descent Method (SDM) of [15], [16]. The idea of combining these two types of features has been developed in our paper [17], which was applied to heartbeat estimation from facial video. Here we look at another application of this idea for physical fatigue estimation. The experiments are conducted on realistic datasets collected at the lab and a commercial fitness center for fatigue measurement. The chapter's contributions are as follows:

- We identify the limitations of the GFT-based tracking used in previous methods for physical fatigue detection and propose a solution using SDM-based tracking.
- We provide evidence for the necessity of combining the trajectories from the GFT and the SDM, instead of using the trajectories from either the GFT or the SDM.

- We introduce the notion of FQA in the physical fatigue detection context and demonstrate empirical evidence for its effectiveness.

The rest of the chapter is organized as follows. Section 3 describes the proposed method. The results are summarized in section 4. Finally, section 5 concludes the chapter.



a b

Figure 5.2: a. Facial feature points (total numbers can vary) in a face obtained by GFT-based tracking [9], b. 49 facial landmarks in a face obtained by SDM-based tracking [15].

5.3. The Proposed Method

The block diagram of the proposed method is shown in Figure 5.3. The steps are explained in the following subsections.

5.3.1. Face Detection and Face Quality Assessment

The first step of the proposed motion-based physical fatigue detection system is face detection from facial video, which has been accomplished by Viola and Jones object detection framework using Haar-like features obtained from integral images [18].

Facial videos captured in real-life scenarios can be subject to the problems of pose variation, varying levels of brightness, and motion blur. When the intensity of these problems increases, the face quality decreases. A low quality face produces erroneous results in detecting facial features using either GFT or SDM [11]. To solve this problem, we pass the detected face to a FQA module. The FQA module assesses the quality of the face in the video frames. As investigated in [17], [19], four quality metrics can be critical for facial geometry analysis and detection of landmarks: resolution, brightness, sharpness, and out-of-plan face rotation (pose). Thus, the low quality faces can be discarded by calculating these quality metrics and employing thresholds to check whether the face needs to be discarded. The formulae to obtain these quality scores from a face are listed in [11]. Resolution score is calculated in terms of number of pixels, pose score is calculated by detecting the center of mass in the binary image

of the face region, sharpness is calculated by employing a low-pass filter to detect motion blur or unfocused capture, and brightness is calculated from the average of the illumination component of all the pixels in the face region. When we obtain the quality scores, following [17], we discard the low quality faces by the thresholds as follows: face resolution- 150x150, brightness- 0.80, sharpness- 0.80, and pose- 0.20 by following [11]. As we detect the fatigue time offsets over a long video sequence (e.g. 30 secs to 180 secs) for maximum muscle activity, discarding few frames (e.g. less than 5% of the total frames) does not affect much the regular characteristic of the trajectories, but removes the most erroneous segments coming from low quality faces. In fact, no frames are discarded if the quality score is not less than the thresholds. Missing points in the trajectory are removed by concatenating trajectory segments. The effect of employing FQA will be illustrated in the experimental evaluation section.

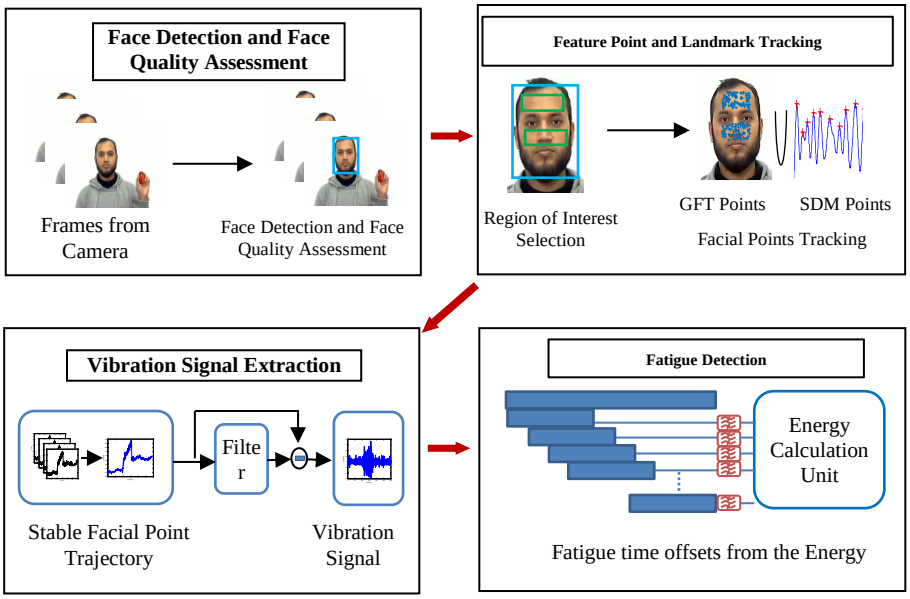


Figure 5.3: The block diagram of the proposed system.

5.3.2. Feature Points and Landmarks Tracking

As mentioned earlier, muscles start shaking when a subject becomes tired physically (the occurrence of physical fatigue from maximal muscle activity) [20]. The energy dispersed from this shaking is distinctively intense then the other types of head motion and is reflected in the face. Thus, physical fatigue can be determined from head motion by estimating the released shaking energy. Tracking facial feature points and generating trajectory help to record the head motion in facial video. This task was accomplished in [9] using merely a GFT-based method (utilizes KLT tracker). In

order to detect and track facial feature points in consecutive video frames, the GFT-based method uses an affine motion model to express changes in the intensity level in the face. It defines the similarity between two points in two frames using a so called 'neighborhood sense' or window of pixels. Tracking a window of size $w_x \times w_y$ in the frame I to the frame J is defined on a point velocity parameter $\boldsymbol{\delta} = [\delta_x \ \delta_y]^T$ for minimizing a residual function f_{GFT} as follows:

$$f_{GFT}(\boldsymbol{\delta}) = \sum_{x=p_x}^{p_x+w_x} \sum_{y=p_y}^{p_y+w_y} \big(I(\boldsymbol{x}) - J(\boldsymbol{x}+\boldsymbol{\delta})\big)^2 \quad (1)$$

where $(I(\mathbf{x}) - J(\mathbf{x}+\boldsymbol{\delta}))$ stands for $(I(x,y) - J(x+\delta_x, y+\delta_y))$, and $\mathbf{p} = [p_x, p_y]^T$ is a point to track from the first frame to the second frame. According to the observations in [14], the quality of the estimate by this tracker depends on three factors: the size of the window, the texture of the image frame, and the amount of motion between frames. The GFT-based fatigue detection method assumes that the head does not have voluntary head motion during data capture. However, voluntary head motion (both external and internal) and low texture in facial videos are usual in real life scenarios. Thus, the GFT-based tracking of facial feature points exhibits four problems. **First problem** arises due to low texture in the tracking window. This difficulty can be overcome by tracking feature points in corners or regions with high spatial frequency content, instead of forehead and cheek. **Second problem** arises by losing track in long video sequences due to point drifting in long video sequences. **Third problem** occurs in selecting an appropriate window size (i.e. $w_x \times w_y$ in (1)). If the window size is small, a deformation matrix to find the track is harder to estimate because the variations of motion within it are smaller and therefore less reliable. On the other hand, a bigger window is more likely to straddle a depth discontinuity in subsequent frames. **Fourth problem** comes when there is large optical flow in consecutive video frames. When there is voluntary motion or expression change in a face, the optical flow or face velocity in consecutive video frames is very high and GFT-based method misses the track due to occlusion [21]. Higher video frame rate may able to address this problem, however this will require specialized camera instead of simple webcam. Due to these four problems, the GFT-based trajectory for fatigue detection leads to erroneous result in realistic scenarios where lighting changes and voluntary head motions exist.

A viable way to enable the GFT-based systems to detect physical fatigue in a realistic scenario is to track the facial landmarks by employing a face alignment system. Face alignment is considered as a mathematical optimization problem and a number of methods have been proposed to solve this problem. The Active Appearance Model (AAM) fitting [22] and its derivatives [23] were some of the early solutions in this area. The AAM fitting works by estimating parameters of an artificial model that is sufficiently close to the given image. In order to do that AAM fitting was formulated as a Lukas-Kanade (LK) problem [24], which could be solved using Gauss-Newton optimization [25]. A fast and effective solution to this was proposed recently in [15],

which develops a Supervised Descent Method (SDM) to minimize a non-linear least square function for face alignment. The SDM first uses a set of manually aligned faces as training samples to learn a mean face shape. This mean shape is then used as an initial point for an iterative minimization of a non-linear least square function towards the best estimates of the positions of the landmarks in facial test images. The minimization function can be defined as a function over Δx as:

$$f_{SDM}(x_0 + \Delta x) = \left\| g\big(d(x_0 + \Delta x)\big) - \theta_* \right\|_2^2 \quad (2)$$

where x_0 is the initial configuration of the landmarks in a facial image, $d(x)$ indexes the landmarks configuration (x) in the image, g is a nonlinear feature extractor, $\theta_* = g(d(x_*))$, and x_* is the configuration of the true landmarks. The Scale Invariant Feature Transform (SIFT) [11] is used as the feature extractor g. In the training images Δx and θ_* are known. By utilizing these known parameters the SDM iteratively learns a sequence of generic descent directions, $\{\partial_n\}$, and a sequence of bias terms, $\{\beta_n\}$, to set the direction towards the true landmarks configuration x_* in the minimization process, which are further applied in the alignment of unlabelled faces [15]. This working procedure of SDM in turns addresses the four previously mentioned problems of the GFT-based approach for head motion trajectory extraction by as follows. **First**, the 49 facial landmark point tracked by SDM are taken only around eye, lip, and nose edges and corners, as shown in Figure 5.2.b. As these landmarks around the face patches have high spatial frequency and do not suffer from low texturedness, this eventually solves the problem of low texturedness. We cannot simply add these landmarks in the GFT based tracking, because the GFT based method has its own feature point selector. **Second**, SDM does not use any reference points in tracking. Instead, it detects each point around the edges and corners in the facial region of each video frame by using supervised descent directions and bias terms. Thus, the problems of point drifting do not occur in long videos. **Third**, SDM utilizes the 'neighborhood sense' on a pixel-by-pixel basis instead of a window. Therefore, window size is not relevant to SDM. **Fourth**, the use of supervised descent direction and bias terms allows the SDM to search selectively in a wider space and look after it from large optical flow problem. Thus, large optical flow cannot create occlusion in the SDM-based approach.

As in realistic scenarios the subjects are allowed to have voluntary head motion and facial expression change in addition to the natural cyclic motion, the GFT-based method results to either of the two consequences for videos having challenging scenarios: i) completely missing the track of feature points and ii) erroneous tracking. We observed more than 80% loss of feature points by the system in such cases. The GFT-based method, in fact, fails to preserve enough information to estimate fatigue from trajectories even though the video have minor expression change or head motion voluntarily. On the other hand, the SDM does not miss or erroneously track the land-

marks in the presence of voluntary facial motions or expression change or low texturedness as long as the face is qualified by the FQA. Thus, the system can find enough trajectories to detect fatigue. However, the GFT-based method uses a large number of facial points to track when compared to SDM. This matter causes the GFT-based method to generate a better trajectory than SDM when there is no voluntary motion or low texturedness. Following the above discussions Table 5.1 summarizes the behavior of GFT, SDM and a combination of these two methods in facial point tracking. We observe that a combination of trajectories obtained by GFT and SDM-based methods can produce better results in cases where subjects may have both motion and non-motion periods. We thus propose to combine the trajectories. In order to generate combined trajectories, the face is passed to the GFT-based tracker to generate trajectories from facial feature points and then appended with the SDM trajectories.

5.3.3. Vibration Signal Extraction

To obtain vibration signal for fatigue detection, we take the average of all the trajectories obtained from both feature and landmark points of a video by as follows:

$$T(n) = \frac{1}{M}\sum_{m=1}^{M}(y_m(n) - \bar{y}_m) \tag{3}$$

where $T(n)$ is the shifted mean filtered trajectory, $y_m(n)$ is the n-th frame of the trajectory m, M is the number of the trajectories, N is the number of the frames in each trajectory, and $\bar{y}_m$ is the mean value of the trajectory m given by:

$$\bar{y}_m = \frac{1}{N}\sum_{n=1}^{N} y_m(n) \tag{4}$$

The vibration signal that keeps the shaking information is calculated from T by using a window of size R by:

$$V_s(n) = T(n) - \frac{1}{R}\sum_{r=0}^{R-1} T(n-r) \tag{5}$$

The obtained signal is then passed to the fatigue detection block.

Table 5.1: Behaviour of the GFT, SDM and a combination of both methods for facial points tracking in different scenarios

Scenario	Challenge	GFT	SDM	Combination

Low texture in video	Number of facial points available to effectively generating motion trajectory	*Bad*	*Good*	*Better*
Long video sequence	Facial point drifting during tracking	*Bad*	*Good*	*Better*
Appearance of voluntary head motion in video	Optical occlusion and depth discontinuity of window based tracking	*Bad*	*Good*	*Better*
Perfect scenario	None of the aforementioned challenges	*Good*	*Good*	*Better*

5.3.4. Physical Fatigue Detection

To detect the released energy of the muscles reflected in head shaking we need to segment the vibration signal V_s from (5) with an interval of Δt_{sec}. Segmenting the signal V_s helps detecting the fatigue in temporal dimension. After windowing, each block is filtered by a passband ideal filter. Figure 5.4.a shows the power of the filtered vibrating signal with a cut-off frequency interval of *[3-5]* Hz. The cutoff frequency was determined empirically in [9]. We observe that the power of the signal rises when fatigue happens in the interval of *[16.3–40.6] seconds* in this figure. After filtering, the energy of i_{th} block is calculated as:

$$E_i = \sum_{j=1}^{M} |Y_{ij}|^2 \tag{6}$$

where E_i is the calculated energy of the i-th block, Y_{ij} is the FFT of the signal V_s, and M is the length of Y. Finally, fatigue occurrence is detected by:

$$F_i = k \frac{E_i}{\frac{1}{N}\sum_{j=1}^{N} E_j} tanh(\gamma(\frac{E_i}{\frac{1}{N}\sum_{j=1}^{N} E_j} - 1)) \tag{7}$$

where F_i is the fatigue index, N is the number of the initial blocks in the normal case (before starting the fatigue), K is the amplitude factor, and γ is a slope factor. Experimentally, we obtained reasonable results with k = 10 and γ = 0.01. As observed in [9], employing a bipolar sigmoid (tangent hyperbolic) function to E_i in (7) suppresses the noise peaks out of fatigue region that appear in the results because of the facial expression and/or the voluntary motion. Figure 5.4.b, c illustrates the effect of the sigmoid function on the output results and Figure 5.4.d, e depicts the effect in values.

To realize the effect of such noise suppression in percentage, by following [26] we use the following metric:

$$SUP_i = \frac{F_i}{F_{max}} \times 100\% \quad (8)$$

where SUP_i is the ratio of the noise to the released fatigue energy. If we employ (8) on Figure 5.4.d, e, we obtain values 8.94%, 11.65% and 0.77%, 1.38%, respectively, for the noise datatips shown in the figures. It can be noticed that before employing the suppression the noise to fatigue energy were ~10%, however reduced to ~1% after employing the suppression. When we obtain the fatigue index, the starting and ending time of fatigue occurrence in a subject's video are detected by employing a threshold with value: 1.0 to the normalized fatigue index, as the bipolar sigmoid suppresses the signal energy out of fatigue region to less than 1.0 by (7). Fatigue starts when the fatigue index exceeds the threshold upward and fatigue ends when fatigue index exceeds the threshold downward.

5.4. Experimental Results

5.4.1. Experimental Environment

The proposed method was implemented using a combination of Matlab (2013a) and C++ environments. We integrated the SDM [15] with the GFT-based tracker from [9], [11] to develop the system as explained in the methodology section. We collected four experimental video databases to generate results: a database for demonstrating the effect of FQA, a database with voluntary motions in some moments for evaluating the performance of GFT, SDM and the combination of GFT and SDM, a database collected from the subjects in a natural laboratory environment, and a database collected from the subjects at a real-life environment in a commercial fitness center. We named the databases as "FQA_Fatigue_Data", "Eval_Fatigue_Data" "Lab_Fatigue_Data" and "FC_Fatigue_Data" respectively. All the video clips were captured in VGA resolution using a Logitech C310 webcam. The videos were collected from 16 subjects (including both male and female from different ethnicities with the ages between 25 to 40 years) after adequately informing the subjects about the concepts of maximal muscle fatigue and the experimental scenarios. Subjects exposed their face in front of the cameras while performing maximal muscle activity by using a handgrip dynamometer for about 30-180 seconds (varies from subject to subject). Subjects were free to have natural head motion and expression variation due to activity prompted by using the dynamometer. Both setups (in the laboratory and in the fitness center) used indoor lighting for video capturing and the dynamometer reading to measure ground truth for fatigue. The FQA_Fatigue_Data has 12 videos, each of which contains some low quality face in some moments. The Eval_Fatigue_Data has 17 videos with voluntary motion, Lab_Fatigue_Data has 54 videos and the FC_Fatigue_Data has 11 videos in natural scenario.

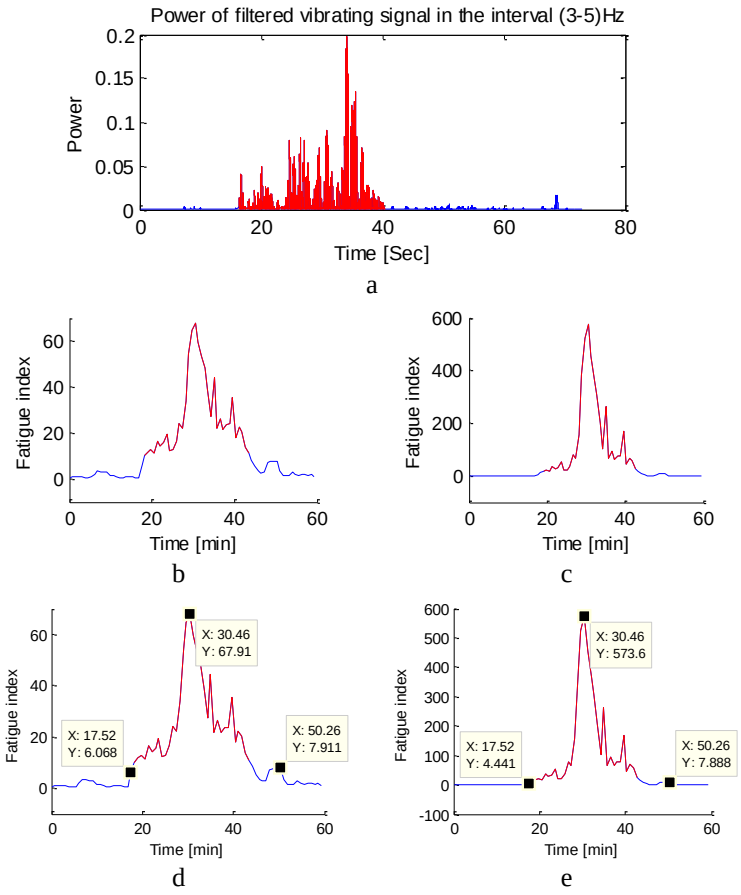


Figure 5.4: Analyzing trajectory for fatigue detection: a. The power of the trajectory where the blue region is the resting time and the red region shows the fatigue due to exercise in the interval (16.3–40.6) seconds, b. and c. before and after using a bipolar sigmoid function to suppress the noise peaks, respectively, and d. and e. depicts the effect of bipolar sigmoid in values corresponding to b. and c., respectively.

As physical fatigue in a video clip occurs between a starting time and an ending time, the starting and ending times detected from the video by the experimental methods should match with the starting and ending times of fatigue obtained from the ground truth dynamometer data. Thus, we analyzed and measured the error between the ground truth and the output of the experimental methods for starting and ending time agreement by defining a parameter μ. This parameter expresses the average of the total of starting and ending point distances of fatigue occurrence for each subject in the datasets, and is calculated as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\left(\left|G_S^i - R_S^i\right| + \left|G_E^i - R_E^i\right|\right) \tag{9}$$

where, n is the number of video (subjects) in a dataset, G_S^i is the ground truth of the starting point of fatigue, G_E^i is the ground truth of the ending point of fatigue, R_S^i is the calculated starting point of fatigue, and R_E^i is the calculated ending point of fatigue.

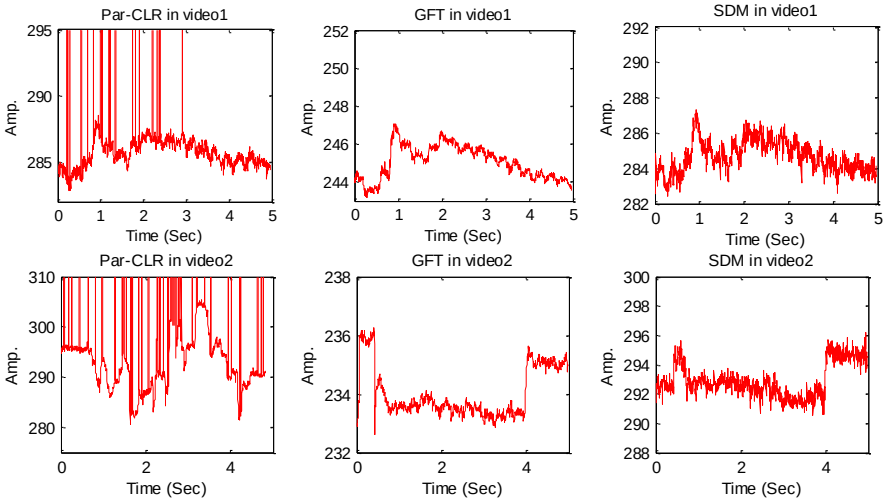


Figure 5.5: Trajectories of tracking points extracted by Par-CLR [27], GFT [14], and SDM [15] from 5 seconds of two experimental video sequences with continuous small motion (for video1 in the first row) and large motion at the beginning and end (for video2 in the second row).

5.4.2. Performance Evaluation

The proposed method used a combination of the SDM- and GFT-based approaches for trajectory generation from the facial points. Figure 5.5 shows the calculated average trajectories of tracked points in two experimental videos. We depicted the trajectories obtained from GFT-based tracker, SDM and another recent face alignment algorithm Par-CLR [27] for two facial videos with voluntary head motion. As observed from the figure, the GFT and SDM-based trackers provide similar trajectories when there is little head motion (video1, first row of Figure 5.5). On the other hand, Par-CLR provides a trajectory very different than the other two because of tracking on false positive face in the video frames. When the voluntary head motion is sizable (beginning of video2, second row of Figure 5.5), GFT-based method fails to track the point accurately and thus produces an erroneous trajectory. However, SDM provides stable trajectory in this case. Thus, lack in proper selection of method(s) for trajectory generation can contribute to erroneous results in estimating fatigue time offsets, as

we observe for the GFT-based tracker and the recently proposed Par-CLR in comparison to SDM.

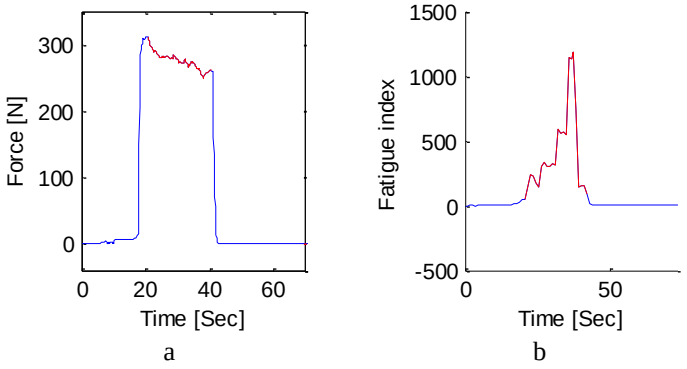


Figure 5.6: Detection of physical fatigue due to maximal muscle activity: a. dynamometer reading during fatigue event, and b. fatigue time spectral map for fatigue time offset measurement. The blue region is the resting time and the red region shows the fatigue due to exercise.

For the fatigue time offset measurement experiment we asked the test subjects to squeeze the handgrip dynamometer as much as they could. As they did this we recorded their face. The squeezed dynamometer provides a pressure force, which is used as the ground truth data in fatigue detection. Figure 5.6.a displays the data recorded while using the dynamometer, where the part of the graph with a falling force indicates the fatigue region. The measured fatigue level from the dynamometer reading is shown in Figure 5.6.b. Fatigue in this figure happens when the fatigue level sharply goes beyond a threshold defined in [9]. Comparative experimental results for fatigue detection using different methods are shown in the next section.

We conducted experiments to evaluate the effect of employing FQA, and a combination of GFT and SDM in the proposed system. Figure 5.7 shows the effect of employing FQA on a trajectory obtained from a subject's video. It is observed that low quality face region (due to pose variation) shows erroneous trajectory and contributed to the wrong detection of fatigue onset (Figure 5.7.a). When, FQA module discarded this region, the actual fatigue region was detected as shown in Figure 5.7.b. Table 5.2 shows the results of employing FQA on the FQA_Fatigue_Data, and evaluating the performance of GFT, SDM and the combination of these two on the Eval_Fatigue_Data. From the results it is observed that when videos have low quality faces (which are true for all the videos of the FQA_Fatigue_Data), automatic detection of fatigue time stumps exhibited very high error due to wrong place of detection. When we employed FQA the fatigue was detected in the expected time with minor error. While comparing GFT, SDM and the combination of these two, we observe that the SDM minimally outperformed the GFT, however the combination worked better.

These observations came with the agreement of the characteristics we listed in Table 5.1.

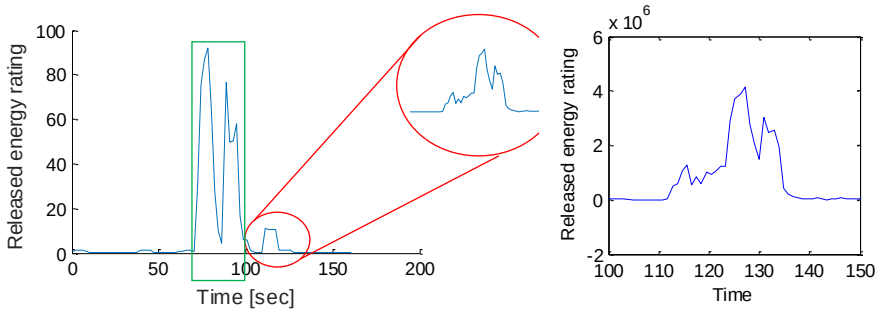


Figure 5.7: Analyzing the effect of employing FQA on a trajectory obtained from an experimental video: a. without employing FQA (red circle presents the real fatigue location and green rectangle presents the moments of low quality faces), b. employing FQA (presenting the area within the red circle of a.).

Table 5.2: Analysing the effect of the FQA and evaluating the performance of GFT, SDM, and the combination of GFT and SDM in fatigue detection

Dataset	**Scenario**	**Average of the total of the starting and ending point distance of fatigue occurrence for each subject in a dataset (μ)**
FQA_Fatigue_Data	Without FQA	65.32
	With FQA	3.79
Eval_Fatigue Data	GFT	6.81
	SDM	6.35
	Combination of GFT and SDM	5.16

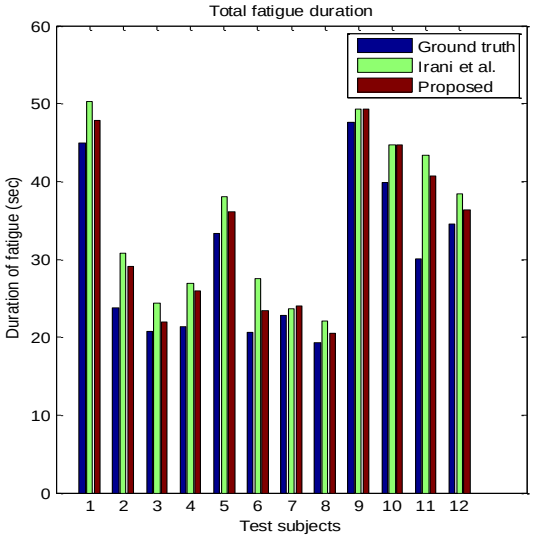
5.4.3. Performance Comparision

To the best of our knowledge, the method of [9] is the first and the only work in the literature to detect physical fatigue from facial video. Other methods for facial video based fatigue detection work for driver mental fatigue [2], and use different scenarios than what is used in physical fatigue detection environment. Thus, we have compared the performance of the proposed method merely against the method of [9] on the experimental datasets. Figure 5.8.a shows the physical fatigue detection duration for a subset of database Lab_Fatigue_Data in a bar diagram. The height of the bar shows

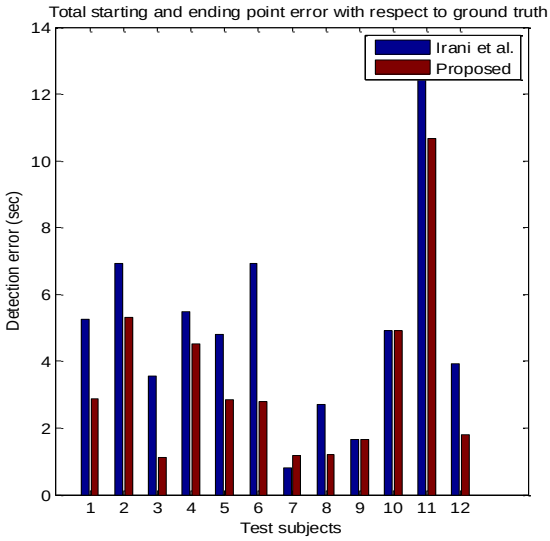
the duration of fatigue in seconds. Figure 5.8.b shows the total detection error in seconds for starting and ending points of fatigue in the videos. From the result it is observed that the proposed method detected the presence of fatigue (expressed by fatigue duration) more accurately than the previous method of [9] in comparison to the ground truth as shown in Figure 5.8.a and demonstrated better agreement with the starting and ending time of fatigue with the ground truth as shown Figure 5.8.b. Table 5.3 shows the fatigue detection results on both Lab_Fatigue_Data and FC_Fatigue_Data, and compares the performance between the state of the art method of [9] and the proposed method. While analyzing the agreement with the starting and ending time of fatigue with the ground truth, we observed that the proposed method shows more consistency than the method of [9] both in the Lab_Fatigue_Data experimental scenario and the FC_Fatigue_Data real-life scenario. However, the performance is higher for Lab_Fatigue_Data than FC_Fatigue_Data. We believe the realistic scenario of a commercial fitness center (in terms of lighting and subject's natural behavior) contributes to lower performance. The computational time of the proposed method suggest that the method is doable for real-time application, because it requires only 3.5 milliseconds (app.) processing time for each video frame in a platform with 3.3 GHz processor and 8GB RAM.

5.5. **Conclusions**

This chapter proposes a physical fatigue detecting system from facial video captured by a simple webcam. The proposed system overcomes the drawbacks of previous facial video based method of [9] by extending the application of SDM over GFT based tracking and employing FQA. The previous method works well only when there is neither voluntary motion of the face nor change of expression, and when the lighting conditions help keeping sufficient texture in the forehead and cheek. The proposed method overcomes these problems by using an alternative facial landmarks tracking system (the SDM-based system) along with the previous feature points tracking system (the GFT-based system) and provides competent results. The performance of the proposed system showed very high accuracy in proximity to the ground truth not only in a laboratory setting with controlled environment, as considered in [9], but also in a real-life environment in a fitness center where faces have some voluntary motion or expression change and lighting conditions are normal.



a.



b.

Figure 5.8: Comparison of physical fatigue detection results of the proposed method with the Irani's method [9] on a subset of the Lab_Fatigue_Data: a. total duration of fatigue, and b. total starting and ending point error in detection.

The proposed method has some limitations. The camera was placed in close proximity to the face (about one meter away) because the GFT-based feature tracker in the combined system does not work well if the face is far from the camera during video

capture. Moreover, the fatigue detection of the proposed system does not take into account the sub-maximal muscle activity due to lack of reliable ground truth data for fatigue from sub-maximal muscle activity. Future work should address these points.

Table 5.3: Performance comparison between the proposed method and the state of the art method of contact-free physical fatigue detection (in the case of maximal muscle activity) on experimental datasets

No	**Dataset name**	**Average of the total of the starting and ending point distance of fatigue occurrence for each subject in a dataset (μ)**	
		Irani et al. [9]	**The proposed method**
1.	Lab_Fatigue_Data	*7.11*	*4.59*
2.	FC_Fatigue_Data	*3.35*	*2.65*

5.6. References

[1] Y. Watanabe, B. Evengard, B. H. Natelson, L. A. Jason, and H. Kuratsune, *Fatigue Science for Human Health*. Springer Science & Business Media, 2007.

[2] M. H. Sigari, M. R. Pourshahabi, M. Soryani, and M. Fathy, "A Review on Driver Face Monitoring Systems for Fatigue and Distraction Detection," *Int. J. Adv. Sci. Technol.*, vol. 64, pp. 73–100, Mar. 2014.

[3] R. R. Baptista, E. M. Scheeren, B. R. Macintosh, and M. A. Vaz, "Low frequency fatigue at maximal and submaximal muscle contractions," *Braz. J. Med. Biol. Res.*, vol. 42, no. 4, pp. 380–385, Apr. 2009.

[4] N. Alioua, A. Amine, and M. Rziza, "Driver's Fatigue Detection Based on Yawning Extraction," *Int. J. Veh. Technol.*, vol. 2014, pp. 1–7, Aug. 2014.

[5] M. Sacco and R. A. Farrugia, "Driver fatigue monitoring system using Support Vector Machines," in *2012 5th International Symposium on Communications Control and Signal Processing (ISCCSP)*, 2012, pp. 1–5.

[6] W. D. McArdle and F. I. Katch, *Essential Exercise Physiology*, 4th edition. Philadelphia: Lippincott Williams and Wilkins, 2010.

[7] N. S. Stoykov, M. M. Lowery, and T. A. Kuiken, "A finite-element analysis of the effect of muscle insulation and shielding on the surface EMG signal," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 1, pp. 117–121, Jan. 2005.

[8] M. B. I. Raez, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications," *Biol. Proced. Online*, vol. 8, pp. 11–35, Mar. 2006.

[9] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Contactless Measurement of Muscles Fatigue by Tracking Facial Feature Points in A Video," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 1–5.

[10] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting Pulse from Head Motions in Video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3430–3437.

[11] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Quality-Aware Estimation of Facial Landmarks in Video Sequences," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015, pp. 1–8.

[12] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera," in *10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2013, pp. 443–448.

[13] J. Klonovs, M. A. Haque, V. Krueger, K. Nasrollahi, K. Andersen-Ranberg, T. B. Moeslund, and E. G. Spaich, *Distributed Computing and Monitoring Technologies for Older Patients*, 1st ed. Springer International Publishing, 2015.

[14] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[15] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.

[16] G. Tzimiropoulos and M. Pantic, "Optimization Problems for Fast AAM Fitting in-the-Wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 593–600.

[17] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Heartbeat Rate Measurement from Facial Video (accepted)," *IEEE Intell. Syst.*, Dec. 2015.

[18] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int J Comput Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[19] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, "Constructing Facial Expression Log from Video Sequences using Face Quality Assessment," in *9th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014, pp. 1–8.

[20] R. R. Young and K.-E. Hagbarth, "Physiological tremor enhanced by manoeuvres affecting the segmental stretch reflex," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 43, pp. 248–256, 1980.

[21] J. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker," *Intel Corp. Microprocess. Res. Labs*, 2000.

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

[23] A. U. Batur and M. H. Hayes, "Adaptive active appearance models," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1707–1721, Nov. 2005.

[24] S. Baker and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[25] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, "Fourier Lucas-Kanade Algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1383–1396, Jun. 2013.

[26] K. Nasrollahi, T. B. Moeslund, and M. Rahmati, "Summarization of Surveillance Video Sequences using Face Quality Assessment," *Int. J. Image Graph.*, vol. 11, no. 02, pp. 207–233, Apr. 2011.

[27] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1–8.

Chapter 6

Contact-Free Heartbeat Signal for Human Identification and Forensics

Kamal Nasrollahi, Mohammad Ahsanul Haque, Ramin Irani, and Thomas B. Moeslund

This paper has been published as a book chapter in

Handbook of Biometrics for Forensic Science. ed. / Massimo Tistarelli; Christophe Champod. Springer, pp. 289-302, 2017.

The layout has been revised.

6.1. **Abstract**

The heartbeat signal, which is one of the physiological signals, is of great importance in many real-world applications, for example, in patient monitoring and biometric recognition. The traditional methods for measuring such this signal use contact-based sensors that need to be installed on the subject's body. Though it might be possible to use touch-based sensors in applications like patient monitoring, it won't be that easy to use them in identification and forensics applications, especially if subjects are not cooperative. To deal with this problem, recently computer vision techniques have been developed for contact-free extraction of the heartbeat signal. We have recently used the contact-free measured heartbeat signal, for bio-metric recognition, and have obtained promising results, indicating the importance of these signals for biometrics recognition and also for forensics applications. The importance of heartbeat signal, its contact-based and contact-free extraction methods, and the results of its employment for identification purposes, including our very recent achievements, are reviewed in this chapter.

6.2. **Introduction**

Forensic science deals with collecting and analyzing information from a crime scene for the purpose of answering questions related to the crime in a court of law. The main goal of answering such questions is identifying criminal(s) committing the crime. Therefore, any information that can help identifying criminals can be useful. Such information can be collected from different sources. One such a source, which has a long history in forensics, is based on human biometrics, i.e., human body or behavioral characteristics that can identify a person, for example, DNA [1], fingerprints [2], [3], and facial images [4]. Besides human biometrics, there is another closely related group of human features/characteristics that cannot identify a per-son, but can help the identification process. These are known as soft biometrics, for instance, gender, weight, height, gait, race, and tattoo.

The human face, which is of interest of this chapter, is not only used as a bio-metric, but also as a source for many soft biometrics, such as gender, ethnicity, and facial marks and scars [5], [6], [8]. These soft biometrics have proven great values for forensics applications in identification scenarios in unconstrained environments wherein commercial face recognition systems are challenged by wild imaging conditions, such as off frontal face pose and occluded/covered face images [7], [8]. The mentioned facial soft-biometrics are mostly based on the physical features/characteristics of the human. In this chapter, we look into heartbeat signal which is a physiological feature/characteristic of the human that similar to the mentioned physical ones can be extracted from facial images in a contact-free way, thanks to the recent advances in computer vision algorithms. The heartbeat signal is one of the physiological signals that is generated by the cardiovascular system of the human body. The phys-

iological signals have been used for different purposes in computer vision applications. For example, in [9] electromyogram, electrocardiogram, skin conductivity and respiration changes and in [10] electrocardiogram, skin temperature, skin conductivity, and respiration have been used for emotion recognition in different scenarios. In [11] physiological signals have been used for improving communication skills of children suffering from Autism Spectrum Disorder in a virtual reality environment. In [12], [13], and [14] these signals have been used for stress monitoring. Based on the results of our recent work, which are reviewed here, the heartbeat signal shows promising results to be used as a soft biometric.

The rest of this chapter is organized as follows: first, the measurements of heartbeat signal using both contact-based and contact-free methods are explained in the next section. Then, employing these signals for identification purposes is discussed in section 4. Finally, the chapter is concluded in section 5.

6.3. **Measurement of Heartbeat Signal**

The heartbeat signal can be measured in two different ways: contact-based and contact-free. These methods are explained in the following subsections.

6.3.1. Contact- based Measurement of Heartbeat Signal

The heartbeat signal can be recorded in two different ways using the contact-based sensors:

- By monitoring electrical changes of muscles during heart functioning, by a method that is known as Electrocardiogram which records ECG signals.
- By listening to the heart sounds during its functioning, by a method that is known as Phonocardiogram which records PCG signals.

The main problem of the above mentioned methods is obviously the need for the sensors to be in contact (touch) with the subject's body. Depending on the application of the measured heartbeat signal, this requirement may have different consequences, for example, for:

- Constant monitoring of patient, having such sensors on the body may cause some skin irritation.
- Biometric recognition, the subjects might not be cooperative to wear the sensors properly.

Therefore, contact-free measurement of heartbeat signals can be of great advantage in many applications. Thanks to the recent advances in computer vision techniques, this has been possible recently to measure heartbeat signals using a simple

webcam. Methods developed for this purpose are reviewed in the following subsection.

6.3.2. Contact- Free Measurement of Heartbeat Signal

The computer vision techniques developed for heartbeat measurement mostly utilize facial images. The reason for this goes back to this fact that heart pulses generate some periodic changes on the face, as well as other parts of the human body. However, since the human face is mostly visible, it is usually this part of the body that has been chosen by the researchers to extract the heartbeats from. The periodic changes that are caused by the heartbeat on the face are of two types:

- Changes in head motion which is a result of periodic flow of the blood through the arteries and the veins for delivering oxygenated blood to the body cells.
- Changes in skin color which is a result of having a specific amount of blood under the skin in specific periods of time.

None of these two types of changes are visible to the human eyes, but they can be revealed by computer vision techniques, like Eulerian magnification of [16] and [17]. The computer vision techniques developed for heartbeat measurement are divided into two groups, depending on the source (motion or color) they utilize for the measurement. These two types of methods are reviewed in the following subsections.

6.3.2.1 Motion for Contact-Free Extraction of Heartbeat Signal

The first motion-based contract-free computer vision method was just recently released by [18]. This system utilizes the fact that periodic heart pules, through aorta and carotid arteries, produce periodic subtle motions on the head/face which can be detected from a facial video. To do that, in [18] some stable facial points, known as good features to track, are detected and tracked over time. The features they track are located on the forehead area and the region between the mouth and then nose are as these areas are less affected by internal facial expressions and their moments should thus be from another source, i.e., heart pules. Tracking these facial points' results in a set of trajectories, which are first filtered by a Butterworth filter to re-move the irrelevant frequencies. The periodic components of these trajectories are then extracted by PCA and considered as the heartbeat rate. Their system has been tested on video sequences of 18 subjects. Each video was of resolution of 1280x720 pixels, at a frame rate of 30 with duration of 70-90 seconds.

To obtain the periodicity of the results of the PCA (applied to the trajectories), in [18] a Fast Fourier Transform (FFT) has been applied to the obtained trajectories from the good features to track points. Then, a percentage of the total spectral power of the signal accounted for by the frequency with the maximal power and its first

harmonic is used to define the heartbeat rate of the subject [18]. Though this produces reasonable results when the subjects are facing the camera, it fails when there are other sources of motion on the face. Such sources of motion can be for instance, changes in facial expressions and involuntary head motion. It is shown in [19] that such motions makes the results of [18] far from correct. The main reason is because the system in [18] uses the frequency with the maximal power as the first harmonic when it estimates the heartbeat rate. However, such an assumption may not always be true, specifically when the facial expression is changing [19]. To deal with these problems [19] has detected good features to track Figure 6.1(right) from the facial regions shown in Figure 6.1(left).

Figure 6.1: The facial regions (yellow areas in the left image) that have been used for detecting good feature to track (blue dots in the right image) for generating motion trajectories in [19].

Then, [19] tracks the good features to track to generate motion trajectories of these features. Then, it replaces the FFT with a Discrete Cosine Transform (DCT), and has employed a moving average filter before the Butterworth filter of [18]. Figure 6.2 shows the effect of the moving average filter employed in [19] for reducing the noise (resulting from different sources, e.g., motion of the head due to facial expression) in the signal that has been used for estimating the heartbeat rate.

Experimental results in [19] show that the above mentioned simple changes have improved the performance of [19] compared to [18] in estimating the heartbeat signal, specifically, when there are changes in facial expressions of the subjects. The system in [19] has been tested on 32 video sequences of five subjects in different facial expressions and poses with duration about 60 seconds.

To the best of our knowledge, the above two motion based systems are the only two methods available in the literature for contact-free measurement of the heartbeat rate using motion of the facial features. It should be noted that these systems do not report their performance/accuracy on estimating the heartbeat signal, but do so only for the heartbeat rate. The estimation of heartbeat signals are mostly reported in the color based systems which are reported in the next subsection.

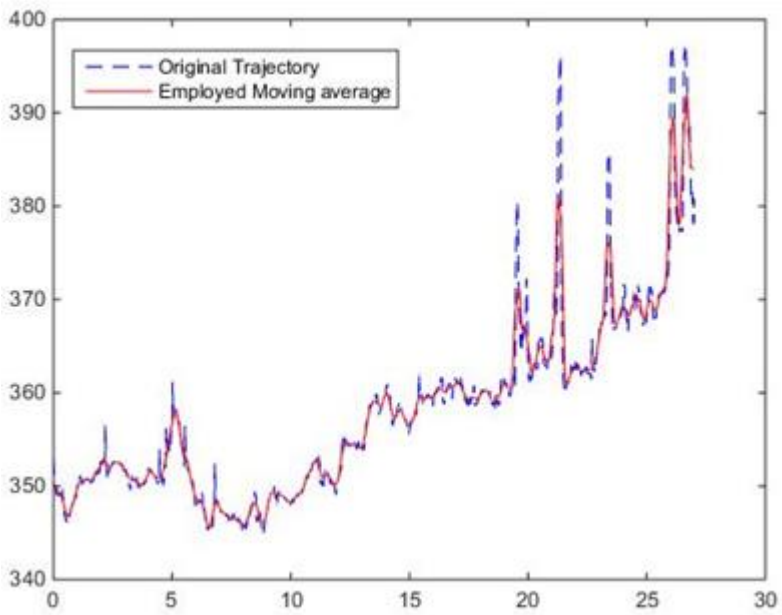


Figure 6.2: The original signal (red) vs. the moving averaged one (blue) used in [19] for estimating the heartbeat rate. On x and y-axis are the time (in seconds) and the amplitude of a facial point (from (good features to track)) that has been tracked, respectively.

6.3.2.2 Color for Contact-Free Extraction of Heartbeat Signal

Using expensive imaging techniques for utilizing color of facial (generally skin) regions for the purpose of estimating physiological signal has been around for decades. However, the interest in this field was boosted when the system of [20] reported its results on video sequences captured by simple webcams. In this work, [20], Independent Component Analysis (ICA) has been applied to the RGB separated color channels of facial images, which are tracked over time, to extract the periodic components of these channels. The assumption here is that periodic blood circulation makes subtle periodic changes to the skin color, which can be revealed by Eulerian magnification of [16]. Having included a tracker in their system, they have measured heartbeat signals of multiple people at the same time [20]. Further-more, it has been discussed in [20] that this method is tolerant towards motion of the subject during the experiment as it is based on the color of the skin. They have reported the results of their systems on 12 subjects facing a webcam that was about 0.5 meters away in an indoor environment. Shortly after in [21] it was shown that besides heartbeat, the methods of [20] can be used for measuring other physiological signals, like respiratory rate.

The interesting results of the above systems motivated others to work on the weakness of those systems, which had been tested only in constrained conditions. Specifically,

- In [22], it has been discussed the methods of [20] and [21] are not that efficient when the subject is moving (questioning the claimed motion tolerance of [20] and [21]) or when the lightning is changing, like in an outdoor environment. To compensate for these they have performed a registration prior to calculating the heartbeat signals. They have tested their system in an outdoor environment in which the heartbeat signals are computed for subjects driving their vehicles.
- In [23] auto-regressive modelling and pole cancellation have been used to reduce the effect of the aliased frequency components that may worsen the performance of a contact-free color based system for measuring physiological signals. They have reported their experimental results on patients that are under monitor in a hospital.
- In [24] normalized least mean square adaptive filtering has been used to reduce effect of changes in the illumination in a system that tracks changes in color values of 66 facial landmarks. They reported their experimental results on the large facial database of MAHNOB-HCI [25] which is publicly available. The reported results show that the system of [24] outperforms the previously published contact-free methods for heartbeat measurement including the color-based meth-ods of [20] and [21] and the motion-based of [18].

6.4. **Using Heartbeat Signal for Identification Purposes**

Considering the fact the heartbeat signals can be obtained using the two different methods explained in the previous section, different identification methods have been developed. These are reviewed in the following subsections.

6.4.1. Human Identification using Contact-based Heartbeat Signal

The heartbeat signals obtained by contact-based sensors (in both ECG and PCG forms) have been used for identification purposes for more than a decade. Here we only review those methods that have used ECG signals as they are more common than PCG ones for identification.

There are many different methods for human identification using ECG signals: [27]-[44], to mention a few. These systems have either extracted some features from the heart signal or have used the signal directly for identification. For example, it has been discussed in [27] that the heart signal (in an ECG form) composes of three parts: a P wave, a QRS complex, and a T wave (Figure 6.3). These three parts and their related key points (P, Q, R, S, and T, known as fiducial points) are then found and

used to calculate features that are used for identification, like the amplitude of the peaks or valleys of these point, the onsets and offsets of the waves, and the duration of each part of the signal from each point to the next. Similar features have been combined in [29] with radius curvature of different parts of the signal. It is discussed in [38] that one could ultimately extract many fiducial points from an ECG signal, but not all of them are equally efficient for identification.

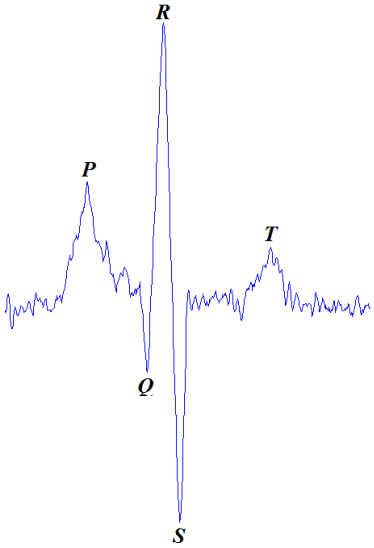


Figure 6.3: A typical ECG signal in which the important key points of the signal are labeled.

In [30], [31], [35], [37], and [39] it has been discussed that fiducial points detection based methods are very sensitive to the proper detection of the signal boundaries, which is not always possible. To deal with this, in:

- [30] the ECG signals have directly been used for identification in a Principal Component Analysis (PCA) algorithm.
- [31] the ECG signal has been first divided into some segments and then the coefficients of the Discrete Cosine Transform (DCT) of the autocorrelation of these segments have been obtained for the identification.
- [35] a ZivMerhav cross parsing method based on the entropy of the ECG signal has been used.
- [37], [40], [43] the shape and morphology of the heartbeat signal has been directly used as feature for identification.
- [39] sparse representation of the ECG signal has been used for identification.
- [33], [34], [36] frequency analysis methods have been applied to ECG signals for identification purposes. For example, in [33] and [36] wavelet

transformation has been used and it is discussed that it is more effective against noise and outliers.

6.4.2. Human Identification using Contact-free Heartbeat Signal

The above mentioned systems use contact-based (touch-based) sensors for measuring heartbeat signals. These sensors provide accurate, however sensitive to noise, measurements. Furthermore, they suffer from a major problem: these sensors need to be in contact with the body of the subject of interest. As mentioned before, this is not always practical, especially in identification context if subjects are not cooperative. To deal with this, we in [45] have developed an identification system that uses the heartbeat signals that are extracted using the contact-free technique of [21]. To the best of our knowledge this is the first system that uses the contact-free heart-beat signals for identification purposes. In this section we review the details of this system and its findings.

Having obtained the heartbeat signals, in form of RGB traces, from facial images using the method of [21] in [45] first a denoising filter is employed to reduce the effect of the external sources of noise, like changes in the lightning and head motions. To do that, the peak of the measured heartbeat signal is found and used to discard the outlying RGB traces. Then, the features that are used for the identification are extracted from this denoised signal. Following [26] the features that are used for identification in [45] are based on Radon images obtained from the RGB traces. To produce such images from the RGB traces, in [45] first the tracked RGB traces are replicated to the same number as the number of the frames that is available in the video sequence that is used for the identification, to generate a waterfall diagram. Figure 6.4 shows an example of such a waterfall diagram obtained for a contact-free measured heartbeat signal.

The Radon image, which contains the features that are going to be used for the recognition in [45], is then generated from the waterfall by applying a Radon transform to the waterfall diagram. Figure 6.5 shows the obtained Radon image from the waterfall diagram of Figure 6.4. The discriminative features that are used for identification purposes from such a Radon image are simply the distance between every two possible pixels of the image.

The experimental results in [45] on a subset of the large facial database of MAHNOB-HCI [25] shown in Table 6.1 indicate that the features extracted from the Radon images of the heartbeat signals that were collected using a contact-free computer vision technique carries some distinctive properties.

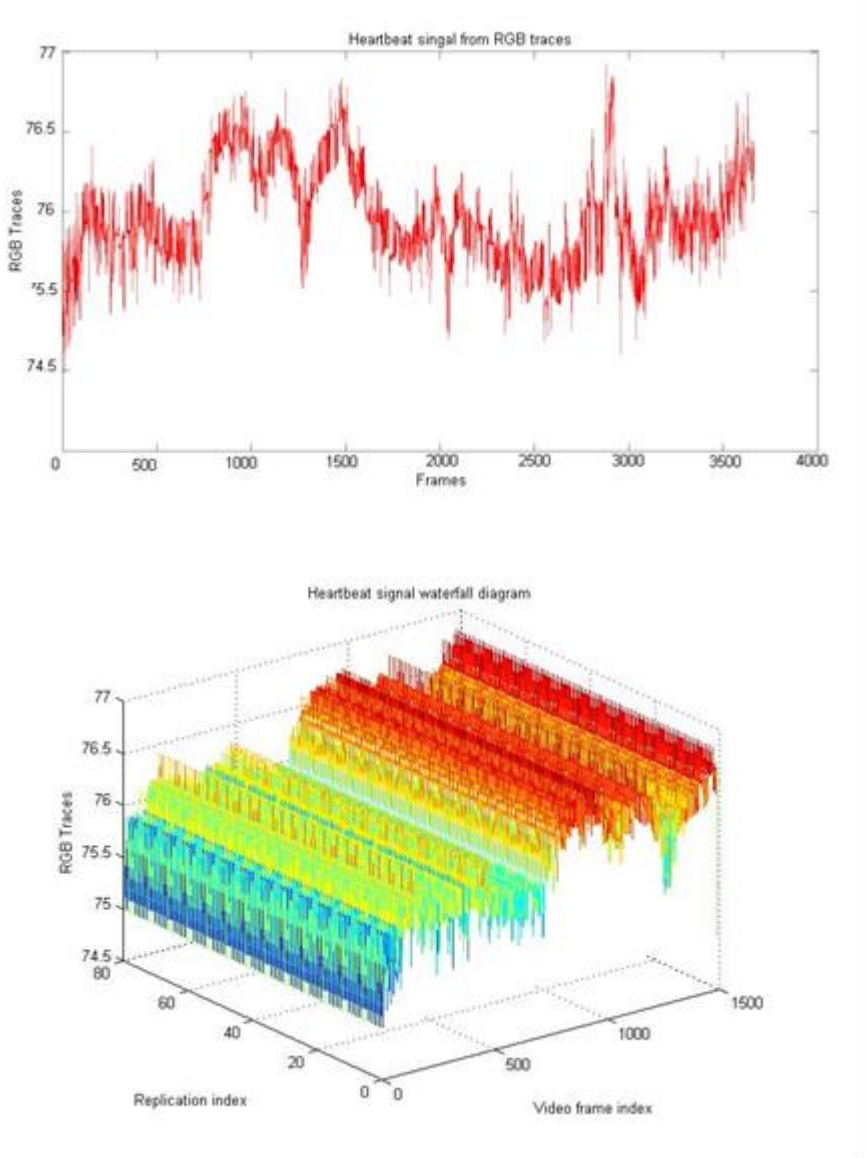


Figure 6.4: A contact-free measured heartbeat signals (on the top) and its waterfall diagram (on the bottom).

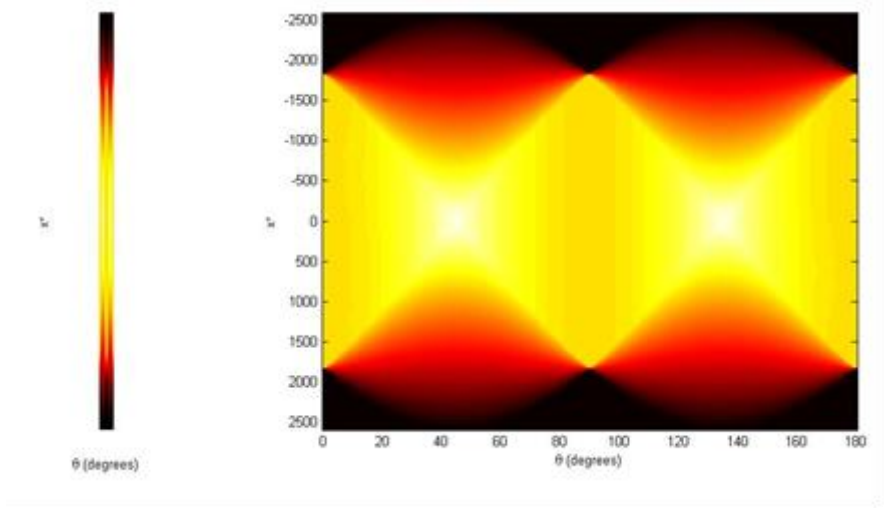


Figure 6.5: The Radon image obtained from the waterfall diagram of Figure 6.4. The discriminative features for the identification purposes are extracted from such Radon images [45]. The image on the right is the zoomed version of the one on the left.

Table 6.1: The identification results using distance features obtained from Radon images of contact-free measured heartbeat signals from [45] on a subset of MAHNOB-HCI [25] containing 351 video sequences of 18 subjects

Measured Parameter	Results
False positive identification rate	1.28 %
False negative identification rate	1.30%
True positive identification rate	98.70 %
Precision rate	80.86%
Recall rate	80.63%
Specificity	98.63%
Sensitivity	97.42%

6.5. Discussions and Conclusions

If the heartbeat signals are going to be used for identification purposes, they should serve as a biometric or a soft-biometric. A biometric signal should have some specific characteristics, among others, it needs to be:

- Collectible, i.e., the signal should be extractable. The ECG-based heartbeat signal is obviously collectible, though one needs to install ECG sensors on the body of subjects, which is not always easy, especially when subjects are not cooperative.
- Universal, i.e., everyone should have an instance of this signal. An ECG heartbeat signal is also universal as every living human has a beating heart. This also highlights another advantage of heartbeat signal which liveness. Many of the other biometrics, like face, iris, and fingerprints, can be spoofed by printed versions of the signal, and thus need to be accompanied by a liveness detection algorithm. Heartbeat signal however does not need liveness detection methods [35] and is difficult to disguise [29].
- Unique, i.e., instances of the signal should be different from one subject to another.
- Permanent, i.e., the signal should not change over time [15].

Regardless of the method used for obtaining the heartbeat signal (contact-free or contact-based), such a signal is collectible (according to the discussions in the previous sections) and obviously universal. The identification systems that have used the contact-based sensors (like ECG), [27]-[44] have almost all reported recognition accuracies that are more than 90% on datasets of different sizes from 20 people in [27] to about 100 subjects in the others. The lowest recognition rate, 76.9 %, has been reported in [34]. The results here are however reported on a dataset of 269 subjects for which the ECG signals have been collected in different sessions. Some of the sessions are from the same day and some are form different days. If both the training and the testing samples are from the same day (but still different sessions) the system report 99% recognition rate, but when the training and testing data is coming from different sessions recorded in different days, the recognition rate drops to 76.9% for the rank-1 recognition, but still as high as 93.5% for rank-15 recognition. This indicates that an ECG based heartbeat signal might be considered as a biometric, though there is not that much study on the permanent dimensions of such signals for the identification purposes, reporting their results on very large databases.

On the other hands, due to the measurement techniques that contact-free methods provide for obtaining the heartbeat signals, the discriminative properties of contact-free obtained heartbeat signals is not as high as their peer contact-based ones. However, it is evident from the results of our recent work, which was reviewed in the previous section, that such signals have some discriminative properties that can be utilized for helping identification systems. In another words, a contact-free obtained heartbeat signal has the potential to be used as soft-biometric.

To conclude, the contact-free heartbeat signals seem to have promising applications in forensics scenarios. First of all, because according to the above discussions they seem to carry some distinguishing features, which can be used for identification purposes. Furthermore, according to the discussion presented in [18] the motion

based methods, which do not necessarily need to extract these signals from a skin region, can extract the heartbeat data from the hairs of the head, or even from a masked face. This will be of great help in many forensics scenarios, if one could extract a biometric or even a soft-biometric from a masked face because in many such scenarios the criminals are wearing a mask to hide their identity from, among others, surveillance cameras.

6.6. **References**

[1] Ehrlich, D., Carey, L., Chiou, J., Desmarais, S., El-Difrawy, S., Koutny, L., Lam, R., Matsu-daira, P., Mckenna, B., Mitnik-Gankin, L., ONeil, T., Novotny, M., Srivastava, A., Streechon, P., and Timp, W.: MEMS-based systems for DNA sequencing and forensics. In: Sensors, Proceedings of IEEE, vol. 1, pp. 448-449, 2002.

[2] Lin, W.S., Tjoa, S.K., Zhao, H.V., Liu, K.J.R.: Digital image source coder forensics via intrinsic fingerprints. In: Information Forensics and Security, IEEE Transactions on, vol. 4, no. 3, pp. 460-475, 2009.

[3] Roussev, V.: Hashing and data fingerprinting in digital forensics. In: Security and Privacy, IEEE, vol. 8, no. 2, pp. 49-55, 2009.

[4] Peacock, C., Goode A, and Brett A.: Automatic forensic face recognition from digital images. In: Science and Justice, vol. 44, no. 1, pp. 29-34, 2004.

[5] Jain, A.K. and Unsang P.: Facial marks: soft biometric for face recognition. In: Image Processing (ICIP) 16th IEEE International Conference on, pp. 37-40, 2009.

[6] Unsang, P. and Jain, A.K.: Face matching and retrieval uses soft biometrics. In: Information Forensics and Security, IEEE Transactions, vol. no. 3, pp. 406-415, 2010.

[7] Jain, A.K., Klare, B., Unsang P.: Face recognition: Some challenges in forensics. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 726-733, 2011.

[8] Han, H., Otto, C., Liu, X., and Jain, A.: Demographic estimation from face images: human vs. machine performance. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. PP, no. 99, 2014.

[9] Wagner, J., Jonghwa K., Andre, E.:From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: Multimedia and Expo, IEEE International Conference on, pp. 940-943, 2005.

[10] Li, Lan and Chen, Jihua: Emotion Recognition Using Physiological Signals. In: Advances in Artificial Reality and Tele-Existence, Lecture Notes in Computer Science, Springer, vol. 4282, pp. 437-446, 2006.

[11] Kuriakose, S. and Sarkar, N. and Lahiri, U.: A step towards an intelligent Human Computer Interaction: Physiology-based affect-recognizer. In: Intelligent Human Computer Interaction (IHCI), 4th International Conference on, pp. 1-6, 2012.

[12] Liao, W., Zhang, W., Zhu, Z., and Ji, Q.: A real-time human stress monitoring system using dynamic Bayesian network. In: Computer Vision and Pattern Recognition - Workshops, IEEE Computer Society Conference on, 2005.

[13] Zhai J., Barreto A.: Stress detection in computer users through non-invasive monitoring of physiological signals. In: Biomedical sciences instrumentation, vol. 42, pp: 495-500, 2006.

[14] Barreto, A., Zhai, J., and Adjouadi, M.: Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In: Human-Computer Interaction, Lecture Notes in Computer Science, Springer, vol. 4796, pp. 29-38, 2007.

[15] Van de Haar, H., Van Greunen, D., and Pottas, D.: The characteristics of a biometric. In: Information Security for South Africa, 2013.

[16] Liu, C., Torralba, A., Freeman, W.T., and Durand, F., and Adelson, E.H.: Motion magnification. In: ACM Trans. Graph, vol. 24, no. 3, pp. 519-526, 2005.

[17] Wu, H.Y., Rubinstein, M. Shih, E., Guttag, J., Durand, F., and William T.F.: Eulerian video magnification for revealing subtle changes in the world. In: ACM Transactions on Graphics, proceedings of SIGGRAPH, vol. 31, no. 4, 2012.

[18] Balakrishnan, G., Durand, F., and Guttag, J.: Detecting pulse from head motions in video. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 3430-3437, 2013.

[19] Irani, R., Nasrollahi, K., and Moeslund, T. B.: Improved pulse detection from head motions using DCT. In: Computer Vision Theory and Applications, 9th International Conference on, vol. 3, pp. 118-124, 2014.

[20] Poh, M.Z., McDuff, D.J., and Picard, R.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. In: Opt. Express 18, pp. 10762-10774, 2010.

[21] Poh, M.Z., McDuff, D.J., and Picard, R.: Advancements in noncontact, multiparameter physiological measurements using a webcam. In: IEEE Transaction on on Biomedical Engineering, vol. 58, pp. 7-11, 2011.

[22] Sarkar, A., Abbott, A.L., and Doerzaph, Z.: Assessment of psychophysiological characteristics using heart rate from naturalistic face video data. In: Biometrics (IJCB), IEEE Interna-tional Joint Conference on, 2014.

[23] Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D.A., and Pugh, C.: Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. In: Physiol Meas vol. 35, pp. 807-831, 2014.

[24] Li, X., Chen, J., Zhao, G., and Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, pp. 4264-4271, 2014.

[25] Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M.: A multimodal database for affect recognition and implicit tagging. In: IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42-55, 2012.

[26] Hegde, C., Prabhu, H. R., Sagar, D. S., Shenoy, P. D., Venugopal, K. R., and Patnaik, L. M.: Heartbeat biometrics for human authentication. In: Signal Image Video Process., vol. 5, no. 4, pp. 485-493, Nov. 2011.

[27] Biel, L., Pettersson, O., Philipson, L., and Wide, P.: ECG analysis: a new approach in human identification. In: Instrumentation and Measurement, IEEE Transactions on, vol. 50, no. 3, pp, 808-812, 2001.

[28] Hoekema, R., Uijen, G.J.H., and van Oosterom, A.: Geometrical aspects of the interindividual variability of multilead ECG recordings. In: Biomedical Engineering, IEEE Transactions on, vol. 48, no. 5, pp. 551-559, 2001.

[29] Israel, S.A., Irvine, J.M., Cheng A., Wiederhold, M.D., Wiederhold, B.K.: ECG to identify individuals. In: Pattern Recognition, vol. 38, no. 1, pp. 133-142, 2005.

[30] Wang, Y., Plataniotis, K.N., Hatzinakos, D.: Integrating analytic and appearance attributes for human identification from ECG signals, In: Biometric Consortium Conference, Biometrics Symposium: Special Session on Research at the, 2006.

[31] Plataniotis, K.N., Hatzinakos, D., Lee, J.K.M.: ECG biometric recognition without fiducial detection. In: Biometric Consortium Conference, Biometrics Symposium: Special Session on Research at the, 2006.

[32] Singh, Y.N., Gupta, P.: ECG to individual identification, Biometrics: Theory, Applications and Systems, 2nd IEEE International Conference on, 2008.

[33] Fatemian, S.Z., Hatzinakos, D.: A new ECG feature extractor for biometric recognition. In: Digital Signal Processing, 2009 16th International Conference on, 2009.

[34] Odinaka, I., Po-Hsiang, L., Kaplan, A.D., O'Sullivan, J.A., Sirevaag, E.J., Kristjansson, S.D., and Sheffield, A.K., Rohrbaugh, J.W.: ECG biometrics: A robust short-time frequency analysis. In: Information Forensics and Security (WIFS), IEEE International Workshop on, 2010.

[35] Coutinho, D.P., Fred, A.L.N., Figueiredo, M.A.T.: One-lead ECG-based personal identification using Ziv-Merhav cross parsing. In: Pattern Recognition (ICPR), 20th International Conference on, pp. 3858-3861, 2010.

[36] Can, Y., Coimbra, M.T., Kumar, B.V.K.V.: Investigation of human identification using two-lead Electrocardiogram (ECG) signals. In: Biometrics: Theory Applications and Systems (BTAS), Fourth IEEE International Conference on, 2010.

[37] Islam, M.S., Alajlan, N., Bazi, Y., and Hichri, H.S.: HBS: A novel biometric feature based on heartbeat morphology. In: Information Technology in Biomedicine, IEEE Transactions on, vol. 16, no. 3, pp. 445-453, 2012.

[38] Tantawi, M., Revett, K., Tolba, M.F., Salem, A.: A novel feature set for deployment in ECG based biometrics. In: Computer Engineering Systems (ICCES), 7th International Conference on, pp. 186-191, 2012.

[39] Wang, J., She, M., Nahavandi, S., Kouzani, A.: Human identification from ECG signals via sparse representation of local segments, Signal Processing Letters, IEEE, vol. 20, no. 10, pp. 937-940, 2013.

[40] Fratini, A., Sansone, M., Bifulco, P., Romano, M., Pepino, A., Cesarelli, M., and D'Addio, G.: Individual identification using electrocardiogram morphology. In: Medical Measurements and Applications Proceedings (MeMeA), 2013 IEEE International Symposium on, pp. 107-110, 2013.

[41] Rabhi, E., Lachiri, Z.: Biometric personal identification system using the ECG signal. In: Computing in Cardiology Conference (CinC), pp. 507-510, 2013.

[42] Ming, L., Xin, L.: Verification based ECG biometrics with cardiac irregular conditions using heartbeat level and segment level information fusion. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, pp. 3769-3773, 2014.

[43] Lourenco, A., Carreiras, C., Silva, H., Fred, A.: ECG biometrics: A template selection approach. In: Medical Measurements and Applications (MeMeA), IEEE International Symposium on, 2014.

[44] Nomura, R., Ishikawa, Y., Umeda, T., Takata, M., Kamo, H., Joe, K.: Biometrics authentication based on chaotic heartbeat waveform, In: Biomedical Engineering International Conference (BMEiCON), 2014.

[45] Haque, M.A., Nasrollahi, K., and Moeslund, T.B.: Heartbeat signal from facial video for biometric recognition. In: 19th Scandinavian Conference on Image Analysis, Proceedings of, 2015.

PART III

ESTIMATION OF PSYCHOLOGICAL INDICATORS

Chapter 7

Pain Recognition using Spatiotemporal Oriented Energy of Facial Muscles

Ramin Irani, Kamal Nasrollahi, and Thomas B. Moeslund

This paper has been published in

IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), Boston, United States, 2015. pp. 80-87.

The layout has been revised.

7.1. **Abstract.**

Pain is a critical sign in many medical situations and its automatic detection and recognition using computer vision techniques is of great importance. Utilizes this fact that pain is a spatiotemporal process, the proposed system in this paper employs steerable and separable filters to measures energies released by the facial muscles during the pain process. The proposed system not only detects the pain but recognizes its level. Experimental results on the publicly available pain database of UMBC shows promising outcome for automatic pain detection and recognition.

7.2. **Introduction**

Pain is an unpleasant sensation that informs us about some (potential) damages or danger in the structure or the function of the body. It causes emotional effects like anger and depression and may even impact on the quality of life, social activities, relationships and our job. Yet pain is one of the most common reasons for seeking medical care, over 80% of patients complain about some sorts of pain [16]. So, for clinical trials and physicians, pain, similar to blood pressure, body temperature, heartbeat rate and respiration, is an important indicator of health. Therefore, reliable assessment of pain is essential for health related issues. That is why in 1995 Dr. James Campbell called the pain assessment as the fifth vital sign and suggested that "quality care means that pain is measured and treated" [20].

The most popular technique for pain assessment is Patient self-report. It is convenient and does not require special skills, but has some limitations. It includes inconsistent metrics, reactivity to suggestions, efforts at impression management and differences in conceptualizations of pain between clinicians and sufferers [15]. Moreover, self-reporting cannot be used, e.g., with children and those patients who cannot communicate properly due to neurological impairment or those who require breathing assistant. Craig et al. in [6] evidenced that changes in facial appearance can be a very useful cue for recognizing the pain. In Atul Gawande's recent book [9], it has been shown that periodically monitoring of patient's pain level by medical staff improves patients' treatment. However, sustained monitoring of patients by this way is difficult, unreliable and stressful. To solve this issue, automatic recognizing of pain using computer vision techniques, mostly from facial images, has received great attention over the past few years [6-14]. Brahman et al. [3] proposed a binary pain detection approach (pain versus no-pain) using Principal Component Analysis (PCA) and Support Vector Machines (SVM). Ashraf et al. [1] detected the pain using Appearance Active Model (AAM). Littlewort et al. [13] employed a two-layer SVM-based approach in order to detect real pain or posed pain. The above mentioned systems implement a binary classifier, meaning they recognize only two cases of pain versus no-pain, while based on the Prkachin and Solomon Pain Intensity metric [15], pain can be quantized into 16 discrete levels ranging from no-pain (0) to maximal pain (15).

To the Best of our knowledge, there are only few research articles that have estimated the pain level automatically, like those in [10, 11, 14, and 15]. In [14] a system has been developed which can detect three levels of pain intensity. It uses gemotry-based and appearance-based features with a separate SVM classifier for each intensity level of pain. Kaltwang et al. [11] proposed an approach using a combination of appearance-based features, Local Binary Pattern (LBP), and Cosine Discrete Transform (DCT), for detecting intensity levels of pain. They applied a Relevance Vector Regression (RVR) model to predict the pain intensity from each feature set. The above mentioned systems use handcrafted features like LBP and try different classifiers like PCA, SVM, and RVR to detected and recognize the pain. Though they produce interesting results, they do not consider the dynamics of the face. We have observed during our experiments that pain is exposed on the face through changes and motions of some of the facial muscles. These motions obviously release some energy. The level of the released energy is in direct relationship with the level of the pain. This is exactly the point that we want to exploit in this paper: we develop a system for pain recognition that measures the level of the released energy of the facial muscles over the time. Changes (activation) of facial muscles during the pain have been previously used for pain recognition in Prkachin and Solomon [18]. However, they do not consider the released energy of the facial muscles but detect the facial Action Units (AU)s and combine them to measure the pain.

There is not that many research work neither on exploiting the temporal axis nor on exploiting the released energy of the facial muscles for detecting and recognizing the pain. For example, [19] measures the pain over the temporal axis. However, it does not use the released energy of the muscles and is more focused on developing a classifier for pain recognition, which is based on Conditional Ordinal Random Fields (CORF). The only system that uses the released energy of facial muscles is the one developed by Hammal et al. [10]. This system uses a combination of AAM and an energy based filter, Log-normal filter, to estimate four intensity levels of pain. Though this system exploits the released energy of the facial muscles, it does that only on a frame by frame basis, in a spatial domain. The proposed system in this paper exploits the released energy of the facial muscles not only on in the spatial domain, but also in the temporal one. To do that, we use a specific type of spatiotemporal filter which is shown to be very useful for extracting information in both spatial and temporal domains at the same time, for other applications, like region tracking in [4], [7].

The rest of this paper is organized as follows: the employed filter and the other details of the proposed system are given in the following section. Section 4, explains the performed experiments and discusses the results that are obtained on a public facial database. Finally, section 5 concludes the paper.

7.3. The Proposed System

The block diagram of the proposed system is shown in figure 7.1. Following the diagram, given an input video sequence, the faces are first detected. Then, an Active Appearance Model (AAM) algorithm is used to align the detected faces in different frames of the video to a fixed framework using the provided landmarks (we assume the landmarks are already provided, as it is the case for the database employed). This registration to the fixed framework will cause losing some of the areas of the face, in some of the frames, which appear as holes or lines on the registered faces. To compensate for this, we use an inpanting algorithm. Then, the spatiotemporal filtering is performed in both *x*, *y*, and *t* dimensions to detect the energy released by the facial muscles motion of the aligned faces. Finally, the pain is detected and its level is recognized. These steps are explained in the following subsections.

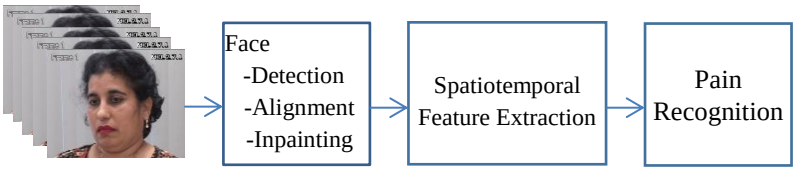


Figure 7.1: The block diagram of the proposed system.

7.3.1. Face Detection and Alignment

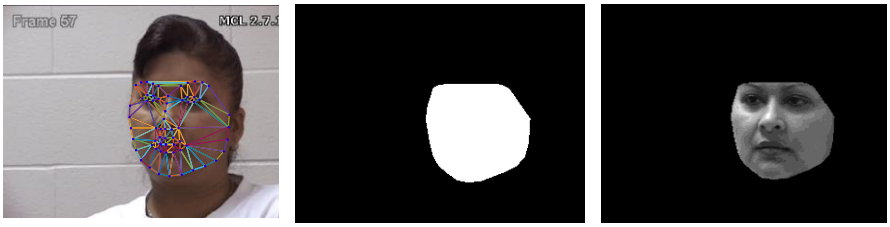
Detecting the face is an essential step in any facial analysis system, including, pain recognition. In this paper faces are detected using 66 facial landmark points that are provided with the employed database [15]. To do so, as it is shown in Ffigure 7.2.a, the facial landmarks are used as vertices of triangles which cover the entire face area, as it is done in [12]. This detected face needs to be segmented from the rest of the image. For this purpose, first, a binary mask (figure 7.2.b) is generated such that:

$$Mask = \bigcup_{k=1}^{K} I_k \tag{1}$$

Where

$$I_k = \begin{cases} 1, & P_{ij} \in T_k \\ 0, & Otherwise \end{cases} \tag{2}$$

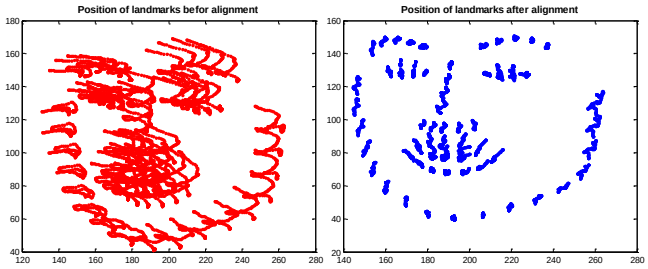
where T_k is the *k*th triangle created by landmark points, P_{ij} is a pixel on the image located at *(i, j)*, I_k is a binary image corresponding to T_k *and U* is a union function. Finally, the face can be segmented from the rest of the image by applying the mask on the image (figure 7.2.c).



a b c

Figure 7.2: a. Triangles generated from facial landmarks using the algorithm of [19] for face detection, b. the used mask for segmenting the face, and c. the segmented face.

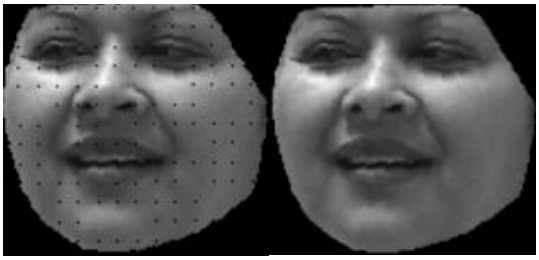
As mentioned before, the proposed system measures the energy that is released due to the motion of the facial muscles. However, in a video sequence, such motions are not the only type of motion. For example, figure 7.3.a shows the positions of 66 facial landmarks in a video sequence of 100 frames. If there was no motion in the video at all, one could only see 66 facial landmarks, but as it can be seen in figure 7.3.a, the position of each landmark is changing from one frame to another. This indicates the presence of other motions on the face, like motions resulting from the head pose. Such motions should be filter out. To do that, we employ the face alignment algorithm of [17]. The faces in this algorithm are aligned using the facial landmarks. The results of this alignment, applied to figure 7.3.a, can be seen in figure 7.3.b.



a b

Figure 7.3: Facial landmarks of 100 face images of a sequence: a. before and b. after alignment.

The alignment algorithm first finds the alignment parameters using the facial landmarks. Then, it uses these alignment parameters to wrap the face images of the input video sequence into a common framework using the wrapping algorithm of [5]. The reader is referred to [5] for the details of the alignment and wrapping algorithms. The result of this wrapping for a face image is shown in figure 7.4.a. It can be seen from figure 7.4.a that there are usually some holes (or even lines) in the results of the wrapped image, which indicate unknown pixel values. This is due to the wrapping of the facial images that are of different head poses. To deal with this, we use the inpainting algorithm of [2] which uses a series of up-sampling followed by down-sampling. The results of this algorithm applied to figure 7.4.a can be seen in figure 4.b.



a b

Figure 7.4: The wrapped aligned face image: a. before and b. after inpainting.

Having aligned the facial images of the input video sequence and generating an aligned facial video, using the above mentioned steps, the next step is to extract the spatiotemporal features. These features extract the direction and the level of the energies released by the facial muscles. These directions and levels are different for different facial expressions. For example, for a neutral face one should not expect too much energy to be released, while for a laughing face or a face suffering from pain, different levels of energy will be released by the facial muscles in different directions. Extracting of orientation and level of the released energy of the facial muscles are explained in the following subsection.

7.3.2. Spatiotemporal Feature Extraction

The extraction of the orientation and the level of the energies released by the facial muscles are done through steerable and separable filters of [4]. These filters compose of a second derivative Gaussian $G_2(\theta, \gamma)$ followed by a Hilbert transform $H_2(\theta, \gamma)$, in different directions θ, and scales γ. We do not use a multiscale method, because the level of the energy is not that much visible in coarse scales, hence $\gamma=1$. During the pain, however, the facial muscles can move in any directions, but such motions can be decomposed into four main directions. Therefore, we measure the released energies in four main directions corresponding to θ= 0, 90, 180, and 270, degrees. The released energy from every pixel is then calculated by:

$$E(x, y, t, \theta, \gamma) = [G_2(\theta, \gamma * I(x, y, t)]^2 + [H_2(\theta, \gamma * I(x, y, t)]^2 \qquad (3)$$

where * stands for a convolution operator, (x, y, t) shows the pixel value located at the position of x and y of the tth frame (temporal domain) of the aligned video sequence of I, and $E(x, y, t, \theta, \gamma)$ shows the energy released by this pixel at the direction of θ and the scale of γ. To make the above obtained energy measure comparable in different facial expressions, we normalize it using:

$$\hat{E}(x,y,t,\theta,\gamma) = \frac{E(x,y,t,\theta,\gamma)}{\sum_{\theta_i} E(x,y,t,\theta_i,\gamma) + \epsilon} \tag{4}$$

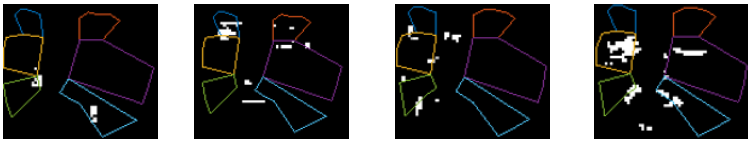
where θ_i considers all the directions and ϵ is a small bias used for preventing numerical instability when the overall estimated energy is too small. Finally, to improve the localization, we weight the above normalized energy using [15]:

$$\dot{E}(x,y,t,\theta,\gamma) = \hat{E}(x,y,t,\theta,\gamma).z(x,y,t,\theta) \tag{5}$$

where:

$$(x,y,t,\theta) = \begin{cases} 1, & if \sum_{\gamma_i} \hat{E}(x,y,t,\theta,\gamma_i) > Z_\theta \\ 0, & otherwise \end{cases} \tag{6}$$

in which Z_θ is a threshold for keeping energies at the direction θ, as too small energies are likely to be noise. The weighted normalized energy obtained in Eq. 5 assigns a number to each pixel (corresponding to the level of the released energy by that pixel) in each of the four chosen directions of θ= 0, 90, 180, and 270. Figure 7.5 shows these pixel-based energies for a facial image computed at the four different orientations.



θ= 0 θ= 90 θ= 180 θ= 270

Figure 7.5: Pixel-based energies of Fig. 4b computed at four different orientations. The different colors represent different facial regions.

The above obtained pixel-based energies can be converted into a more understandable form, if we study the regional changes/motions of the facial muscles of different parts of the face. Based on our observations, different regions of the face contribute differently to the level and direction of the energy in different facial statuses. We have observed that facial muscles that are actively participating to the facial motions during the pain are coming from the three regions that are highlighted in figure 7.6. Besides this, the facial muscles on the left side and the right side of each of these three regions are participating differently in motions during the pain. Because of this, inside each region we have used different colors to distinguish between the left and the right side.

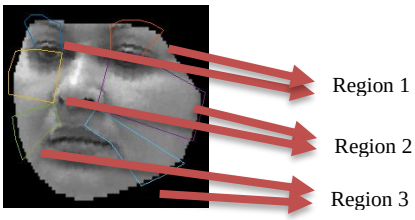


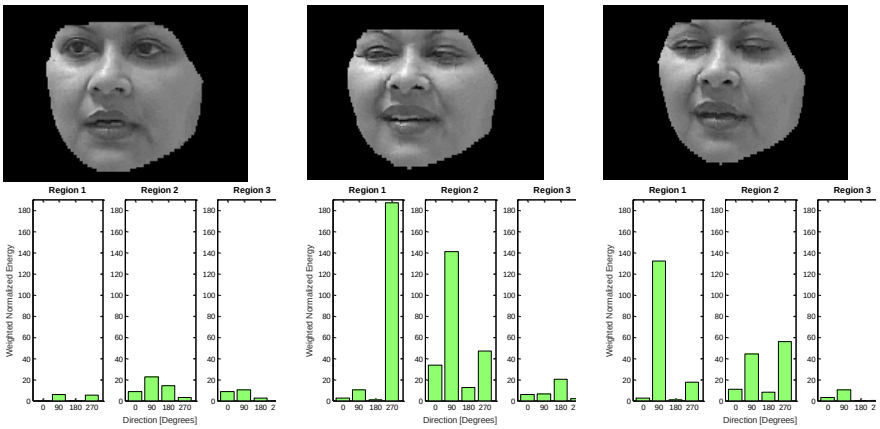
Figure 7.6: The facial muscles of these three regions are actively contributing to the facial motions during the pain.

To convert the pixel-based energies into region based energies, we obtain the histograms of the directions of the pixel-based energies, computed above, for all the three mentioned facial regions. We calculate the histogram of the directions, H_{R_i} by:

$$H_{R_i}(t;\theta_i,\gamma) = \sum_{R_i} \dot{E}(x,y,t;\ \theta_i,\gamma): \qquad i = 1,2,3 \tag{7}$$

where R_i $i=1,2,$ or 3 is the ith region of the face. Figure 7.7 a to figure 7.7.c each show three histograms of directions of the weighted normalized energies that are obtained using Eq. 7 at three different stages of a pain process. Figure 7.7.a (top) shows a neutral face, therefore, there is not that much energy in either of the directions. It can be verified by figure 7.7a (bottom). Figure 7.7.b (top) shows just the beginning of a pain, it can be seen in figure 7.7.b (bottom) that muscles in region 1 release energy in direction 270 degree (downwards), those in region 2 release energy in direction 90 degree (upwards) and those in region 3 release some energies to sides (here direction 180 degree). Figure 7.7.c (top) shows the face just before revealing from the pain. It can be seen from figure 7.7.c (bottom) that muscles are releasing energy in the opposite direction of figure 7.7.b (bottom) to get back to their original locations.

The above obtained energies can inform us only about some muscles activities (motions), but we need a specific interpretation to see if these motions are due to the pain or not. To do that, we need to study the effect of the pain on the motions of the muscles in the temporal domain. To consider the time domain, we simply obtain the histograms of the directions for each facial region in the aligned input video (figure 7.8 (left)). However, as mentioned, since the muscles will move back to their original locations at the end of the pain, instead of the measured directions, we simply consider the changes of the released energies of the muscles in two main orientations, up-down (UD) and left-right (LR). This will convert the histograms of directions in Fig. 8 (left) into histograms of orientations, as shown in figure 7.8 (right).



a b c

Figure 7.7: Histograms of directions of the normalized released energies from different facial regions (bottom row), for three different images (top row) at different stages of a pain process: a. neutral face, b. the beginning of the pain, and c. just before the end of the pain.

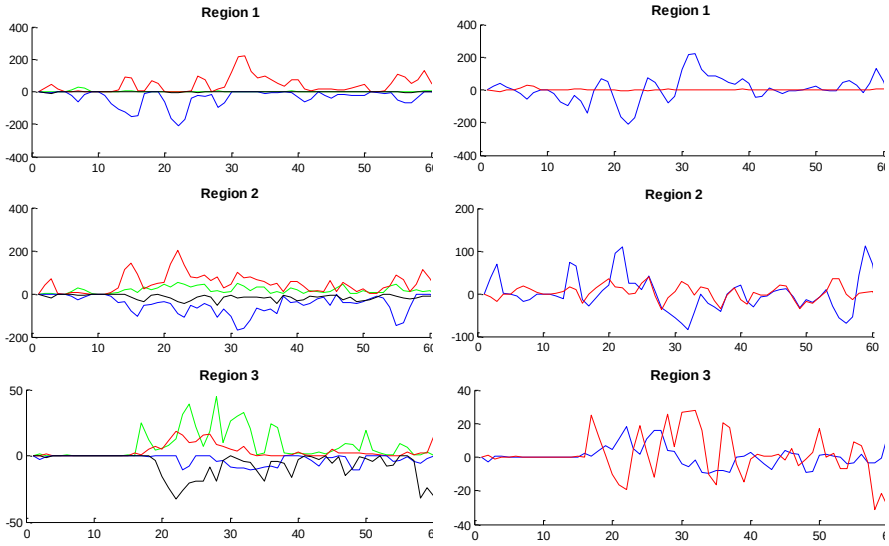


Figure 7.8: Histograms of directions (left) and histograms of orientations (right) for three facial regions for the aligned facial images of a video sequence of 60 frames. The red, green, blue, and dark colors in the left column show the released energies at directions, 0, 90, 180, and 270, respectively. The red and blue colors in the right column show the UD, and LR histograms of orientations, respectively.

7.3.3. Pain Recognition

Having obtained the histograms of orientations from each region, the final step is to combine them by considering the temporal domain and recognize the pain. To consider the temporal domain for monitoring the changes in the released energy we take the integral of the two *UD* and *LR* histograms of orientations in each region, using:

$$A_{R_{i_{UD}}} = \sum_{t=1}^{n} UD_t \qquad A_{R_{i_{LR}}} = \sum_{t=1}^{n} LR_t \tag{8}$$

where $A_{R_{i_{UD}}}$ and $A_{R_{i_{LR}}}$ are the integrals of *UD* and *LR* for the *i*th region (*i=1,2,3*), respectively, and *n* is the number of the frames in the aligned video. Finally, the pain intensity (*PI*) is obtained by calculating the above two integrals for each of the three regions:

$$PI = \sum_{i=1}^{3} w_{R_{i_{UD}}} A_{R_{i_{UD}}} + \sum_{i=1}^{3} w_{R_{i_{LR}}} A_{R_{i_{LR}}} \tag{9}$$

where $w_{R_{i_{UD}}}$ and $w_{R_{i_{LR}}}$ define some experimentally obtained weights of the corresponding regional histogram of orientations. The *PI* gives us an indication of the presence of the pain in each frame of the video. Depending on the value of PI we find some experimentally achievable thresholds to classify the pain into three class of no-pain, weak, and strong. The experimental results are given in the next section.

7.4. Experimental results

The proposed system has been implemented in Matlab 2014b. We've used the publicly available UNBC-MacMaster Shoulder Pain Expression Archive Database [15] for evaluating of our model. This database is composed of 25 participants who suffer from pain in their shoulders. They have been filmed during series of movements in two different scenarios (active and passive). In the active scenario participants move their arms themselves, but in the passive scenario a physiotherapist is responsible for this. Videos were captured at a resolution of 320×240. The total number of the recorded frames is 48398. In this database, the ground truth pain information has been provided using AUs (see figure 7.9), by:

$$Pain = AU4 + \max(AU6, AU7) + \max(AU9, AU10) + AU43 \tag{10}$$

For each frame in the database, the AU intensities were coded on a 6 level scale except the AU number 43 which was coded on two levels [15].

Figure 7.9: Active AUs of Pain, the image is from the UNBC database of [14].

To evaluate our system we selected randomly 50 sequences from 12 participants, containing 4926 frames. Table 7.1 shows the results of the proposed system against the ground truth data obtained using Eq. 14, and the results of the system developed in [10]. It can be seen from Fig. xx and this table, that our system not only detects the pain but also recognizes three different levels of the pain. These three levels are no-pain where $PI = 0$, weak pain where $PI = 1$ or 2, and strong pain where $PI \geq 3$. The proposed system actually outperforms the system of [10] in terms of the accuracy of recognizing the level of the pain. It should be mentioned that system [10] is the only energy based system in the literature for calculating the pain, but it is working in the spatial domain. Outperforming this system by our proposed system means that including the temporal information in an energy-based pain recognition system results in better outcomes.

Table 7.1: Comparing the results of the proposed system against the system of [10], against the ground truth applied to the images of the UNBC database [15].

Semantic Ground Truth	Pain Index Ground Truth	Number of Frames	System of [10] (in%)	Proposed System (in%)
No pain	0	4230	65	77
Weak	1, 2	387	36	62
Strong	≥3	309	70	70

We should take into account that the pain intensity in [10] has been classified into four levels. Two pain levels of Trace and Weak in [10] corespond to the level Weak in our system. Therefore, for comparision purposes mean of the Trace and Weak levels has been reported as Weak in table 7.1.

Finally, figure 7.10 shows the *PI* values obtained by the proposed system (using Eq. 9) for two different video sequences each containing100 facial images, along with their ground truth data taken from the database employed. It can be seen from this figure, that there is a good overlap between the peaks of the estimated pain intensity curve and the ground truth. It should be noted that the negative values in the estimated values will be considered as zero (no pain).

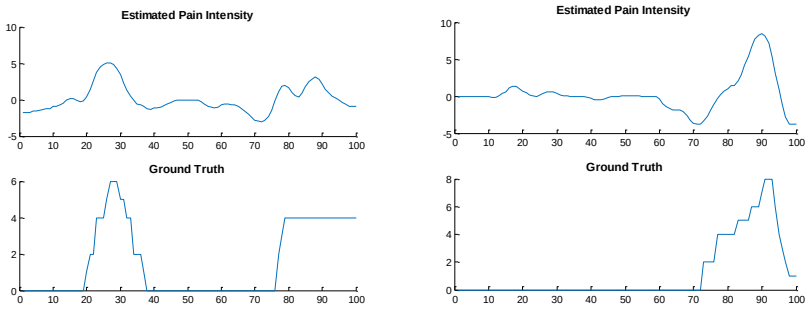


Figure 7.10: Comparing the pain levels obtained by the proposed system (top row) against the Ground truth (bottom row), for two different pain scenarios.

7.5. **Conclucion**

The proposed system in this paper uses separable steerable filters for automatic detection and recognition of pain. To do that, it applies these filters in both spatial (*x*, and *y* axises) and temporal (time axis) domains and measures the energies released by the facial muscles that are during the pain process. The porposed system has produced promising experimental results on the publicly available dataset of UNBC [14]. However, it still needs further improvement especially in the wrapping step to compensate for the head poses, for which we plan to use 3D information of facial landmarks in our future works.

7.6. **References**

[1] A. Ashraf, S. Lucey, T. Chen, K. Prkachin, P. Solomon, Z. Ambadar, J. Cohn, "The painful face: pain expression recognition using active appearance models", *In Proceedings of the 9th International Conference on Multimodal interfaces*, pp. 9-14, 2007.

[2] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, "Image inpainting", *In Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417-424, 2000.

[3] S. Brahnam, C. Chuang, F. Shih, M. Slack, "Machine recognition and representation of neonatal facial displays of acute pain", *Artificial Intelligence in Medicine*, vol. 36, pp. 211-222, 2006.

[4] K. Cannons, R. Wildes, "The applicability of spatiotemporal oriented energy features to region tracking", *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 36, no. 4, pp. 784-796, 2014.

[5] T. Cootes, C. Taylor, "Statistical models of appearance for computer vision", *Technical report Imaging Science and Biomedical Engineering University of Manchester*, 2004.

[6] K. Craig, K. Prkachin, R. Grunau, "The facial expression of pain", *Handbook of pain assessment Guilford New York*, 2001.

[7] K. Derpanis, J. Gryn, "Three-dimensional nth derivative of gaussian separable steerable filters", *In IEEE International Conference on Image Processing*, pp. 553-556, 2005.

[8] R. Donner, M. Reiter, G. Langs, P. Peloschek, H. Bischof, "Fast active appearance model search using canonical correlation analysis", *IEEE Trans. Pattern Anal. Mach. In tell*, vol. 28, no. 10, pp. 1690-1694, 2006.

[9] A. Gawande, "The checklist manifesto: How to get things right?", *Metropolitan Books New York*, 2010.

[10] Z. Hammal, J. Cohn, "Automatic detection of pain intensity", *In Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 22-26, 2012.

[11] S. Kaltwang, O. Rudovic, M. Pantic, "Continuous pain intensity estimation from facial expressions", *In International Symposium on Advances in Visual Computing*, pp. 368-377, 2012.

[12] D. Lee, B. Schachter, "Two algorithms for constructing a delaunay triangulation", *International Journal of Computer and Information Sciences*, vol. 9, no. 3, pp. 219-242, 1980.

[13] G. Littlewort, M. Bartlett, K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain", *Image and Vision Computing*, vol. 27, pp. 1797-1803, 2009.

[14] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, S. Chew, I. Matthews, "Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database", *Image and Vision Computing*, vol. 30, pp. 197-205, 2012.

[15] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database", *In Automatic Face Gesture Recognition and Workshops (FG 2011) 2011 IEEE International Conference on*, pp. 57-64, 2011.

[16] M. S. Walid, S. N. Donahue, D. M. Darmohray, L. A. Hyer, J. S. Robinson, "The fifth vital sign-what does it mean?", Pain Practice, vol. 8, no. 6, pp. 417-422, 2008.

[17] C. Miaskowski, M. Bair, R. Chou, "Principles of analgesic use in the treatment of acute pain and cancer pain", *Chicago: American Pain Society*, 2008.

[18] K. Prkachin, P. Solomon, "The structure reliability and validity of pain expression: Evidence from patients with shoulder pain", *Pain*, vol. 139, pp. 267-274, 2008.

[19] O. Rudovic, V. Pavlovic, M. Pantic, "Automatic pain intensity estimation with heteroscedastic conditional random fields", *Advances in Visual Computing*, vol. 8034, pp. 234-243, 2013.

[20] “Pain Management,” Pain Management - Canadian Hemophilia Society. [Online]. Available: http://www.hemophilia.ca/en/care-and-treatment/pain-management/. [Accessed: 07-Apr-2017].

Chapter 8

Spatiotemporal Analysis of RGB-D-T Facial Images for Multimodal Pain Level Recognition

Ramin Irani, Kamal Nasrollahi, Marc O. Simon, Ciprian A. Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H. Lundtoft, Thomas B. Moeslund, Tanja L. Pedersen, Maria-Louise Klitgaard, and Laura Petrini

This paper has been published in

IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), Boston, United States, 2015. pp. 88-95.

The layout has been revised.

8.1. Abstract

Pain is a vital sign of human health and its automatic detection can be of crucial importance in many different contexts, including medical scenarios. While most available computer vision techniques are based on RGB, in this paper, we investigate the effect of combining RGB, depth, and thermal facial images for pain intensity level recognition. For this purpose, we extract energies released by facial pixels using a spatiotemporal filter. Experiments on a group of 12 elderly people applying the multimodal approach show that the proposed method successfully detects pain and recognizes between three intensity levels in 82% of the analyzed frames, improving by more than 6% the results that only consider RGB data.

8.2. Introduction

Pain plays an essential role as part of a complex system for dealing with injury [27]. Distinguishing harmful from harmless situations, prompting avoidance of harm and its associated cues, giving a high priority to escape from danger, and promoting healing by inhibiting other activities that might cause further tissue damage has great adaptive value [4]. Pain serves to promote the organism's health and integrity, to the extent that congenital absence of pain on injury significantly shortens human life [10]. There are rare cases of people with no pain sensation. An oftencited case is that of F.C., who did not exhibit a normal pain response to tissue damage. She repeatedly bit the tip of her tongue, burned herself, did not turn over in bed or shift her weight while standing, and showed a lack of autonomic response to painful stimuli. She died at the age of [29], [25].

Physiological measures of pain vary significantly from person to person, failing to reflect its intensity [24]. Indicators of pain include changes in heart and respiratory rates, blood pressure, vagal tone, and palmar sweating [8]. In addition to physiological responses, facial expressions play a critical role for communicating pain [30, 28 and 12].

The main way of assessing pain in clinical contexts is by self report. This can be sometimes problematic because of the lack of reliability and consistency and in the case of persons with limited communication abilities (infants, young children, patients with certain neurological impairments, intubated and unconscious persons) this option is simply not available [14]. One alternative is to have human experts to ass's pain. Unfortunately, a considerable amount of training is needed and the process is burdensome. Another alternative is the automatic assessment of pain. Even though there is no consensus about the physiological measures of pain [6] brow-lowering, tightening the eyelids, raising the cheeks (orbit tightening), nose wrinkling or upper-lip raising, and eye closure were identified as a core set of actions during facial expression of pain [29, 26].

Pain is a complex behavioural expression. While until now research has mostly concentrated on facial pain recognition from RGB, in the case of human affect perception there is strong evidence supporting the integration of multiple modalities over using a single modality [2, 32, 35]. For instance, depth information offers important advantages. It is more invariant to rotation and illumination and it captures more subtle changes on the face. It also facilitates recognizing a wider range of expressions which would be more difficult to detect from only RGB. Pain experience is highly correlated with cardiovascular changes such as increase in heart rate, blood pressure, cardiac output and blood flow [17], which in turn has a direct effect on skin temperature. Radiance at different facial regions, captured through thermal infrared imaging, varies according to emotions [38]. While we could not find a similar study in the case of pain, thermal imaging could add valuable information in the automatic assessment of pain. Because the different modalities can be redundant, concatenating features might not be efficient. A common solution is to use fusion. Many studies have demonstrated the advantage of classifier fusion over the individual classifiers [19]. Finally, another important aspect of pain assessment from facial expressions is temporal information [9, 1]. For instance, it has been shown that temporal dynamics of facial behavior represent a critical factor for distinction between spontaneous and posed facial behavior [3, 11, 36] and for categorization of complex behaviors like pain [11, 42].

In this paper, we present a multimodal dynamic pain recognition method from RGB, depth and thermal facial images. We extract energies released by facial pixels in these three modalities using a spatiotemporal filter and test the methodology in a real case scenario consisting of 12 subjects, obtaining high recognition rates measuring the level of pain, and showing the benefits of including multimodal information.

The rest of this paper is organized as follows. In Section 3 we review related methods in the field. Methodology is described in Section 4. Section 5 presents the experimental results. Finally, Section 6 concludes the paper.

8.3. Related Work

While the vast majority of related work described in the literature focuses on pain recognition from RGB (e.g. [21, 3, 31], works based on 3D [39] or multimodal [40, 41] also exist. A first category of RGB methods do not include temporal information [13, 22, 21, 23]. In [13] a method for automatically detecting four levels of pain intensity is proposed based on SVM trained with the responses of Log-normal Filters. In [22], Littlewort et al. classify real and fake pain with 88% accuracy compared to the 49% obtained by naive human subjects. In [3], a pain no pain classification is proposed, achieving a hit rate of 81% using AAM and SVM. Similar to [22], [21] obtains above naive human discrimination between posed and genuine facial expression of pain (72% compared to 52%) by using boosted Gabor filters and SVM. Finally, in [23], Lucey et al. show that detecting pain by fusing pain associated AUs is more efficient than using extracted features to directly detect pain/no-pain. A distinct

group of RGB methods use temporal information [31, 40, 18, 15]. In [18] Kaltwang et al. propose what they claim to be the first fully automatic continuous pain intensity method. It is based on a late fusion of a set of regression functions learned from appearance (DCT and LBP) and geometric (shape) features. In [31] facial expressions of pain intensity are detected by using Conditional Ordinal Random Fields (CORF). In [15] an approach based on the Transferable Belief Model are proposed capable of obtaining above human observer's performance when recognizing the pain expression among the six basic facial expressions and neutral on acted and spontaneous sequences. Examples of using other modalities include 3D [39] and physiological signals in a multimodal context [41, 40]. [39] is based on a SVM classifier and a function model for intensity rating. The intensity model is trained using Comparative Learning, a technique that simplifies labelling of data. In [40, 41] it is proposed a multimodal (RGB+physiological) dataset (BioVid Heat Pain Database) and dynamic methods for recognizing pain by combining information from video and biomedical signals, namely facial expression, head movement, galvanic skin response, electromyography and electrocardiogram. In contrast to previous works, the method we propose here is the first one to combine RGB, depth and thermal facial images for pain recognition.

8.4. **Methodology**

The block diagram of the proposed system is shown in figure 8.1. Having a trimodal input video, first the RGB modality is used to detect the face and facial landmark positions. The registration information between the three modalities is used to estimate the positions of the landmarks in the other two modalities. Afterwards, an energy-based method using steerable separable spatiotemporal filter, which uses the landmark positions, is applied to each modality. This gives an indication of visible pain in each of these modalities. Finally, a fusion unit is used to combine the results of the three modalities to recognize the pain level. These steps are explained in the following subsections.

8.4.1. Landmark detection in RGB

In order to develop a fully automatic pain recognition method, first a landmark detection approach is applied over the RGB modality. Two steps are required in order to efficiently detect landmarks in video sequences. Firstly, the Viola&Jones [37] face detection algorithm is applied to the first frame. Landmarks are then located inside the facial region by using the Supervised Descent Method (SDM)

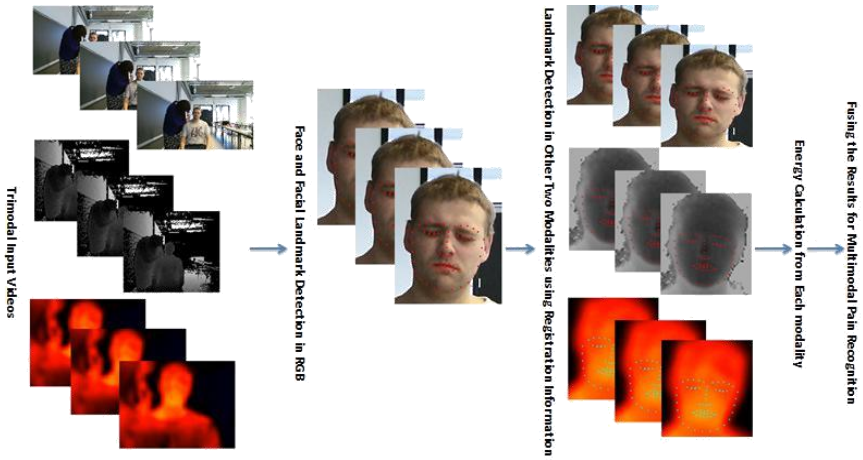


Figure 8.1: The block diagram of the proposed system.

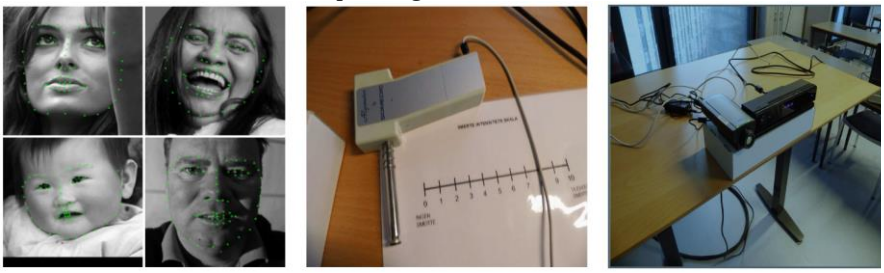
[43]. In the subsequent frames, the facial region is obtained from the previous frame geometry, applying SDM inside that region to estimate the new landmark locations. The SDM algorithm consists on a custom implementation trained for the detection of 68 landmarks (see figure 8.2.a). For training, a combination of the LFPW [5], HELEN [20], AFW [44] and IBUG [33] datasets is used, amounting to 3837 in-stances. The ground truth of the 68 facial landmarks over these datasets is obtained from the 300 Faces In-The-Wild Challenge (300-W) [34].

Since the tracking approach uses the previous frame landmarks to select the facial region, it is important to have a robust algorithm for landmark localization. SDM is less prone to local minima when compared to other minimization methods [43], learning the descent direction and step size towards the global minima regardless of the gradient at the current estimate. This is achieved with a cascaded approach, where an initial shape estimate S^0 is iteratively adjusted to the image though linear regressors. At each step a simplified version of SIFT is used to extract features from the landmarks. These are concatenated into a feature vector F_{SIFT}^t, where the dimensionality has been reduced by using PCA to keep 95% of the original variance. A linear regression w^t estimates the displacement between the current shape estimate S^t and the face geometry, as shown in Equation 1.

$$S^{t+1} = S^t + F_{SIFT}^t . w^t \quad (1)$$

This robustness can be further improved by using multiple initializations. For this purpose, n = 10 plane rotations of the mean shape are homogeneously sampled from the range [=2, + =2] and fit to the image during test. The distance between each pair of fits $D_{i,j} = d(S_i, S_j)$ is stored into a matrix $D^{<n \times n>}$ of distances, being $d(x, y)$ the the sum of euclidean distances between

corresponding landmarks. The



a b c

Figure 8.2: a. Sample images from the 300-W dataset labeled with 68 facial landmarks, b. The hand-held device used for introducing the pain, c. The cameras used for capturing the three modalities.

fit minimizing the sum of distances to the others, i.e. the centroid fit, is selected as the best one. This criterion is used because it corresponds to the fit towards which most other alignments, regardless of the initialization orientation, tend to converge, thus having a higher probability of corresponding to the global minima.

8.4.2. Landmark detection in depth and thermal

The landmarks obtained from the RGB frames are translated to the corresponding frames in the depth and thermal modalities by first finding a registration between the three modalities. Once the registration is found, we transform the landmarks coordinate system from RGB to both depth and thermal, representing the geometry in those spaces. The registration of these modalities is explained in the following subsection.

8.4.2.1 Registration of different modalities

The registration between the RGB and depth modalities uses the built-in calibration tool of the KinectT M for Windows 2.0 SDK. Although the calibration parameters used for the registration are not directly visible, it is possible to obtain an accurate registration of each depth image to the corresponding RGB frame. Registration of the thermal modality to RGB requires considering two modalities captured by two separate devices, whose relative positions are not know beforehand. Therefore, we obtain a separate calibration of the thermal and RGB modalities by moving a custom-made multimodal checkerboard in the region where the upper body of test participants is located. The multimodal calibration board consists of a white, A3-sized 10 mm polystyrene backdrop which is heated by a heat gun immediately before the calibration, and thick card board plate where a chessboard pattern is cut out. The differences in temperature and color of the two boards enable the detec-tion of point correspondences between RGB and thermal. The point correspondences from the

calibration stage is used to obtain a homography which is accurate for points near the face of the participants.

8.4.3. Feature extraction

Having found the positions of the landmarks in all the three modalities, the next step is to use these positions to extract a feature that can give us an indication of the pain in each modality. The following steps should be performed for all the three modalities similarly, thus we will explain them only for RGB modality.

Since changes due to pain in facial expression are spa-tiotemporal phenomena, we need to employ a descriptor that considers both spatial and temporal domains, and can be independently applied to all three modalities. For these reasons a steerable separable spatiotemporal filter has been chosen, which considers the second derivative of a Gaussian filter and their corresponding Hilbert transforms. This filter measures the orientation and level of energy in the 3D space of x, y, and t, representing the spatial and time spaces, respectively. The spatial responses of the filter describe the spatial texture of the face, while the temporal responses describe the dynamic of the features, e.g., the velocity. For each pixel, the energy is calculated by:

$$E(x, y, t, \theta, \gamma) = [G_2(\theta, \gamma * I(x, y, t)]^2 + [H_2(\theta, \gamma * I(x, y, t)]^2 \quad (2)$$

where * stands for a convolution operator, (x, y, t) shows the pixel value located at the position of x and y of the tth frame (temporal domain) of the aligned video sequence of I, and $E(x, y, t, \theta, \gamma)$ shows the energy released by this pixel at the direction of θ and the scale of γ. To make the above obtained energy measure comparable in different facial expressions, we normalize it using:

$$\hat{E}(x, y, t, \theta, \gamma) = \frac{E(x, y, t, \theta, \gamma)}{\sum_{\theta_i} E(x, y, t, \theta_i, \gamma) + \epsilon} \quad (3)$$

where θ_i considers all the directions and ϵ is a small bias used for preventing numerical instability when the overall estimated energy is too small. Finally, to improve the localization, we weight the above normalized energy using [15]:

$$\dot{E}(x, y, t, \theta, \gamma) = \hat{E}(x, y, t, \theta, \gamma).z(x, y, t, \theta) \quad (4)$$

where:

$$z(x, y, t, \theta) = \begin{cases} 1, & if \ \sum_{\gamma_i} \hat{E}(x, y, t, \theta, \gamma_i) > Z_\theta \\ 0, & otherwise \end{cases} \quad (5)$$

in which Z_θ is a threshold for keeping energies at the direction θ, as too small enegies are likely to be noise. The weighted normalized energy obtained in Eq. 5 assigns a

number to each pixel (corresponding to the level of the released energy by that pixel) in each of the four chosen directions of θ= 0, 90, 180, and 270. Following [16] these pixel based energies are then combined into region based energies using their histograms of directions by:

$$H_{R_i}(t;\theta_i,\gamma) = \sum_{R_i} \dot{E}(x,y,t;\ \theta_i,\gamma): \quad i = 1,2,3 \tag{6}$$

where H_{R_i} is the histogram of the directions, and R_i, $i = 1; 2$, or 3 is the ith region of the face [16]. Since the muscles are moving back to their original locations, after they are moved due to, e.g., pain, we need to combine the regional histograms, by considering the regions that are directly related to each other during the pain process. Following [16] two directions of up-down (UD) and left-right (LR) are considered for combining the histograms. These directional histograms are obtained for each modality of RGB, depth, and thermal. Then, they will be separately used to obtain the pain level, which is explained in the next section.

8.4.4. Pain recognition

In the previous subsection two histograms where obtained for the energy orientation of facial regions of each modality, resulting in six directional histograms of energy. The two histograms of the RGB modality are combined by:

$$PI_{RGB} = \sum_{i=1}^{3} w_{R_{i_{UD}}} A_{R_{i_{UD}}} + \sum_{i=1}^{3} w_{R_{i_{LR}}} A_{R_{i_{LR}}} \tag{7}$$

in which PI_{RGB} is the pain index in RGB modality and $A_{R_{i_{UD}}}$ and $A_{R_{i_{LR}}}$ are defined as the integrals of *UD* and *LR* for the ith region (i = 1; 2; 3), respectively [16]:

$$A_{R_{i_{UD}}} = \sum_{t=1}^{n} UD_t \qquad A_{R_{i_{LR}}} = \sum_{t=1}^{n} LR_t \tag{8}$$

where n is the number of the frames in the video. Similarly pain indexes are determined for the other two modalities, resulting in PI_D and PI_T, representing pain indexes obtained for depth and thermal modalities, respectively. These three pain indexes are then fused together to recognize the pain level using:

$$PI = w_{RGB}.PI_{RGB} + w_D.PI_D + w_T.PI_T \tag{9}$$

where w_{RGB}, w_D, w_T are the weights associated to corresponding modalities, and PI is the fused pain index. It should be noted that $w_{RGB} + w_D + w_T = 1$. In the following subsection, it is explained how P I is used to determine the pain level based on experimentally found thresholds.

8.5. Experimental results

In order to present the results, we first discuss the setup and data considered for the experiments, and evaluation measurements and parameters.

8.5.1. Setup and data

12 healthy elderly volunteers (all females) between the ages of 66 and 90 years (mean age 73.6 years) participated in the study. Participants were screened with an interview prior participation to exclude conditions that could affect pain perception and pain report. Exclusion criteria were, if the participant reported the presence of severe on-going pain, neuropsychological and psychiatric disorders, diabetes, or had signs of a rheumatic or arthritic disease, especially on the neck/shoulders. During the interview, subjects were also tested with the Mini Mental State Examination (MMSE) in order to ensure intact cognitive capabilities.

All subjects were pain-free and none of them had taken any analgesic or sedative for at least 48 hours prior to the experiment. The study protocol was approved by the regional ethics committee. Experimental pressure pain was applied on the subjects' trapezius muscle. Eight stimuli of different intensities: No-Pain, Light-Pain, Moder-ate Pain and Strong Pain were applied on left and right trapezius muscles of the par-ticipants.

An electronic handheld pressure algometer (Somedic AB, Stockholm, Sweden) was used to produce noxious mechanical pressure (figure 8.2.b). A force gauge fitted with a rubber disk with a surface of 1 cm2 was used in this study. Pain and no-pain stimuli were determined for each subject on the base of the individual pain detection thresh-old (PDT). Pain stimuli were calculated as follow: No-Pain: 0.2 X PDT, Light Pain: 1.10 X PDT, Moderate Pain: 1.30 X PDT, and Strong Pain: 1.5 X PDT.

Subjects' pain self-reports were recorded using a numerical rating scale (NRS) that measured the perceived intensity of the stimulation. The NRS ranges from 0 (no pain) to 10 (the worst pain you can imagine). Participants NRS was recorded after each stimulus.

During each pain and no-pain stimulation subjects face was video-recorded in order to identify specific pain behaviors on the participants face. During the process the subjects were filmed using a device capturing all the three modalities of RGB, depth, and thermal. The first two modalities were recorded by a Microsoft Kinect for Win-dows V2 device and the thermal modality by an AXIS Q1921 thermal camera. Setup is shown in figure 8.2.c).

Figure 8.3 shows a test subject in these three modalities during our experimental pro-cess. The first, second, and third rows show the RGB, depth, and thermal modalities of this test subject.

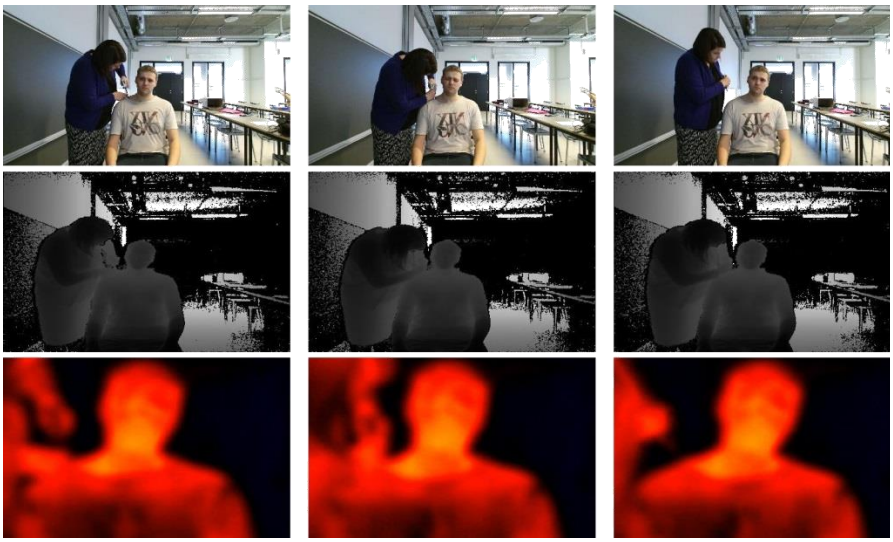


Figure 8.3: RGB (first row), depth (second row), and thermal (third row) of a test subject during the experimental process: first column (before introducing pain), second column (under pain), and third column (after the pain).

8.5.2. Evaluation measurements and parameters

Having obtained pain index *PI* for each of the subjects using Eq. 9, if this pain is smaller than 1, it is considered as no-pain, if it is between 2 and 5, it is considered as a weak pain, and if it is larger than 6, it is considered as a strong pain. These thresholds were experimentally defined in agreement with the team of psychologists. Furthermore, the weights w_{RGB}, w_D and w_T in Eq. 9 have been set via cross-validatio, being 0.6, 0.35, and 0.05, respectively. The obtained weights indicate that RGB provide the primary source of information, followed by depth features and finally by the thermal ones. The obtained weights that are greater than zero show that the use of all three modalities can be useful and complementary for pain recognition.

8.5.3. Results and discussion

Table 8.1 shows the results of the proposed system compared to [16], which considers the same recognition approach only using the RGB modality. One can observe that the proposed system achieves high recognition rates for the three levels of pains, improving the results provided in [16], and showing the benefits of the multimodal approach over just considering RGB data.

Tabel 8.1, Comparing the results of the proposed system against the system of [16] applied to our RGB-D-T facial data base

Semantic Ground Truth	Pain Index Ground Truth	Number of Frames	System of [16] (in %)	Proposed System (in %)
No Pain	0 and 1	757	72	88
Weak	2,3,4,5	427	79	87
Strong	>6	1204	76	76
		sum: 2388	**weighted avg:**75.26	**weighted avg:** 81.77

More specifically, one can see in Table 1 that for the first two levels of pain (no pain and weak pain) the proposed system outperforms the previous system of [16] with a large margin of 16% and 8%, respectively. However, there is no difference in detecting pains of the strong level among the two systems for the considered data. It is mainly produced because for the strong level of pain almost all the details and changes of the facial expressions can be observed in the RGB modalities, and thus adding the other two modalities, at least for the collected video sequences, do not provide any further improvement. However, for the two levels of no pain and weak pain, depth and thermal features are useful to complement visual information from RGB modality and extract more subtle visual features, being useful to discriminate among categories with lower interclass variability.

Finally, figure 8.4 shows the results of the proposed system against the RGB-based one of [16] for a small clip within a sequence. Once can observe that the results of our multimodal system is closer to the ground truth compared to the results of [16].

8.6. **Conclusion and future works**

Pain is a temporal process that can usually be detected from facial images. The proposed system in this paper uses a spatiotemporal approach using a filter which extracts released energies of facial pixels in three modalities, RGB, depth, and thermal, and groups them into histograms of orientations for different facial regions. The integrals of each of these histograms of orientations over time are then used to find a pain index for each modality. These different pain indexes are then fused into a final pain index. The experimental results on a group of 12 elderly people show that the proposed system can accurately detect the pain and recognize its level into three classes of no-pain, weak and strong pain, improving results of single RGB sequence analysis by more than 6%.

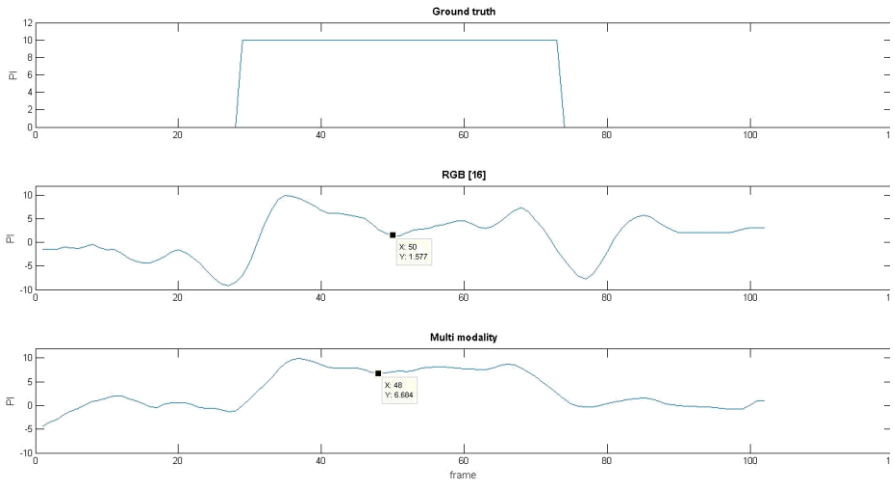


Figure 8.4: Comparing the results of the proposed system against that of [16] against the ground truth.

8.7. **References**

[1] Z. Ambadar, W. Schooler, and J. Cohn. The importance of facial dynamics in interpreting subtle facial expressions. Psychological Science, 16(5):403–410, 2005.

[2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A metaanalysis. Psychological Bull., 111(2):256–274, 1992.

[3] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, M. Prkachin, and P. E. Solomon. The painful face–pain expression recognition using active appearance models. Image and vision computing, 27(12):1788–1796, 2009.

[4] P. Bateson. Assessment of pain in animals. Animal Behaviour, 52:827–839, 1991.

[5] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In CVPR, pages 545–552. IEEE, 2011.

[6] S. Brahnam, C. F. Chuang, F. Y. Shih, and M. R. Slack. Machine recognition and representation of neonatal facial displays of acute pain. Artificial Intelligence in Medicine, 36(3):211–222, 2006.

[7] K. Cannons and R. Wildes. The applicability of spatio-temporal oriented energy features to region tracking. IEEE Trans. Pattern Anal. Mach. Intell., 36(4):784–796, 2014.

[8] S. Coffman, Y. Alvarez, M. Pyngolil, R. Petit, C. Hall, and Smyth. Nursing assessment and management of pain in critically ill children. Heart Lung, 26(221), 1997.

[9] J. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. Journal of Wavelets, Multiresolution & Information Processing, 2(2):121–132, 2004

[10] A. R. Damasio. The feeling of what happens: Body, emotion and the making of consciousness. 1999.

[11] P. Ekman and E. Rosenberg. What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system, 2nd edn. Oxford University Press, London, 2005.

[12] T. Hadjistavropoulos. Social influences and the communcation of pain. Pain: psychological perspectives, page 87, 2004.

[13] Z. Hammal and J. F. Cohn. Automatic detection of pain intensity. In ACM-ICMI, pages 47–52. ACM, 2012.

[14] Z. Hammal and J. F. Cohn. Towards multimodal pain assessment for research and clinical use. In Roadmapping the Future of Multimodal Interaction Research, Business Opportunities and Challenges, pages 13–17. ACM, 2014.

[15] Z. Hammal and M. Kunz. Pain monitoring: A dynamic and context-sensitive system. Pattern Recognition, 45(4):1265– 1280, 2012.

[16] R. Irani, K. Nasrollahi, and T. B. Moeslund. Pain recognition using spatiotemporal oriented energy of facial muscles. In Computer Vision and Pattern Recognition Workshop, 2015 IEEE Conference on, pages 679–692. IEEE, 2015.

[17] W. Janig. The sympathetic nervous system in pain. Eur. J. Anaesthesiol, (10):53–60.

[18] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. Advances in Visual Computing, pages 368–377, 2012.

[19] L. I. Kuncheva. Combining Pattern Classifier: Methods and Algorithms. John Wiley & Sons, 2004.

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In Computer Vision–ECCV 2012, pages 679–692. Springer, 2012.

[21] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. In Proceedings of the 9th international conference on Multimodal interfaces, pages 15–21. ACM, 2007.

[22] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. ICV, 27(12):1797–1803, 2009.

[23] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and M. Prkachin. Automatically detecting pain using facial actions. In Affective Computing and Intelligent Interaction, pages 1–8. IEEE, 2009.

[24] P. McGrath. Pain in children: nature, assessment and treatment. Guildford Press, New York.

[25] N. B. Patel. Physiology of pain. Guide to pain management in low resource settings, page 13, 2010.

[26] K. M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. Pain, 51(3):297–306, 1992.

[27] K. M. Prkachin. Assessing pain by facial expression: facial expression as nexus. Pain Research & Management: The Journal of the Canadian Pain Society, 14(1):53, 2009.

[28] K. M. Prkachin and K. D. Craig. Expressing pain: The communication and interpretation of facial pain signals. Journal of Nonverbal Behavior, 19(4):191–205, 1995.

[29] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. Pain, 139(2):267–274, 2008.

[30] K. M. Prkachin, P. E. Solomon, and J. Ross. Underestimation of pain by healthcare providers: towards a model of the process of inferring pain in others. CJNR (Canadian Journal of Nursing Research), 39(2):88–106, 2007.

[31] O. Rudovic, V. Pavlovic, and M. Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. Advances in Visual Computing, pages 234– 243, 2013.

[32] J. A. Russell, J. Bachorowski, and J. Fernandez Dols. Facial and Vocal Expressions of Emotion. 54:329–349, 2003.

[33] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Computer Vision Workshops (IC-CVW), 2013 IEEE International Conference on, pages 397– 403. IEEE, 2013.

[34] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi automatic methodology for facial landmark annotation. In CVPR Workshops (CVPRW), 2013, pages 896–903. IEEE, 2013.

[35] K. R. Scherer. Appraisal theory. Handbook of cognition and emotion, pages 637–663, 1999.

[36] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn. Spontaneous versus posed facial behavior: automatic analysis of Brow actions. Proc eight intl conf multimodal interfaces (ICMI06), pages 162–170, 2006.

[37] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001, volume 1, pages I–511. IEEE, 2001.

[38] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. T. Multimedia, 12(7):682–691, 2010.

[39] P. Werner, A. Al-Hamadi, and R. Niese. Pain recognition and intensity rating based on comparative learning. In ICIP, pages 2313–2316. IEEE, 2012.

[40] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and C. Traue. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In BMVC, pages 119–1, 2013.

[41] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and C. Traue. Automatic pain recognition from video and biomedical signals. In ICPR, pages 4582–4587. IEEE, 2014.

[42] A. C. d. C. Williams. Facial expression of pain, empathy, evolution, and social learning. Behavioral and brain sciences, 25(04):475–480, 2002.

[43] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In CVPR, pages 532–539. IEEE, 2013.

[44] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR.

Chapter 9

Application of Automatic Energy-based Pain Recognition in Functional Electrical Stimulation

R. Irani, D. Simonsen, O. K. Andersen, K. Nasrollahi, and T. B. Moeslund,

This paper has been published in

Internatinal J. Integr. Care, vol. 15, no. 7, pp. 1–2, 2015.

The layout has been revised

9.1. Abstract

Purpose: Validate and test a method for automatic detection of painful electrical stimulation using computer vision techniques.

Context: In rehabilitation of stroke patients with motor deficits, functional electrical stimulation (FES) of muscles and nerves is a method used for activating the muscles and assisting body movements [1]. The stimulation can be painful, e.g. if the intensity is too high, electrode positioning is wrong or if the electrode/skin contact becomes poor. This is not desirable as it, most likely, discourages the patient to continue using the system. Therefore, a tele-rehabilitation FES system should incorporate methods for automatic detection of painful stimulation and ask the patient to reassess electrode contact and/or positioning upon detection.

9.2. Methods

Two electrodes (Pals Platinum Round 3.2cm, Axelgaard Ltd., USA) were placed on the forearm, targeting the finger extensor muscles. Initially, the pain threshold (PT) of the subject was assessed, followed by 20 stimulation trains of 5 seconds (30Hz pulse train, 200μs pulse duration) with an intensity of 1.5×PT. Subjects were asked to rate each stimulation interval [0 10] (0=no perception, 3=PT, 10=worst imaginable pain), while being filmed by a Logitech 310 webcam. For automatic detection of the pain level, we applied the algorithm of [2], which is based on energy released by facial expressions. The estimated pain intensity was divided into 3 levels (<1=no pain, 2-5=weak and >5=strong pain). Eight healthy subjects (20-42 years) participated in the study. Signed consent was obtained from all subjects and the Declaration of Helsinki was respected. The study was approved by the local ethical committee (N-20130053).

9.3. Results and discussion

The system successfully detected pain responses from changes in facial expressions with an accuracy of 79%. Changes in facial expressions due to stimulation varied greatly between subjects and were not always occurring even though the stimulation was rated as being painful, which might explain why the system did not always detect a behavioral pain response. The proposed work builds on the state of the art pain recognition system of [2] which is able to detect the pain level with accuracy of 75.25% [3]. It means that the overall recognition rate of the pain perception in the proposed system has been improved by 3.74%.

Keywords
Functional electrical stimulation, pain detection, computer vision

9.4. **References**

[1] Schuhfried O et al. Non-invasive neuromuscular electrical stimulation in patients with central nervous system lesions: an educational review. Journal of Rehabilitation Medicine 2012; 44:99-105.

[2] Irani R et al. Pain recognition using spatiotemporal oriented energy of facial muscles. IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2015; 679–692.

[3] Irani R et al. Spatiotemporal Analysis of RGB-D-T Facial Images for Multimodal Pain Level Recognition. IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2015

Chapter 10

Design of 4D Spatiotemporal Oriented Energy Filter for Kinect-based Pain Recognition (Technical Report)

Ramin Irani, Kamal Nasrollahi, Thomas B. Moeslund

10.1. **Abstract**

The following report details the design of 4D energy-based spatio-temporal filters used to recognize the subjects' pain signals collected via Kinec. Normally, since the Kinect data contains two modals (RGB and depth frame sequences), the recognition is restricted to two separate 3D spatio-temporal filters as feature extraction. In the discussion that follows, the proposed 4D spatio-temporal filter can provide a direct measure of facial muscles motions rather than analyzing them separately in each modal. We also propose a 3D landmark alignment and face alignment algorithm, which is necessary to have as an input of the 4D filter.

Key words: *4D spatio-temporal filter, 3D landmarks, Pain recognition, Pose estimation.*

10.2. **Introduction**

Since the facial emotions are highly dynamic phenomenon, dynamic scene analysis provides more accurate system than static scene analysis for analysing emotional recognition tasks. Such a system must be able to track faces and read muscles motion in the scenes. The scenes captured by a camera in a sequence of image frames are used as the input for the system. Each frame represents the facial changes in a particular instance of time and is called spatio-temporal data.

Recently, the spatio-temporal approaches for facial analysing has drawn interest of many researchers [1- 8]. For instance, one of the approaches proposed in [6] applies a 3D steerable separable quadrature filter pairs [9] for detecting the energy of motion of the facial muscles in specific directions (Up, Down, Left and right). Then, utilizes the extracted spatio-temporal information to recognize the pain related feeling from subjects' faces. In the next attempt, the authors applied the same approach for pain recognition to a multi-modality database (RGB-D-T) collected by Kinect and a thermal camera [7]. They applied three individual 3D spatio-temporal filters with three different classifiers that increased processing time.

In this chapter, we expand on this approach by designing a N-dimensional energy-based steerable separable spatio-temporal quadrature filter pairs. The state of the art oriented quadrature filter pairs designed in [9] is a 3D filter which can be employed to estimate the energy of motion on sequence of 2D images. In this report, not only we made use of higher dimensions of filter possible, but also a new mathematical algorithm was introduced. The mathematical algorithm has been constructed based on vector and matrix concepts that make the design and understanding of the filter clearer and more descriptive. To this day, no research has aimed to design the filter in higher dimensions. Another advantage of the proposed method in this work is reduced processing cost. According to the paper [7], thermal changes on the face have less correlation with pain feeling. Therefore, we removed the thermal modality before

applying the proposed spatio-temporal filter to RGB-D data. As a result, only one 4D spatio-temporal filter will be required for pain recognition which brings about significant decrease in the processing cost.

The rest of the report is presented through following sections: Section 3: collecting 3D pain database (RGB-D database) and the device setup, section 4: developing an algorithm for Kinnect-based 3D face alignment, section 5: proposing N-dimensional spatio-temporal steerable separable filter and finally, section 6 concludes the paper.

10.3. **Setup and kinect-based pain database**

The collected database involves thirty volunteers (9 male and 21 female) who participated in the study. All subjects were interviewed prior to the experiment and excluded if they had conditions that could affect pain perception. In order to ensure intact cognitive function capabilities of the subjects, we tested them with the Mini Mental State Examination (MMSE) during the interview. No subject had taken any sedatives or had felt any pain for at least 48 hours prior to experiment. Figure 10.1 shows the environment of the pain experiment and the experiment process. We set the light intensity and position so that fewer shadows are generated on the subjects' face.

For the pain stimulation, we employed a mechanical pressure using a device named "electronic hand-held pressure algometer" (Somedic AB, Stockholm, Sweden). The

Figure 10.1: Pain experiment lab

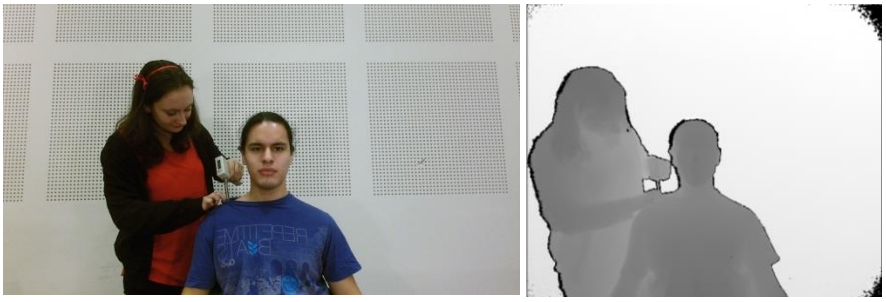
device (figure 2) includes a rubber with a 1_{cm^2} rubber tip, which was placed on the trapezius muscle located on the shoulder. Next, we placed the rubber on the subject's

shoulder and manually applied pressure. We used a digital force gauge fitted with an algometer to record the maximum pressure. The experiment was conducted eight times on both shoulders (4 for each) with four different intensity levels of the pain as follows:

1. No-Pain: 0.2 X PDT,
2. Light Pain: 1.10 X PDT,
3. Moderate Pain: 1.30 X PDT,
4. Strong Pain: 1.5 XPDT

Figure 2, The hand-held pressure algometer used for introducing the pain.

The perceived intensity of the stimulation is measured utilizing a numerical rating scale (NRS). The NRS ranges from 0 to 10. We assign 0 to a status of no pain and 10 to the worst pain that the subjects felt. A Microsoft Kinect V2 sensor, which is placed around 80_{cm} away from the participants (figure 1), filmed their faces during the experimental process. Figure 3 shows a test subject in the two (RGB and Depth) modalities captured by the Kinect.



a. *b.*

Figure 3, a. RGB , b. Depth, of a test subject during the experimental process

10.4. **3D Alignment of kinect-based facial data**

For 3D recognizing of facial region using landmarks-based algorithms, the data collected via kinect should be modeled in a 3D scene. Articles [6] and [7] successfully applied 2D landmark-based face detection on sequences of RGB images in pain recognition algorithm. This section aims to provide a procedure which detects, models and finally aligns 3d facial regions on Kinect-based facial data utilizing the applied 2D face detection algorithm in [6]. To do so, registration of both RGB and depth modalities is necessary [7]. Therefore, the first step in the current research work starts with assigning the corresponding points in the two modal images such that one involves color information of each pixel and another one involves the depth information for each corresponding pixel. Registration processes were done utilizing Look-up tables created by the Kinect software and a procedure described in [10]. A 3D landmark localization approach is followed simply by projecting each pixel from the images with a 2D coordinate system to a 3D world coordinate system using the following equation:

$$X_w = z_d . R . S . x_p \tag{1}$$

$$\begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = z_d . R . [I|t] . \begin{bmatrix} {f_x}^{-1} & 0 & -c_x \\ 0 & {f_y}^{-1} & -c_y \\ 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{2}$$

Where:

- u and v are coordinates of each landmark versus pixel in RGB images.
- f_x and f_y are focal lengths in x and y directions,
- c_x and c_y are coordinates of principal points.
- z_d is Depth of landmarks obtained from Depth images and
- x_w , y_w , z_w are world coordinates of the face on 3D space.
- R is a rotation matrix which will be utilized for landmarks alignment in the next step.
- t is Translation Vector which, like *R*, will be utilized for landmarks alignment in the next step.

Focal lengths and principal points of the employed Kinect, determines from table 10-1. The values are calculated based on the calibration procedure introduced in [11].

Table 10-1, Standard values of focal length and principal points in Kinect v2

	Focal length	**Principal point**
u *direcion*	$f_x = 367.5608$	$c_x = 256.9131$
v *direcion*	$f_y = 367.5608$	$c_y = 207.8238$

The motion of the landmarks on the space is affected due to two different functions: facial muscles motion (soft motion) and head pose (rigid motion). For analyzing motions of the facial muscles in pain recognition, we need to align the landmarks in order to exclude rigid motions. It is done with rotating and translating of all the head poses to the first scene as a reference using the rotation matrix *R* and translation vector *t* in the equation 1. *R* and *t* is obtained due to the fact that the landmarks on the eyes' corners and those on the nose are more stable than other landmarks in muscle movement. Therefore, they are useful when estimating the *R* and *t*. After estimating the *R* and *t*, we can apply it to the rotation and translation of the rest of the landmarks. They are calculated based on the method proposed in [12] as follows:

$$[U\ S\ V\ ] = SVD(\sum_{i=1}^{N}(P_r^i - G_r).(P_s^i - G_s)^T) \tag{3}$$

$$R = V.U^T \tag{4}$$

$$t = G_s - R.G_r \tag{5}$$

Where:

- P_r^i and P_s^i coordinates of i_{th} landmarks in the reference and the rest of 3D facial pose respectively,
- G_s and G_r are given:

$$G_s = \frac{1}{N}.\sum_{i=1}^{N} P_s^i,$$
$$G_r = \frac{1}{N}.\sum_{i=1}^{N} P_r^i, \tag{6}$$

- *R* and *t* required Rotation and translation matrix.

Figure 10.4 illustrates the performance of the alignment method discussed above. Figure 10.4.a shows the position of the landmarks on the 253_{th} facial pose vs. the reference landmarks (on the first facial pose). These landmarks successfully are aligned as shown on the figure 10.4.b. The black circles on the figures mark those landmarks applied for calibration in order to obtain *R* and *t*.

The obtained extrinsic matrix ($[R \mid t]$) can be applied to align the texture as well as the landmarks. It is necessary to remove textures around the facial areas due to the subjects' facial regions being used to recognize pain. This can be done by creating a binary mask for facial area. So we suppose C is a closed contour on the x-y plane

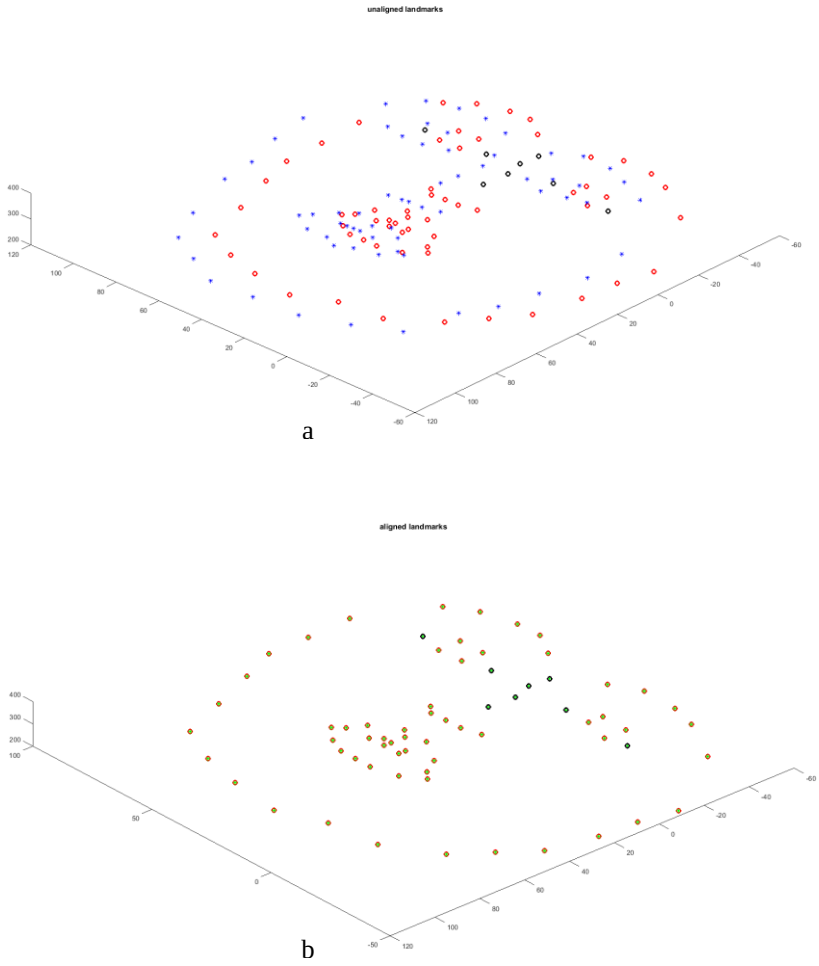


Figure 10.4: a. Landmarks of a facial pose # 253 for a subject (blue stars), b. The Landmarks after alignment (green stars). Red circles on the both figure marks reference landmarks.

that is the projection of a curve passes through the landmarks around the face (figure 10.5). Then to create the Mask, we assign 1 to all points which are inside the contour *C* as follows:

$$Mask = \begin{cases} 1, & P'(x,y) \in C \\ 0, & Otherwise \end{cases} \quad (7)$$

where $P'(x,y)$ is projection of point P on the face surface with coordinate (x, y, z). Figure 5.b. shows the *Mask* with regard to equation 7. All the points in the face

region which are separated by the *Mask* can be aligned along with the landmarks using the *R* and *t* matrix in equations 4 and 5. As a last step of the face alignment, we warp the aligned faces based on the algorithm proposed in [13]. This is done by assigning the intensities in RGB image (which already had been registered with the depth image) to the corresponding points after alignment. After interpolation, the results are as illustrated on figure 10.6.

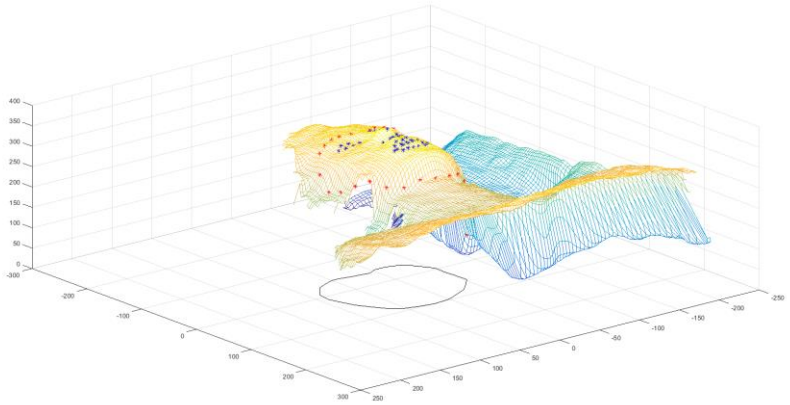


Figure 10.5: Red points the landmarks that surrounded the face and the black line on the x-y plane is the contour of the curve that connects the red landmarks.

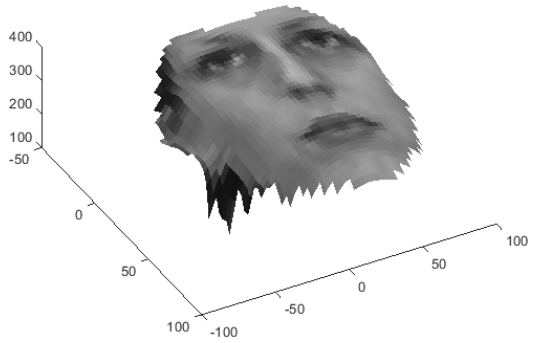


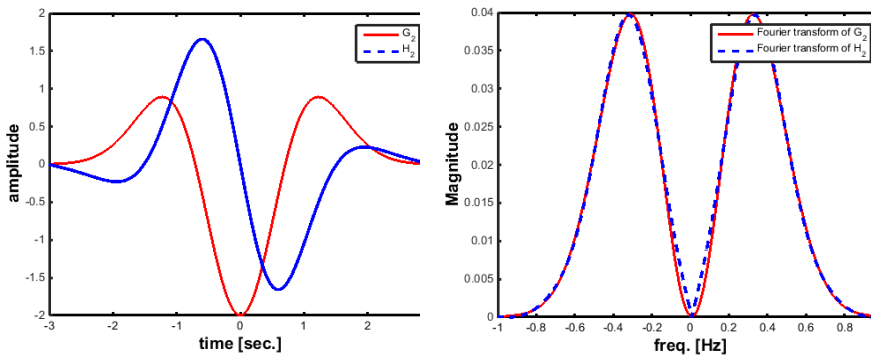
Figure 10.6: The face region which is separated in figure 5 after warping.

10.5. *N-D* Spatio-temporal steerable separable filter

In applied Mathematics, a steerable filter is described as an oriented filter, which expressed via the linear combination of a set of fixed functions. These fixed functions are known as "basis filters". A steerable filter provides an output as a function of orientation [14]. Knutsson et al in [15], proposed a method to design a steerable filter with synthesizing "quadrature pairs" for orientation detection. Quadrature pairs are defined as a pair of functions with similar frequency responses but 90° difference in the phase. Figure 10.7 shows the second derivative of Gaussian in one dimension and its Hilbert transform. According to the figure, such pairs are independent of the phase and allow the filters to be synthesized for a given frequency response and arbitrary phase [16]. Gaussian derivatives are popular functions in many image processing tasks [17-20]. It would therefore be useful to apply their quadrature pairs as a steerable filter in orientation analysis of early vision systems. Article [9] designed a steerable quadrature pair based on the frequency response of the second derivative of a Gaussian G_2 and described the optimal use of the filter to measure the orientation in a particular direction Θ, by squared output of a quadrature pair of the designed filter which steers in angle Θ. This spectral power is called "Orientation energy", $E(\Theta)$. Using the n_{th} derivative of a Gaussian quadrature pair, we have:

$$E_n(\Theta) = [G_n^{\Theta}]^2 + [H_n^{\Theta}]^2 \tag{8}$$

Where G_n^{Θ} is, n_{th} derivative of the Gaussian function in direction of Θ and H_n^{Θ} is its Hilbert transform.



A b

Figure 10.7, a. 2nd derivative of Gaussian (G_2) in 1D and its Hilbert transform (H_2), b. Magnitudes of Fourier transform of G_2 and H_2

The basis filters in the equation 8 are a function of Θ which presents a complex computation in three or more dimensional space. Design of separable basis functions in

Cartesian coordinate system is one way to deal with the complexity problem. It is also very useful in the most applications of machine vision, which are a function of (x, y, z ...). In this section, we aim to design an *N*-Dimensional (*N*-D) spatio-temporal steerable separable filter and next we will apply it in *4*-D space which is useful in the Kinect based spatio-temporal phenomena e.g. pain.

10.5.1. Preliminary Mathematics

In this sub-section, we propose an analytical analysis for *N*-D steerable separable G_2 and its Hilbert transform H_2 in desired orientation. Then we apply them on the equation 8 in order to design an energy-based steerable separable quadrature pairs filter:

$$E_2(\underline{X}) = [G_2^{\theta}(\underline{X})]^2 + [H_2^{\theta}(\underline{X})]^2 \tag{9}$$

10.5.1.1 Steerable separable Gaussion filter Design

We assume that the function to be steered is second derivative of a Gaussian function as follows:

$$G_2^R = \vartheta_{go}.P(x')e^{-\underline{X}^T.\underline{X}} \tag{10}$$

where:

- $P(x')$ is second order polynomial which described as:

$$P(x') = -\frac{\delta^2 \underline{X}^T.\underline{X}}{\delta x^2} \tag{11}$$

- $\underline{X} = \{x_i\}, i = 1,2, \dots, N$ and involves coordinate of each point on the filter in Cartesian system.
- G_2^R is second derivative Gaussian function which rotated by transformation R such that its axis rotational symmetry x' [9] is along arbitrary direction $\underline{\Theta} = \{\theta_i\}\ i = 1,2, \dots, N-1$. Where $\theta_i$is the angle between x' and i_{th} orthogonal surface on Cartesian coordinate system.
- ϑ_{go} is a normalization constant so that the integral over the square of the function G_2^{θ}equals one.

Here, we want to prove that G_2^R in arbitrary direction $\underline{\Theta}$ can be decomposed to combination of separable basis functions in Cartesian coordinate system as follows:

$$G_2^R = \vartheta_{go}.\underline{K}_g^T(\underline{\Theta}).\underline{B}_g(\underline{X}) = \vartheta_{go}.\sum_{j=1}^{M} k_{g_j}(\underline{\Theta}).b_{g_j}(\underline{X}) \quad (12)$$

Such that $\{k_{g_j}\}$ are coefficients, which are function of $\underline{\Theta}$ direction of x' and $\{b_{g_j}\}$ are basis functions decomposed into *N*, 1-D functions as follow:

$$b_{g_j}(\underline{X}) = \prod_{i=1}^{N} f_{g_{ji}}(x_i) \quad (13)$$

We presume $\{\rho, \theta_i\}_{,i=1,2,\dots,N-1}$ are spherical coordinate of unit vector along the axis symmetry x' with $\rho = 1$ then using spherical system to Cartesian system conversion [21], *x'* and direction cosines vector $\underline{A}$ can be obtained as follows:

$$x' = \underline{A}^T.\underline{X} = \sum_{i=1}^{N} a_i.x_i \quad where: \quad (14)$$

$\underline{A} = \{a_i\}$:

$$a_i = \begin{cases} \prod_{j=1}^{N-1} \sin(\theta_j) & if \quad i = 1 \\ (\prod_{j=1}^{N-i} \sin(\theta_j)).\cos(\theta_{n-i+1}) & if\ 2 < i < N \\ \cos(\theta_1) & if \quad i = N \end{cases}$$

and $\|\underline{A}\| = 1$

After substituting equation 8 into 7.1 and expanding, we have:

$$P(x') = \left(4.\left(\underline{A}^T.\underline{X}\right)^2 - 2\right) \quad (15)$$

Since $\underline{A}^T.\underline{A} = \underline{A}^T.I.\underline{A} = 1$ then we have:

$$P(x') = 2.\underline{A}^T\left(2.\underline{X}.\underline{X}^T - I\right).\underline{A} \quad (16)$$

Note that $\left(2.\underline{X}.\underline{X}^T - I\right)$ in equation 16 is Hermitian matrix and $P(x')$ is Rayleigh quotient.

$$P(x') = \frac{2.\underline{A}^T\left(2.\underline{X}.\underline{X}^T - I\right).\underline{A}}{\underline{A}^T.\underline{A}} \quad (17)$$

After expansion of equation 16:

$$P(x') = 2.\left\{\sum_{q=1}^{N}\sum_{p=1}^{N} a_p.a_q\left(2.x_p.x_q - \kappa\right)\right\} \quad where$$
$$\kappa = \begin{cases} 1 & p = q \\ 0 & otherwise \end{cases} \tag{18}$$

$$P(x') = 2.\sum_{i=1}^{N} a_i^2.(2.x_i^2 - 1) + 8.\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i.a_j.x_i.x_j \tag{19}$$

By substituting 19 in 10, G_2^R is given:

$$G_2^R = 2.\vartheta_{go}.\sum_{i=1}^{N} a_i^2.(2x_i^2 - 1).e^{-\sum_{j=1}^{N} x_j^2} + 8.\vartheta_{go}.\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i.a_j.x_i.x_j.e^{-\sum_{i=1}^{N} x_i^2} \tag{20}$$

Then the equation 20 can be simplified as:

$$G_2^R = \sum_{s=1}^{M} \vartheta_g.k_{g_s}(\underline{\Theta}).b_{g_s}(\{x_i\}) = \underline{K_g^T}(\underline{\Theta}).\underline{B_g}(\underline{X}) \tag{21}$$

such that:

$$k_{g_s}(\underline{\Theta}) = \vartheta_g.a_i.a_j : \quad i = 1,2,\dots,N \ \& \ j \geq i \tag{22}$$

$$\vartheta_g = \begin{cases} 2.\vartheta_{go} & i = j \\ 8.\vartheta_{go} & i < j \end{cases} \tag{23}$$

$$b_{g_s}(\{x_i\}) = \begin{cases} (2.x_i^2 - 1).\prod_{s=1}^{N} e^{-x_s{}^2} & if\ i = j \\ x_i.x_j.\prod_{s=1}^{N} e^{-x_s{}^2} & if\ i > j \end{cases} \tag{24}$$

Equation 21, 22, 23 and 24 confirm that the steerable second derivative Gaussian function can be explained by combination of separable basis functions in Cartesian coordinate system. According to equation 24, each basis function is written by *N* multiplication of one-dimensional function $f_{g_i}(x_i)$ corresponded direction *i*.

$$b_{g_s}(\{x_i\}) = \prod_{i=1}^{N} f_{g_i}(x_i) \tag{25}$$

where f_{g_i} is one of the below options:

- $(2.t^2 - 1).e^{-t^2}$
- $t.e^{-t^2}$
- e^{-t^2}

At the end, to calculate the number of basis functions *M* we assume T(.) as a function that gives the number of independent terms in a summation then considering equation 20 we have:

$$\begin{aligned} M_{G_2} = T\big(P(x')\big) &= T\left(\sum_{i=1}^{N} a_i^2.(2.x_i^2 - 1) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i.a_j.x_i.x_j\right) \\ &= T\left(\sum_{i=1}^{N} a_i^2.(2.x_i^2 - 1)\right) + T\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i.a_j.x_i.x_j\right) \\ &= T\left(\sum_{i=1}^{N} a_i^2.(2.x_i^2 - 1)\right) + \sum_{i=1}^{N-1} T\left(\sum_{j=i+1}^{N} a_i.a_j.x_i.x_j\right) \end{aligned} \tag{26}$$

$$M_{G_2} = N + (N-1).N/2 = N.(N+1)/2 \tag{27}$$

10.5.1.2 Hilbert transform of the second derivative Gaussian

The Hilbert Transform of the second derivative of the *N*-D Gaussian (H_2^R) is obtained by finding the least squares fit to a third order polynomial multiplied by Gaussian which satisfied level of approximation by 1% ratio of total error power to the total signal power [9]. Then we have:

$$H_2^R = H[G_2^R] = \vartheta_{ho}.Q(x')e^{-\underline{X}^T.\underline{X}} \tag{28}$$

Where ϑ_{ho} is normalization coefficient and

$$Q(x') = \left(\left(\underline{A}^T.\underline{X}\right)^3 - 2.254.\underline{X}\right) \tag{29}$$

Similar to the previous subsection, we want to describe H_2^R by combination of separable basis functions in Cartesian coordinate system as follows:

$$H_2^R = \underline{K_h^T}(\underline{\Theta})\underline{B_g}(\underline{X}) = \sum_{j=1}^{M} k_{h_j}(\underline{\theta}).b_{h_j}(\underline{X}) \tag{30}$$

Such that $\{k_{h_j}\}$ are coefficients function of $\underline{\Theta}$ the direction of axis symmetry x' and $\{b_{h_j}\}$ are desired basis functions that is defined as:

$$b_{h_j}(\underline{X}) = \prod_{i=1}^{N} f_{h_{ji}}(x_i) \tag{31}$$

Expanding the equation 28 gives:

$$H_2^R = \vartheta_{ho}.\underline{A}^T.\left(\left(\underline{X}.\underline{X}^T - 2.254.I\right).A.\underline{A}^T\right).\underline{X}.e^{-\underline{X}^T.\underline{X}} \tag{32}$$

$$\begin{aligned} H_2^R &= (\vartheta_{ho}.\sum_{i=1}^{N} a_i^3.x_i.(x_i^2 - 2.254).e^{-\sum_{j=1}^{N} x_j^2} \\ &+ 3.\vartheta_{ho}.\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i^2.a_j(x_i^2 - 0.751).x_j\ .e^{-\sum_{j=1}^{N} x_j^2} \\ &+ 3.\vartheta_{ho}.\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i^2.a_j.x_i^2.x_j\ .e^{-\sum_{j=1}^{N} x_j^2} \\ &+ 6.\vartheta_{ho}.\sum_{i=1}^{N-2}\sum_{j=i+1}^{N-1}\sum_{k=i+2}^{N} a_i.a_j.a_k.x_i.x_j.x_k\ .e^{-\sum_{j=1}^{N} x_j^2}) \end{aligned} \tag{33}$$

Equation 33 can be written as vector form:

$$H_2^R = \sum_{s=1}^{M} \vartheta_h.k_{h_s}(\underline{\Theta}).b_{h_s}(\{x_i\}) = \underline{K_h^T}(\underline{\Theta}).\underline{B_h}(\underline{X}) \tag{34}$$

such that:

$$k_{h_s}(\underline{\Theta}) = \vartheta_h . a_i . a_j . a_k : i = 1,2, \dots, N \ \& \ i \leq j \ \& \ j \leq k \tag{35}$$

$$\vartheta_h = \begin{cases} \vartheta_{ho} & if\ i = j\ or\ k\ and\ j \neq k \\ 3.\vartheta_{ho} & if\ i \neq j = k \\ 6.\vartheta_{ho} & if\ i = j = k \end{cases} \tag{36}$$

$$b_{h_s}(\{x_i\}) = \begin{cases} x_i.(x_i^2 - 2.254).\prod_{s=1}^{N} e^{-x_s{}^2} & if\ k = j = i \\ (x_i^2 - 0.751).x_j.\prod_{s=1}^{N} e^{-x_s{}^2} & if\ j > i = k \\ x_i^2.x_j.\prod_{s=1}^{N} e^{-x_s{}^2} & if\ k > i = j \\ x_i.x_j.x_k.\prod_{s=1}^{N} e^{-x_s{}^2} & if\ k > j > k \end{cases} \tag{37}$$

Equations 34, 35, 36 and 37, concludes that Hilbert transform of steerable second derivative Gaussian function in Cartesian coordinate system can be expanded as summation of M_{H_2} separable basis functions are written by N multiplication of $f_{h_i}(x_i)$ which is selected from below options:

- $(t^3 - 2.254.t).e^{-t^2}$
- $(t^2 - 0.751).e^{-t^2}$
- $t.e^{-t^2}$
- e^{-t^2}

The number of basis functions M_{H_2} is obtained by using the function $T(.)$ as follow:

$$M_{H_2} = T\big(Q(x')\big) = T\left(\sum_{i=1}^{N} a_i^3.x_i.(x_i^2 - 2.254) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i^2.a_j(x_i^2 - 0.751).x_j + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} a_i^2.a_j.x_i^2.x_j + \cdots\right.$$

$$\left. \ldots \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=i+2}^{N} a_i . a_j . a_k . x_i . x_j . x_k \right) \tag{38}$$

$$\begin{aligned} M_{H_2} = \; & T\left(\sum_{i=1}^{N} first\ item \right) + T\left(\sum_{i=1}^{N} \sum_{j=i+1}^{N} second\ item \right) \\ & + T\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} third\ item \right) \\ & + T\left(\sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=i+2}^{N} forth\ item \right) \end{aligned} \tag{39}$$

$$\begin{aligned} M_{H_2} = \; & N + (N-1).N/2 \; + (N-1).N/2 \\ & + \; (N^2 - 3.N + 4).(N-2)/4 \end{aligned}$$

$$M_{H_2} = N^2 + \; (N^2 - 3.N + 4).(N-2)/4 \tag{40}$$

10.5.2. Design of energy-based filter in 4D spatio-temporal space

Energy-based steerable separable spatio-temporal filter in 4-D space is achieved by substituting $N = 4$ in the filter proposed in sub-section 10.5.1. In this case, the input vector of the filter $\underline{X}$ includes position of objects (subjects' face) and corresponding time value. In this section, for clarifying the implementation of the filter, we categorize the design process into the three stages: input data, initializing quadrature pairs and energy-based filter generation. These stages are illustrated in figure 10.8.

10.5.2.1 Input data

Kinect-based spatio-temporal analysis of the facial pose sequences given in section 10.4 deals with geometrical position of face in a specific time *t*. Therefore, the input vector $\underline{X}$ in the designed filter is defined as follow:

$$\underline{X} = [x, y, z, t]^T \in \mathcal{R}^3 \times \mathcal{N} \tag{41}$$

Where x, y, z are coordinate of a point P on the facial pose in the time $t = n.T$ (n_{th}frame) and T denotes time sampling. In addition, $\mathcal{R}$ and $\mathcal{N}$ are real and integer space respectively.

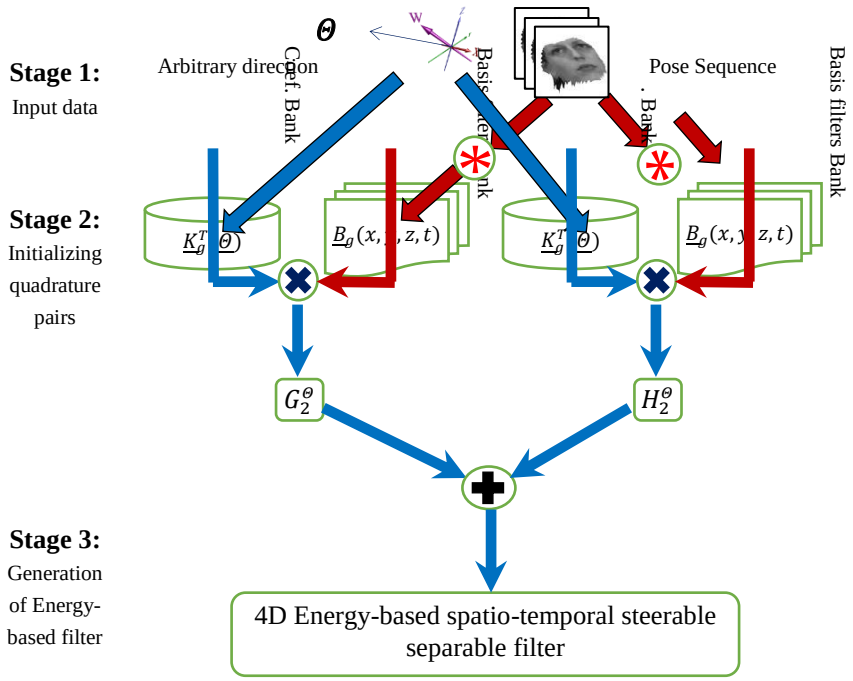


Figure 10.8: Block diagram of a 4D Energy spatio-temporal based filter system

Besides, $\underline{\Theta}$ which is the desired orientation of the axis of symmetry for the filter is also considered as an input that determines the direction cosine vector $\underline{A} = \{a_i\}$ by using the equation 14. It is attained by conversion of the orientation $\underline{\Theta}$ to the Cartesian coordinate system. We assume that spherical components of the desired orientation $\underline{\Theta}$ are $\{\theta_1, \theta_2, \theta_3$ and $\rho = 1\}$, thereby, the components of the direction cosine $[\alpha, \beta, \gamma, \delta]^T$ are given [22]:

$$\begin{aligned} \alpha &= cos(\theta_1).sin(\theta_2).sin(\theta_3) \\ \beta &= sin(\theta_1).sin(\theta_2).sin(\theta_3) \\ \gamma &= cos(\theta_2).sin(\theta_3) \\ \delta &= cos(\theta_3) \end{aligned} \tag{42}$$

10.5.2.2 Initializing quadrature pairs

According to figure 10.8, initialization process is carried out by coefficients and basis functions construction. The coefficients of G_2^{θ} and H_2^{θ} are acquired by substitution of $N = 4$ into equations 22 and 24. These coefficients and their corresponding basis functions set are listed in Tables 12.2 - 12.4. As mentioned earlier, the basis functions are separable functions which are separated to N_{th} = 4_{th} one-dimensional function. These one-dimensional functions are obtained from a set of functions in 25 and 37 where the first set with 3 elements corresponds to G_2^{θ} and second set with 4 elements corresponds to H_2^{θ}. Table 12.2 shows all elements of each set. Table 12.3 summarizes the construction of the G_2^{θ}, in other words, it represents 1D functions contained in each G_2^{θ} basis function used for filtering in one of the directions. In the same way, table 12.4 represents the same for H_2^{θ}. The number of coefficients and corresponding basis functions *M* are calculated using the equation 27 and 40 as shown below:

$$M_{G_2} = N.\frac{N+1}{2} = \frac{4 \times 5}{2} = 10 \tag{43}$$

$$M_{H_2} = N^2 + \ (N^2 - 3.N + 4).(N-2)/4 = 4^2 + \frac{(4^2 - 3 \times 4 + 4).(4-2)}{4} = 20 \tag{44}$$

We can see that the interpolation coefficients of G_2 and H_2 should be 10 and 20 in order.

10.5.2.3 Energy-based filter generation

As seen in figure 8 the steerable separable energy-based filter strength along direction $\underline{\theta}$ is generated by the square output of a quadrature pair of filters G_2 and H_2.They are calculated with substitution of interpolation coefficients and basis functions (summarized in table 12.2 –12. 4) into equations 21 and 34 in sections 10.5.1.1 and 10.5.1.2. Then we have:

$$G_2^{\theta} = \sum_{s=1}^{10} \vartheta_g . k_s(\alpha, \beta, \gamma, \delta) \, . b_s(x, y, z, t) \tag{45}$$

$$H_2^{\theta} = \sum_{s=1}^{20} \vartheta_h . k_s(\alpha, \beta, \gamma, \delta) \, . b_s(x, y, z, t) \tag{46}$$

Where ϑ_{go} and ϑ_{ho} are a normalization constant and equals 0.7351 and 0.7841 respectively. At the end, the 4-D energy-based steerable separable filter results:

$$E_{\alpha,\beta,\gamma,\delta}(x,y,z,t) = [G_{2\,(\alpha,\beta,\gamma,\delta)}^{\theta}(x,y,z,t)]^2 + [H_{2\,(\alpha,\beta,\gamma,\delta)}^{\theta}(x,y,z,t)]^2 \quad (47)$$

Table 10.2 set of 1D filters used in separable basis functions

1D functions for G_2 basis functions		1D functions for H_2 basis functions	
$\mathbf{f_{g1}}$	$(2.t^2-1).e^{-t^2}$	$\mathbf{f_{h1}}$	$(t^3-2.254.t).e^{-t^2}$
$\mathbf{f_{g2}}$	$t.e^{-t^2}$	$\mathbf{f_{h2}}$	$(t^2-0.751).e^{-t^2}$
$\mathbf{f_{g3}}$	e^{-t^2}	$\mathbf{f_{h3}}$	$t.e^{-t^2}$
---	--------	$\mathbf{f_{h4}}$	e^{-t^2}

Table 10.3. Construction of G_2 basis filters and corresponding interpolation coefficients

G_2 Basis filter	coefficients	Applied 1D filter in direction :			
$\boldsymbol{b_{g_s}}$	$\boldsymbol{K_{g_s}(\underline{\Theta})/\vartheta_g}$	x	y	z	t
$\boldsymbol{b_{g1}}$	α^2	f_{g1}	f_{g3}	f_{g3}	f_{g3}
$\boldsymbol{b_{g2}}$	$\alpha.\beta$	f_{g2}	f_{g2}	f_{g3}	f_{g3}
$\boldsymbol{b_{g3}}$	$\alpha.\gamma$	f_{g2}	f_{g3}	f_{g2}	f_{g3}
$\boldsymbol{b_{g4}}$	$\alpha.\delta$	f_{g2}	f_{g3}	f_{g3}	f_{g2}
$\boldsymbol{b_{g5}}$	β^2	f_{g3}	f_{g1}	f_{g3}	f_{g3}
$\boldsymbol{b_{g6}}$	$\beta.\gamma$	f_{g3}	f_{g2}	f_{g2}	f_{g3}
$\boldsymbol{b_{g7}}$	$\beta.\delta$	f_{g3}	f_{g2}	f_{g3}	f_{g2}
$\boldsymbol{b_{g8}}$	γ^2	f_{g3}	f_{g3}	f_{g1}	f_{g3}
$\boldsymbol{b_{g9}}$	$\gamma.\delta$	f_{g3}	f_{g3}	f_{g2}	f_{g2}
$\boldsymbol{b_{g10}}$	δ^2	f_{g3}	f_{g3}	f_{g3}	f_{g1}

Table 1.4. Construction of H_2 basis filters and corresponding interpolation coefficients

H_2 Basis filter	Interpolation coefficients	Applied 1D filter in direction :			
$\boldsymbol{b_{h_s}}$	$\boldsymbol{K_s(\underline{\theta}))/\vartheta_h}$	x	y	z	t
$\boldsymbol{b_{h1}}$	α^3	f_{h1}	f_{h4}	f_{h4}	f_{h4}
$\boldsymbol{b_{h2}}$	$\alpha^2.\beta$	f_{h2}	f_{h3}	f_{h4}	f_{h4}
$\boldsymbol{b_{h3}}$	$\alpha^2.\gamma$	f_{h2}	f_{h4}	f_{h3}	f_{h4}
$\boldsymbol{b_{h4}}$	$\alpha^2.\delta$	f_{h2}	f_{h4}	f_{h4}	f_{h3}
$\boldsymbol{b_{h5}}$	$\alpha.\beta^2$	f_{h3}	f_{h2}	f_{h4}	f_{h4}
$\boldsymbol{b_{h6}}$	$\alpha.\gamma^2$	f_{h3}	f_{h4}	f_{h2}	f_{h4}
$\boldsymbol{b_{h7}}$	$\alpha.\delta^2$	f_{h3}	f_{h4}	f_{h4}	f_{h2}
$\boldsymbol{b_{h8}}$	$\alpha.\beta.\gamma$	f_{h3}	f_{h3}	f_{h3}	f_{h4}
$\boldsymbol{b_{h9}}$	$\alpha.\beta.\delta$	f_{h3}	f_{h3}	f_{h4}	f_{h3}
$\boldsymbol{b_{h10}}$	$\alpha.\gamma.\delta$	f_{h3}	f_{h4}	f_{h3}	f_{h3}
$\boldsymbol{b_{h11}}$	β^3	f_{h4}	f_{h1}	f_{h4}	f_{h4}
$\boldsymbol{b_{h12}}$	$\beta^2\gamma$	f_{h4}	f_{h2}	f_{h3}	f_{h4}
$\boldsymbol{b_{h13}}$	$\beta^2.\delta$	f_{h4}	f_{h2}	f_{h4}	f_{h3}
$\boldsymbol{b_{h14}}$	$\beta.\gamma^2$	f_{h4}	f_{h3}	f_{h2}	f_{h4}
$\boldsymbol{b_{h15}}$	$\beta.\delta^2$	f_{h4}	f_{h3}	f_{h4}	f_{h2}
$\boldsymbol{b_{h16}}$	$\beta.\gamma.\delta$	f_{h4}	f_{h3}	f_{h3}	f_{h3}
$\boldsymbol{b_{h17}}$	γ^3	f_{h4}	f_{h4}	f_{h1}	f_{h4}
$\boldsymbol{b_{h18}}$	$\gamma^2.\delta$	f_{h4}	f_{h4}	f_{h2}	f_{h3}
$\boldsymbol{b_{h19}}$	$\gamma.\delta^2$	f_{h4}	f_{h4}	f_{h3}	f_{h2}
$\boldsymbol{b_{h20}}$	δ^3	f_{h4}	f_{h4}	f_{h4}	f_{h1}

10.6. Conclusion

In this report, we proposed an algorithm for landmark alignment in 3D space using the collected data by Kinect. Next, it is followed by separating face region in 3D data. Additionally, we presented an N-dimensional separable steerable quadrature pair filter and provided a procedure to design a 4D energy-based spatio-temporal filter. This filter can be applied on the pain detection system proposed in [6] where the input of the system should be updated to the 3D Kinect data proposed in section 12.4. The output of the filter and classification process will be the same as the system in [6]. Consequently, the proposed filter and face detection algorithm in this report allows us to upgrade the 2D pain recognition proposed in [6] to 3D pain recognition without requiring an additional filter that was the drawback of the RGB-D-T system in [7].

Having finalized the implementation of the above explained data, the collected data explained in chapter 10.5.2 needs to be analysed using the method proposed in [7] with two separate 3D separable steerable filters, along with the method discussed in this chapter using the new 4D designed separable steerable filter. Comparing the output of both systems in terms of accuracy and processing speed will clarify the advantages of the new approach.

10.7. References

[1] M. Pantic, I. Patras, and M. Valstar, “Learning spatiotemporal models of facial expressions,” in Int’l Conf. Measuring Behaviour 2005, August 2005, pp. 7–10.

[2] S. K. A. Kamarol, M. H. Jaward, J. Parkkinen and R. Parthiban, "Spatiotemporal feature extraction for facial expression recognition," in IET Image Processing, vol. 10, no. 7, pp. 534-541, June 2016.

[3] M. Valstar and M. Pantic, "Fully Automatic Facial Action Unit Detection and Temporal Analysis," 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), 2006, pp. 149-149.

[4] M. F. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 1, pp. 28-43, Feb. 2012.

[5] R. Yang et al., "On pain assessment from facial videos using spatio-temporal local descriptors," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, 2016, pp. 1-6.

[6] R. Irani, K. Nasrollahi and T. B. Moeslund, "Pain recognition using spatiotemporal oriented energy of facial muscles," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 80-87.

[7] R. Irani et al., "Spatiotemporal analysis of RGB-D-T facial images for multimodal pain level recognition," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 88-95.

[8] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal RBM based model for facial expression recognition," Pattern Recognit., vol. 49, pp. 152–161, Jan. 2016.

[9] K. G. Derpanis and J. M. Gryn, "Three-dimensional nth derivative of Gaussian separable steerable filters," IEEE International Conference on Image Processing, 2005, pp. III-553-6.

[10] "aauvap / KinectV2Toolbox — Bitbucket", Bitbucket.org, 2017. [Online]. Available: https://bitbucket.org/aauvap/kinectv2toolbox. [Accessed: 25- Apr- 2017].

[11]"CoordinateMapper Methods", Msdn.microsoft.com, 2017. [Online]. Available: https://msdn.microsoft.com/enus/library/windowspreview.kinect.coordina-te-mapper_methods.aspx. [Accessed: 07- Apr- 2017].

[12] N. Ho., "Finding Optimal Rotation and Translation between corresponding 3D," 10-May-2013. [Online]. Available: http://nghiaho.com/?page_id=671. [Accessed: 07-Apr-2017].

[13] N.A. Gumerov, A. Zandifar, R. Duraiswami, and L.S. Davis, "Structure of Applicable Surfaces from Single Views," Proc. European Conferece on Computer Vision, May 2004, pp. 482–496

[14] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters, " in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 9, pp. 891-906, Sep 1991.

[15] H. Knutsson and G. H. Granlund. "Texture analysis using two-dimensional quadrature filters, " In IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management, pages 206-213, 1983

[16] R. N. Bracewell, The Fourier Transform and Its Applications. New York: McGraw-Hill, 1986.

[17] J. J. Koenderink. "Design for a sensorium," In W. von Seelen, B. Shaw, and U. M. Leinhos, editors, Organization of Neural Networks, pages 185-207. Verlagsgesellschaft mbH, 1988.

[18] J. J. Koenderink. "Operational significance of receptive field assemblies," Biol. Cybern., 58:163-171, 1988.

[19] J. J. Koenderink and A. J. van Doorn. "Representation of local geometry in the visual system," Biol. Cybern., 55:367-375, 1987.

[20] R. A. Young. “Simulation of human retinal function with the Gaussian derivative model,” In Proc. IEEE Computer Society Conferece on Computer Vision and Pattern Recognition, pages 564-569, 1986.

[21] G. W. Collins, “The foundations of celestial mechanics,” The Pachart Foundation dba Pachart Publishing House, 1989.

[22] E. W. Weisstein, Hypersphere, Wolfram MathWorld. [Online]. Available: http://mathworld.wolfram.com/Hypersphere.html. [Accessed: 10-Apr-2017].

Chapter 11

Validation and Test of a Closed-loop Tele-rehabilitation System based on Functional Electrical Stimulation and Computer Vision for Analysing Facial Expressions in Stroke Patients

Daniel Simonsen, Ramin Irani, Kamal Nasrollahi, John Hansen, Erika G. Spaich, Thomas B. Moeslund, and Ole K. Andersen

This paper has been published in

International Conference on NeuroRehabilitation, ICNR, Aalborg, Denmark, 2014, p.p. 741-750

The layout has been revised.

11.1. **Abstract**

The aim of the present study was to validate and test a closed-loop tele-rehabilitation system for training of hand function and analyzing facial expressions in stroke patients. The paper presents the methods for controlling functional electrical stimulation (FES) to assist hand opening and grasping. The main outcome of the FES control was time differences in grip detections performed by the automatic system and by analysis of the output from force sensing resistors. This time difference was in the range of 0 to 0.8 s. Results from analysis of facial expressions were very variable showing that subjects were disgusted, happy and angry during the exercises, which were not in agreement with the observations made during the experimental sessions.

11.2. **Introduction**

Stroke is a leading cause of disability worldwide [1]. Studies on stroke rehabilitation have shown that stroke patients are capable of regaining motor control to some extent by rehabilitative training, especially in the first months post stroke [2, 3]. However, hand function often remains significantly affected [3]. Since proper motor function of the hand, i.e. hand opening and grasping, is related to activities of daily living this has a major impact on the patient's daily life. Therefore, it is of great importance to exploit the time window for rehabilitation post stroke, in order to maximize the outcome of rehabilitation, particularly in relation to hand function. Generally, stroke patients receive intensive rehabilitation subsequent to the acute treatment of the stroke, but the amount of time spent on self-training by the patient in his or hers own home will most likely increase. At home the patient is not supervised and supported by a therapist. This might mean that the patient has an increased risk of performing the exercises wrong or in some cases might not even be able to complete the exercises due to a lack of sufficient motor function. Furthermore, training at distance without continuous supervision obviously makes it difficult for therapists to detect non-spoken social cues (facial expressions/body language), which might provide crucial information about how well the patient mentally complies with the rehabilitation.

Reviews of tele-rehabilitation studies state that the current evidence is insufficient to draw definite conclusions upon the effectiveness of tele-rehabilitation of stroke patients [4, 5]. Common for the studies included in the reviews is that the support given to the patients during self-training is either non-existing or limited to visual feedback. As a consequence it will not make sense for the patients to start using these systems before they can comply with the self-training exercises. By combining a tele-rehabilitation system with a system that can assist the patients in complying both physically and mentally with the self-training exercises a broader range of patients can be targeted and rehabilitation training in the patient's own home might be more efficient.

Functional electrical stimulation (FES) of the muscles is a method for assisting stroke patients in performing functional movements [6]. FES rehabilitation systems are triggered by user input, often by surface electromyographic (EMG) signals, or cyclically according to predefined settings. A meta-analysis by [7], found no significant difference in rehabilitative outcome in stroke patients between use of EMG-triggered FES and conventional care. Although studies show that FES rehabilitation systems might be comparable to conventional care, they are not designed as tele-rehabilitation systems and thus are not suited for use during self-training in the patient's own home. Furthermore, current FES rehabilitation systems require some kind of direct user input for triggering the assistive stimulation, which means that the user cannot solely focus on the execution of the movements during training.

By use of a camera it is possible to monitor the patient's movements during training. By performing real-time analysis on the images captured by the camera it is possible to control FES, thus eliminating the need for direct user input to trigger the electrical stimulations. Furthermore, cameras can also be used for recording facial expressions. Parameters derived from facial expressions are the most effective ones in visual information as they provide clues to recognize the mental state of a person. The majority of the methods reported in the literature use only facial expressions for automatic emotional recognition. These works are mostly based on Charles Darwin's idea [8] which established general principles of expressions and their meanings on the face of both human and animals. In 1978 Ekman [9] defined a new scheme named Facial Action Coding Scheme, which involves 64 basic Action Units (AUs) and combination of AUs representing movement of facial muscles.

In this paper a tele-rehabilitation system for assisting training of hand function and recognizing facial expressions in stroke patients is presented. The system controls FES in a closed loop by a Microsoft Kinect sensor, and records facial expressions by a web camera.

11.3. Methods

11.3.1. Subjects

Four subjects were included in the study. They were aged between 18 to 80 years, previously diagnosed with a cerebrovascular stroke (verified by a MRI scan), had decreased hand function, and were able to sit upright without support. Subjects were excluded if they were pregnant, drug addictive (hash, opioids or other psychedelic drugs), not able to understand the aim of the study and complete the experiment due to cognitive or linguistic deficits, suffering from serious general deterioration, had a pacemaker, or local infection at the stimulation sites. Written informed consent was obtained from all subjects prior to participation and the Declaration of Helsinki was respected. The study was approved by the local ethical committee of the North Denmark Region (approval no. N-20130053).

11.3.2. functional Electrical Stimulation

In the beginning of each experimental session, stimulation sites for delivering FES to assist the subject with hand opening and hand grasping were identified. A total of up to eight self-adhesive surface electrodes (Pals Platinum Round 3.2 cm, Axelgaard Ltd., USA) were placed targeting the following muscle groups: m. flexor digitorum profundus, m. flexor digitorum superficialis, and m. abductor pollicis (hand grasping), m. extensor digitorum communis, m. extensor pollicis longus, and m. abductor pollicis (hand opening). The stimulation consisted of a pulse train with a frequency of 30 Hz and square pulse duration of 200 μs. The intensity of each stimulation channel was set to the level where visible motor activation occurred (motor threshold) plus 2 mA. The onset of stimulation assisting hand opening and grasping was controlled by the system. The duration of the stimulation assisting hand opening was set to 2.5 s, while the duration of stimulation assisting hand grasping was controlled by the system.

11.3.3. Hand Function Exercise

A hand function exercise was performed by the subjects seated on a chair in front of the table. The exercise involved lifting and moving a cylindrical object in the sagittal plane between two squares (located ~70 mm apart) marked on the table.

Two cylindrical objects (denoted as the “small-” and “large- cylinder”) were used in the hand function exercise. The cylinders had grey colored sides, a green colored lid, equal heights (100 mm) and weights (300 g). The diameters of the small and large cylinder were 40 mm and 75 mm respectively.

11.3.4. Monitoring of Grip Force

The small and large cylinder had two or four 38 mm square force sensing resistors (Interlink Electronics FSR® 406) mounted on the side, respectively (figure 11.1).

These FSRs provided a continuous measure of the grip force applied to the cylinder during the hand function exercise (ranging from 0-10 V). Each FSR was sampled at 1000 Hz and data were saved for offline analysis. The activity for all FSRs was summed in the online analysis. A grip was considered to be established when the summed FSR activity exceeded 0.2 V.

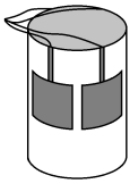


Figure 11.1: Illustration of the large cylinder with FSRs mounted on the side. A total of 4 FSRs were mounted on the side of the large cylinder.

11.3.5. Monitoring of Hand and Cylinder Kinematics

A Microsoft Kinect sensor was used for recording and analyzing the kinematics of the subject's hand and the cylinder during the hand function exercise. The Kinect sensor captured depth images and RGB images in a resolution of 640 x 480 pixels at 30 frames per second [10]. The sensor was mounted on a tripod and positioned 85 cm above the table surface providing a top-down view of the table. The position of the camera resulted in a distance between each pixel of approximately 1.5 mm.

11.3.6. Control of Functional Electrical Stimulation

The control of FES for hand opening once the hand was approaching the cylinder was based on the distance between the cylinder and the hand. FES for hand opening was triggered once the distance was between 200 mm and 30 mm (all distances are Euclidian distances). The control of FES for hand opening once the cylinder was placed in the target area (one of the two squares marked on the table) was based on the distance between the table and the bottom of the cylinder. A distance less than 10 mm triggered FES for hand opening.

FES for hand grasping was triggered once a grip around the cylinder was detected by the system (method for grip detection is described in section *I*). FES continued for each frame where grip was detected.

11.3.7. CylindeR DeTection

The detection of the cylinder was based on the RGB images. In each frame the RGB image was filtered in order to extract green colored pixels. The identified pixels were labelled as pixels representing the cylinder surface.

11.3.8. Hand Detection

Hand detection was based on both the RGB and depth image. Based on the depth image pixels with depth values in the range of the table surface were excluded as the majority of these pixels represented the table surface. Also the pixels representing the

cylinder were excluded. Finally, all connected pixels were grouped (a pixel was considered to be connected to another pixel if it was located exactly on top, below, left or right to the other pixel). Groups of less than 50 pixels were excluded, since most of these were either pixels representing the cylinder or the table. The remaining group(s) of pixels was labelled as pixels representing the hand.

11.3.9. Grip Detection

In frames where the distance between the hand and the cylinder was less than 30 mm, the system would determine whether a grip around the cylinder had been established. Initially a subset of the depth and RGB images for the present frame was used. The dimensions of the quadratic image subset were equal to the diameter of the cylinder plus 30 mm. The center of the quadratic image was matched with the estimated centroid of the cylinder.

The hand labelled pixels in the upper left and lower right part of the image were then used pairwise in a geometric calculation to determine whether the centroid of the cylinder was located left to the line intersecting the combination of points (figure 11.2). This was one of two requirements that had to be fulfilled in one case or more for a grip to be detected.

For each frame the mean of the distances of all hand labelled pixels to the centroid of the cylinder was saved. Initially, a grip was detected if this mean distance was less than the radius of the cylinder plus 15 mm. In the frames following a frame where grip had been detected a grip was still detected only if the mean distance of the present frame was less than the minimum mean distance detected during the session.

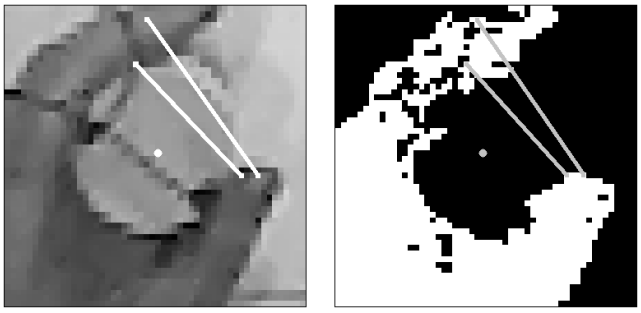


Figure 11.2: On the left the subset of the RGB image is shown in greyscale. The image on the right side shows the pixels representing the hand (the dots mark the centroid of the small cylinder).

11.3.10. Facial Expression Recognition

In order to analyze the emotional state of the patient during the experiment, a Logitech webcam was used for capturing frontal facial images of the patient at 30 frames per second with a resolution of 640 x 480 pixels. Emotional states were analyzed by a commercial facial expression recognition system, "Facereader Ver. 5.1"©, which recognizes six basic expressions: sadness, disgust, happiness, anger, surprise, fear, and neutral. This system is based upon the Active appearance model (AAM), which is typically used for facial emotion recognition [11].

11.4. **Results**

11.4.1. Differences in Grip Detections by FSRs and System

The absolute mean time differences in grip detection based on the FSRs and the Kinect output ranged from 0.18-0.27 s (figure 11.3).

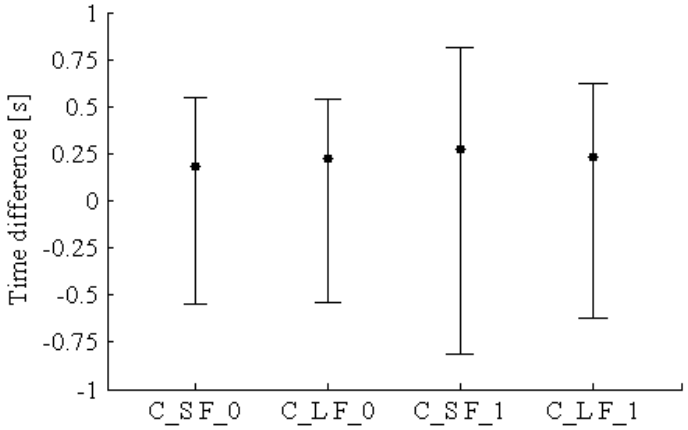


Figure 11.3: Mean time differences and 95 % CI. "C_S": small cylinder, "C_L": large cylinder. "F_0": FES not applied, "F_1" FES applied.

The average number of grips detected under the different conditions is summed up in Table 11.1.

Table 11.1: Average number of detected grips (±standard deviations).

C_S F_0	C_L F_0	C_S F_1	C_L F_1
10.0±10.9	9.0±11.5	13.8±11.5	14.8±9.9

11.4.2. Subjects' Emotional Expression

It can be seen from Table 11.2 that subject 1 is mostly disgusted, subject 2 is mostly happy, subject 3 is mostly neutral, and subject 4 is mostly angry during the exercises.

Table 11.2: Subjects' emotional expression in percentage.

Subject	Fear	Sadness	Surprise	Disgust	Happy	Anger	Neutral
1	<1 %	<1 %	<1 %	52 %	10 %	<1 %	43 %
2	<1 %	<1 %	2 %	<1 %	57 %	8 %	41 %
3	<1 %	<1 %	<1 %	<1 %	4 %	25 %	72 %
4	<1 %	<1 %	<1 %	<1 %	45 %	70 %	20 %

11.5. Discussion

11.5.1. Functional Electrical Stimulation

In two out of four subjects, it was not possible to increase the intensity of FES to a sufficient level to elicit motor responses without causing pain. Therefore, the intensity of FES given to these subjects during the experiment was below motor threshold meaning that the subjects did not experience any assistance in hand opening and grasping.

11.5.2. Monitoring of Grip Force

The method for monitoring grip force during the experiment was based on FSRs placed on the side of the cylinders. In cases where the position of the subjects hand during grasping was on the edge of the FSRs, the force recorded by the FSRs was close to zero. For that reason some of the grasps performed by the subjects had to be excluded from the analysis. Grasps were also excluded in cases where the subject used the other hand to establish the grasp.

11.5.3. Control of Functional Electrical Stimulation

This study used a fixed duration of 2.5 s of FES for assisting hand opening, and as a result of this the subjects had to wait for the stimulation to finish before grabbing the cylinder. Not all subjects had enough patience to wait for the stimulation to finish and consequently they grabbed the cylinder while getting stimulation assisting hand opening.

The onset of FES assisting hand grasping, i.e. facilitating the grip, relied on the ability of the system to detect when a grip was present. When comparing the time for grips detected by the system and the FSR sensors on the objects, an absolute mean difference less than 0.3 s was found. Similarly, another study using the Kinect sensor for detection of hand closing postures compared the difference between detections of movement onset by the Kinect sensor with onset of EMG activity in the hand flexor muscles and found a mean difference less than 0.25 s [12].

11.5.4. Object Detection

The method used for detection of the cylinders was solely based on the color of the lid of the cylinder. Therefore, the detection of the cylinder was sensitive to changes in the background light. This is a common issue for all the methods including analysis of RGB images. The method is also sensitive to objects with a color similar to that of the cylinder in case that this object is located too close to the cylinder or is larger than the cylinder. This problem might be solved by combining the existing method with estimation of the size of the detected objects from the depth images. Then objects that are too small or too large compared to the size of the cylinder can be excluded.

The method for detection of the hand of the subject was based on both the depth and RGB image. It would have been preferable to detect the hand from the depth image alone. However, this was not possible for the frames where the subject had established a grasp around the cylinder. In these frames the depth information of parts of the fingers was not available most likely due to a shadowing effect caused by the cylinder (the infrared beams from the Kinect sensor were reflected by the cylinder before they reached the fingers located closer to the table surface). Therefore these parts of the hand could not be detected.

The clinical value of the system is highly dependent on the precision of the captured kinematics, but the present study has not validated the precision of the object detection. A previous study has shown that the Kinect sensor can be used for detection and tracking of finger joints angles with an average absolute error ranging from 2.4 to 4.8 degrees [13].

11.5.5. Facial Expression Recognition

Our own observations and questioning of the subjects showed that in most cases they were actually neutral even though results from the facial expression recognition yielded that three out of four subjects were mainly non-neutral. The difference between the results of Facereader and our own observations might be due to the insufficient facial muscle control of the patients. In addition to the facial analysis of the patients during the exercises we asked one of the patients to express the six mentioned basic facial expressions in the end of the experimental session (each facial expression was maintained for approximately 15 seconds). We observed that it was very hard for the subject to do that though she tried her best. Even though she was not able to show her emotion at all, Facereader was still detecting wrong emotional states, clearly indicating a shortage in the classification approach.

11.6. Conclusion

In this study, it has been shown that it is possible to control FES assisting hand opening and grasping by a Microsoft Kinect sensor. When comparing the time for grips

detected by the system and the FSR sensors on the objects, an absolute mean difference of less than 0.3 s was found. Such a difference would be functionally useable.

The results from the study also suggest that present facial expression recognition systems are not reliable for recognizing patients' emotional states especially when they have difficulties to control/move their facial muscles. For addressing these issues, one may train facial expression recognition systems with facial images captured directly from patients and then combine the results with physiological signals of facial images, similar to our previous work at [14]. Combining the system with proper facial expression recognition would make it possible for the system to provide the patient different kinds of feedback, e.g. changing the level of difficulty of the task when the patient has been detected as being bored.

11.7. **Acknowledgment**

The research council for Technology and Production supported the study. Træningsenheden, Aalborg Municipality, assisted with the clinical validation studies.

11.8. **References**

[1] World Health Organization, "The World Health Report 2003: shaping the future," (accessed Feb 25, 2014).

[2] H. S. Jørgensen, H. Nakayama, H. O. Raaschou, J. Vive-Larsen, M. Støier, and T. S. Olsen," Outcome and Time Course of Recovery in Stroke. Part II: Time Course of Recovery. The Copenhagen Stroke Study," Arch Phys Med Rehabil, vol. 76, pp. 406–412, May 1995.

[3] S. Lai, S. Studenski, P. W. Duncan, and S. Perera," Persisting Consequences of Stroke Measured by the Stroke Impact Scale," Stroke, vol. 33, pp. 1840–1844, July 2002.

[4] K. E. Laver, D. Schoene, M. Crotty, S. George, N. A. Lannin, and C. Sherrington, "Telerehabilitation services for stroke," Cochrane Database of Systematic Reviews, issue 12, 2013.

[5] T. Johansson, and C. Wild, "Telerehabilitation in stroke care – a systematic review," J Telemed Telecare, vol. 17, January 2011.

[6] O. Schuhfried, R. Crevenna, V. Fialka-Moser, and T. Paternostro-Sluga, "Non-invasive neuromuscular electrical stimulation in patients with central nervous system lesions: an educational review," J Rehabil Med, vol. 44, pp. 99–105, 2012.

[7] A. Meilink, B. Hemmen, H. Seelen, and G. Kwakkel," Impact of EMG-triggered neuromuscular stimulation of the wrist and finger extensors of the paretic hand

after stroke: a systematic review of the literature," Clin Rehabil, vol. 22, pp. 291–305, 2008.

[8] C. Darwin, "The expression of the emotions in man and animal," J. Murray, London, 1872.

[9] P. Ekman and W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," Consulting Psychologists Press, Palo Alto, 1978.

[10] Microsoft Corporation, http://www.microsoft.com/en-us/kinectforwindows/discover/features.aspx (accessed Feb 25, 2014)

[11] R. B. Knapp, J. Kim, and E. André, "Physiological signals and their use in augmenting emotion recognition for human–machine interaction," in Emotion-Oriented Systems, Pt. 2, Springer-Verlag, pp. 133–159, 2011.

[12] R. Scherer, J. Wagner, G. Moitzi, and G. Müller-Putz, "Kinect-based detection of self-paced hand movements: Enhancing Functional Brain Mapping paradigms," 34th Annual International Conference of the IEEE EMBS, 2012.

[13] C. D. Metcalf, R. Robinson, A. J. Malpass, T. P. Bogle, T. A. Dell, C. Harris, and S. H. Demain, "Markerless Motion Capture and Measurement of Hand Kinematics: Validation and Application to Home-Based Upper Limb Rehabilitation," IEEE Transactions on Biomedical Engineering, vol. 60, no. 8, 2013.

[14] R. Irani, K. Nasrollahi, B. Moeslund, "Improved Pulse Detection from Head Motions Using DCT," in 9th International Conference on Computer Vision Theory and Applications, Lisbon, 2014.

PART IV

ESTIMATION OF PSYCHO-PHYSIOLOGICAL INDICATORS

Chapter 12

Thermal Super-Pixels for Bimodal Stress Recognition

Ramin Irani, Kamal Nasrollahi, Abhinav Dhall, Thomas B. Moeslund, and Tom Gedeon,

This paper has been published in

IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, Finland, 2016, pp. 1-6

The layout has been revised.

12.1. **Abstract**

Stress is a response to time pressure or negative environmental conditions. If its stimulus iterates or stays for a long time, it affects health conditions. Thus, stress recognition is an important issue. Traditional systems for this purpose are mostly contact-based, i.e., they require a sensor to be in touch with the body which is not always practical. Contact-free monitoring of the stress by a camera [1], [2] can be an alternative. These systems usually utilize only an RGB or a thermal camera to recognize stress. To the best of our knowledge, the only work on fusion of these two modalities for stress recognition is [3] which uses a feature level fusion of the two modalities. The features in [3] are extracted directly from pixel values. In this paper we show that extracting the features from super-pixels, followed by decision level fusion results in a system outperforming [3]. The experimental results on ANUstressDB database show that our system achieves 89\% classification accuracy.

12.2. **Introduction**

Nowadays stress is a major problem in the human society. Usually the reason for stress is recognized as time pressure. For example, when we want to perform a task within a given period while we do not have enough time a set of physiological reactions, like, heartbeat and respiration rates increase indicate a stressful situation [4]. This situation, however, may not be the same for different people, as stress is subjective. In other words, stress depends more on changes in specific physiological signs, not on conditions/events themselves [4]. It is also argued that different people may experience different conditions or events and hence their physiological signs may change differently.

Traditional stress recognition systems, which are based on self-report or measure physiological signals using invasive sensors have some limitations. For example: these systems are unable to monitor the subjects instantaneously and continuously [5]. Some of the systems are based on self-reporting or saliva test. Self-reporting-based systems at times may not be able to recognize the stress on short duration of time. As a consequence, to overcome these problems researchers nowadays tend to measure stress using contact-less sensors such as RGB and thermal cameras.

Since stress is associated with physical appearance, some researchers utilized the physical symptoms as clues for stress detection. For example, in [6], [7] a model for monitoring the subjects based on deformation of lips, mouth and eyebrows due to the stress has been presented. Liao et al [1] proposed an approach to recognize the stress using some visual features like blinking frequency, average eyes closure speed and percentage of saccadic eye movement. In another work [8] authors detected the stress by tracking 3D facial expressions. In a recent work, Gao et al [9] presented a contact-less real time system for detecting stress in vehicle drivers. It functioned by considering two negative basic facial expressions (anger and disgust) based on the Ekman

theory [10] and applied Local discrete cosine transform [11] and scale invariant feature transform [12] as features.

Since physical appearance is not as reliable as physiological response to stressors, many researchers are interested in employing these symptoms for stress detection. Recently, imaging techniques like RGB video recorder or thermal imaging have been employed for contact-less measuring of physiological signals, e.g., Hearthbeat rate [13], [14], respiration rate [15], [16] and muscle fatigue [17] which promise contact-less-based measuring of the stress as well.

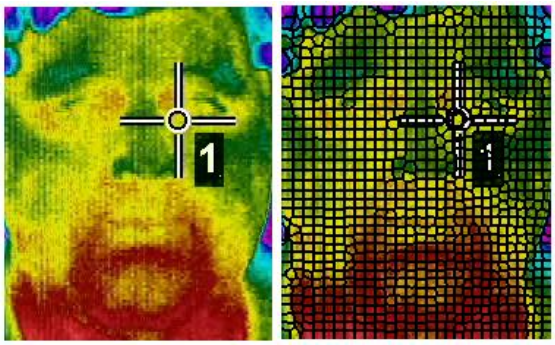
Hyperspectral imaging [5], thermal imaging and RGB Imaging have been used to probe the stress using physiological features. Pavlidis et al [18], [19] are the first researchers who measured the stress with a contact-less thermal sensor. The principle of their work is based on the fact that mental stress increases the blood flow in the forehead region. Thus, they applied a contact-less blood flow measurement on 10% hottest pixel of the ROI for monitoring the stress. To quantify the stress level, in another interesting work, Shastri et al. [20] captured thermal imaging-based data for measuring the transient perspiration, which is also known as a physiological functions. The drawback of this method is that stress cannot be measured in scenarios, where subjects experience heavy sweating due to hot environment or during doing exercise.

In [2] Frédéric et. al. proposed a system based on instantaneous pulse rate signal extracted from imaging photoplethysmography. The proposed algorithm derives the Heart Rate Variability (HRV) from a webcam and detects the stress by analysing the HRV changes due to stress.

Employed vision-based systems for stress recognition usually use only one of the RGB or thermal imaging techniques. To employ the opportunities of fusing the two modalities, a recent literature [3] presented a computational model using the information from both thermal and RGB imaging. They proposed a new descriptor named Histogram of Dynamic Thermal Patterns (HDTP). However, they could not achieve more than 65% accuracy. Nevertheless, such accuracy has been improved to 85% by combining RGB and thermal imaging features as input of a Genetic Algorithm (GA)-Support Vector Machine (SVM) classifier.

In thermal images a unique colour is assigned to the pixels with similar temperature (figure 12.1.a). This gives rise to a formation of regions on the facial block. An advantage of thermal based face analysis over RGB is that it is less affected by noise in the facial parts location detection. Extracting features in the sub-regions (block-wise) instead of the entire face (holistic-level) improves the accuracy [21] in RGB images based face analysis. Generally, a particular block may cover a facial part, or two adjacent blocks may contain a particular facial part. However, this is not guaranteed for thermal images as the sub-regions in thermal areas don’t strictly adhere to facial part

boundaries. Another reason for this can be that if the thermal image is divided into fixed blocks, it may result into a block containing different (complete/incomplete) thermal regions, which may not have any correlation. Furthermore, due to the process of image capturing a sensor quantizes a natural continuous signal (image) into pixels. Motivated by these observations, in this paper we propose to represent a thermal image as a group of super-pixels. A super-pixel is a group of adjacent pixels which have similar characteristic and special information (figure 12.1.b). Super-pixel representation has been used for face recognition [22]. In the case of thermal images, super-pixels are a group of pixels with similar colour (temperature) which seem like a more natural representation for thermal images as compared to dividing images into non-overlapping blocks. This method not only groups the adjacent pixels with high correlation but also increases the speed of processing. Our experimental results show promising outcomes and are in agreement with the state-of-the-art method of [3].



a b

Figure 12.1: A typical facial Region a and its corresponding super-pixels b

Super-pixels are the results of perceptual grouping of pixels and involve more information and provide better image alignment compared to using a single pixel alone [23]. Mapping from a pixel grid to super-pixel, holds desirable properties, like, computational efficiency, perceptual meaningfulness, over segmentation, and efficient graph representations [24]. Super-pixels share some properties like texture distribution or colour similarity. Specially, this attribute can be helpful in thermal image analysing, because we are interested in temperature of sub-regions instead of points.

In the recent years, there has been progress in super-pixel creating algorithms [25], [26]. A detailed pros and cons of various super-pixel algorithms are presented in [23]. In this work, for computing super-pixels, Linear Spectral Clustering (LSC) method [27] is followed. The reason for choosing LSC is its ability to produce fast compact and uniform super-pixels.

The rest of the paper is organized as follows: Section 3 Explains the details of the proposed system, Section 4 Discusses the experimental results, and finally, Section 5 concludes the paper.

12.3. **The proposed system**

The block diagram of the proposed system is shown in figure 12.2. The test subjects are filmed by a RGB camera that is synchronized with a thermal camera in parallel. These two types of video streams go through three different steps: 1) Face region detection and quality assessment, 2) Feature extraction, and 3) Classification and fusion.

Since the data collected by the two types of cameras are different in nature, the applied algorithms in the first two steps are different. For RBG images, recognizing stress is similar to [1], that is, first facial region is detected by the Viola Jones (VJ) face detector. Then, the face regions with less correlation are removed using a face quality assessment algorithm. Finally, Local Binary Patterns (LBP) [28] is extracted from the remaining facial regions and is used as feature points. However, for detecting the face area in the thermal images, we use a template matcher for face region detection as proposed in [31]. Then, we compute the LSC super-pixel algorithm, instead of directly computing a facial descriptor. Further, the mean values of the generated super-pixels are used as the facial features. Having extracted the facial features from two types of inputs, we use a support vector machine (SVM) classifier for producing classification scores for each type of input. These scores are finally fused at decision level to recognize the stress. These steps are explained in detail in the following subsections.

12.3.1. Step 1: face detection and quality assessment

1) RGB Data: The first step of stress recognition in RGB videos is cropping the face region. We used the VJ face detection algorithm [29] for this purpose. In order to decrease the error of the algorithm, if it cannot find a face in the current frame, we use the position of the frame in the previous frame as the position of the face in the current frame. Considering the fact that in our employed database (discussed later) the subjects' face does not have considerable head pose changes and movements within short period of time, this method seems working and reducing the error of the face detection algorithm, when it fails to detect faces. Furthermore, if there is more than one region detected as face, we utilize the information about the setup (discussed later in III) to keep only the one which is closest in size to v×w.

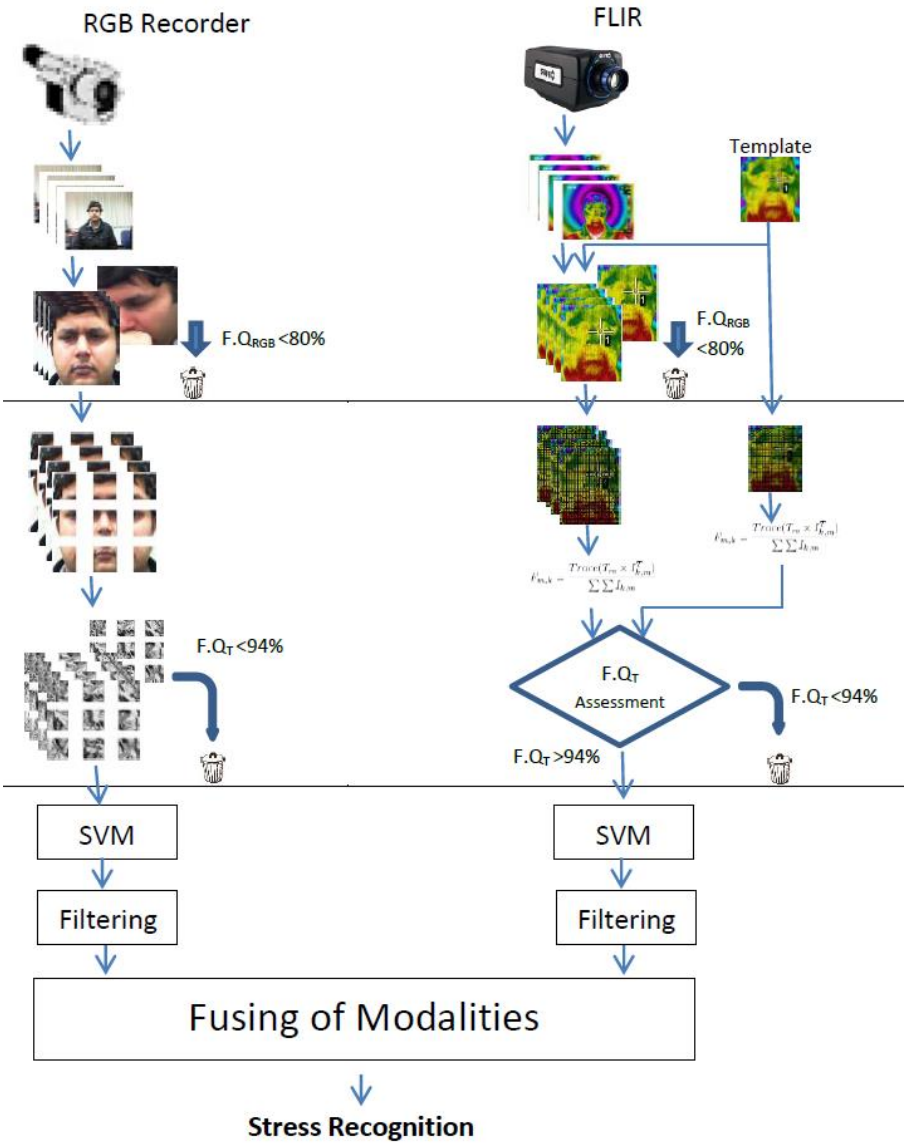


Figure 12.2: The block diagram of proposed bimodal system

The values of v and w are determined experimentally, based on the distance of the subjects from camera.

Finally, we employ a face quality assessment technique for detecting the frames with incorrect face region (figure 12.3). To do so, we use the first detected face in the first frame as a reference face and discard all the other faces that are not similar enough to this reference face (less than 80%). The similarity is calculated using the following correlation:

$$S_{RBG} = \frac{\sum_m^M \sum_n^N (A_{mn} - \bar{A}).(B_{mn} - \bar{B})}{\sqrt{\sum_m^M \sum_n^N (A_{mn} - \bar{A})^2 . \sum_m^M \sum_n^N (B_{mn} - \bar{B})^2}} \times 100 \qquad (1)$$

in which, A is the template/reference face, $\bar{A}$ is the average grey level in the reference image, B is the face in the current frame, $\bar{B}$ is the average grey level of the face in the current frame, and M & N are the number of rows and columns of frames, respectively, (template image size = columns × rows).

a b c

Figure 12.3: Quality assessment, a. Correlation of the frames with a chosen template for all the frames in subject 1 video sequence, b. a frame with correlation less than 80% c. a frame with correlation larger than 80%

Figure 12.3.a shows a correlation curve obtained by the above formula for the entire faces of a video sequence and two faces. The first face (figure 12.3.b) has been discarded while the second one (figure 12.3.c) has been kept.

It should be mentioned that when we discard a frame/face in RGB video sequence, its corresponding frame in the thermal video sequence should also be discarded. It is an essential condition to keep the synchronization between the modalities.

2) Thermal Data: Before applying the LSC super-pixel algorithm on the thermal images, face localization in the thermal images is required. Since the VJ algorithm, which was applied on the RGB frames, is not useful in this case, we used a template matcher [30]. The template is created manually for each thermal video sequence. The facial region on one frame (the reference frame) is cropped and then used to find the

facial regions in the rest of the frames using the Yue Wu algorithm of in [30] which is based on correlation. Figure 12.4.c shows the correlation values between the template (figure 12.4.b) and the face region in the current frame (figure 12.4.a). The brightest point on the correlation map (figure 12.4.c) indicates center of the face region that should be cropped. Figure 12.4.d shows the detected and cropped face region.

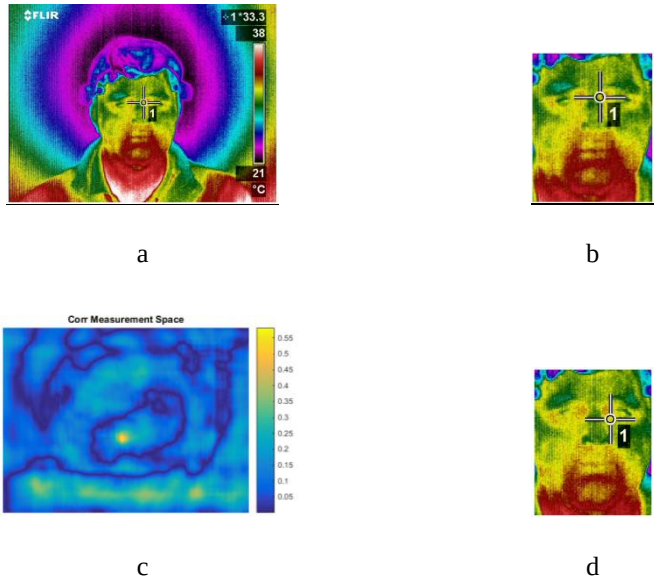


Figure 12.4: Template matching process for detecting facial region in thermal images: a. a frame from the ANU StressDB database, b. Template, c. correlation map, d. the detected Facial region

12.3.2. Step 2: Feature extraction:

1) Extracting features from RGB data: It is discussed in [21] that for RGB images it is be better to extract facial features from individual non-overlapping blocks than at the holistic level. In this work, similar to [3] in each frame the facial region is segmented into a grid of 3×3 blocks. Next, LBP features are computed for each block [28]. LBP has been successfully utilized in many facial analysis systems, like [31], [32]. Figure 12.5.a and 12.5.b show an input image divided into 3 by 3 blocks and their corresponding LBP counterparts.

2) Extracting features from thermal data: To extract features from thermal images, we first apply the super-pixel technique of [27] to segment the face regions to small none-overlapping pieces. This technique divides the facial region to some sub-regions (figure 12.1.b) that unlike blocked RGB images (figure12.5.a), each sub-

a b

Figure 12.5: Facial sub-regions in a. RGB frames, b. Corresponding LBP features

region includes pixels with mostly similar color (hence similar temperature in this case). Such property, in addition to the fact that the stress has direct correlation with skins temperature, made it possible to consider the mean temperature of each super-pixel as feature of the corresponding sub-region. The region of each super-pixel is determined using a matrix named Label with the size equal to the size of a face. Label assigns an integer to each super-pixel such that:

$$Label = \bigcup_{k=1}^{K} k.I_k \tag{2}$$

in which:

$$I_k = \begin{cases} 1 & if\ P_{i,j} = k \\ 0 & otherwise \end{cases} \tag{3}$$

The Thermal feature F_m for each frame is given by:

$$F_m = \frac{Tr(T_m \times I_{k,m}^T)}{\sum\sum I_{k,m}} \tag{4}$$

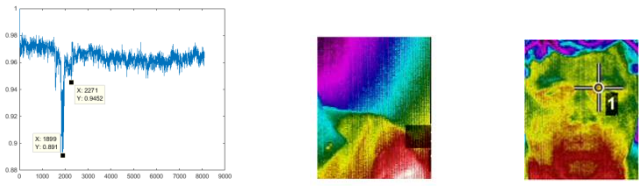
where, F_m is k_{th} element of the feature vector of m_{th} frame, T_m is m_{th} thermal frame, $I_{k,m}$ is transport matrix of k_{th} super-pixel in frame m, $P_{i,j}$ is a pixel on the thermal image of T_m with coordinate *i,j*, ∪ is a union function, and Tr is Trace function.

Similar to RGB modality we apply a quality assessment to thermal modality. The quality assessment is however not applied to the thermal images, but to their features, as these features are one dimensional (average of blocks) and are much easier to process than the images. To do so, we use the correlation scores obtained by Equation 5. The difference between this correlation and the one used for the RGB modality is that

we here have replaced gray levels with the mean of super-pixels. In addition, since the applied features (temperature) involves values with small variation, an exponential function and with a factor of α has been considered to depict the frames with less quality with decreasing their corresponding scores faster than high quality frames, as in:

$$S_T = exp(\alpha \times \left(\frac{\sum_{i=0}^{N-1}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{N-1}(x_i - \bar{x})^2.\sum_{i=0}^{N-1}(y_i - \bar{y})^2}} - 1\right)) \times 100 \quad (5)$$

where, x is the feature vector of the template frame, $\bar{x}$ is the average of template feature vector, y is the feature vector of source frame, $\bar{y}$ is the average of source features vector, and N is the size of feature vector. The value of S_T varies within 0 and 100, such that larger values of "S_T" represent a strong relationship between the two images. The features (also frames) with score (S_T) less than 94% were removed. Figure 12.6.a illustrates the score of a video sequence. Figures 12.6.b and 12.6.c show corresponding frames of the spots marked on figure 12.6.a. It can be seen that the frame with a score less than 94% is not correctly matching with the template.



a b c

Figure 12.6: Quality assessment of thermal images, a. Similarity of a thermal segment for subject 1, b. detected face region with score 89%, c. detected face region with score 95% .

12.3.3. Step 3: Fusing and classification

Motivated by the successful application of SVM in different vision algorithms [33], [34], we have decided to use it for classifying our features. Two separate SVMs have been used for the classification of the features extracted from the different modalities. The output of these SVMs needs to be fused to make a decision if the test subjects are in stress in the current RGB and its corresponding thermal frames or not. Since stress is a continues phenomenon and cannot vary abruptly, we have assumed that the level of the stress does not change in short periods (here experimentally obtained period of four seconds has been used). To reflect this temporal period, we have applied a median and then a mean filter to the output of each SVM. The median filter

removes the outlier scores of the SVMs, while the mean filter (moving average) aggregates the scores over temporal period of the stress. To apply the moving average, the outputs of median filter are windowed with length of N and an overlapping factor of N-1. Finally, the output of moving average filter is weighted and fused using the following equation:

$$S_{Modal} = tanh(\gamma.(\omega_1.S_{RGB} + \omega_2.S_T + Threshold)) \quad (6)$$

where, S_{RGB} is output of RGB SVM modal, S_T is output of thermal SVM modal, S_{Modal} is the final output after fusion, ω_1 and ω_2 are weight coefficients of RGB and thermal inputs of the fusion, and *Threshold* is a threshold for making decision if the frame is stressful or not. Frames with corresponding value less than threshold are stressful frame and those larger than threshold are non-stressful.

12.4. **Experimental results and discussion**

This proposed system, has been tested on the only database for stress recognition that contains images of both RGB and thermal modalities, ANUStressDB. This database has been collected at Australian National University (ANU) [3] and involves 35 subjects, composed of 22 males and 13 females, between 23 and 39 years old. The thermal and RGB modalities were captured by a FLIR infrared camera and a Microsoft webcam, respectively. Both cameras were working at 30 frames per second at a 640x480 pixels resolution. We set the values of v and w (of section II-A.1) to 110 pixels each.

Instructors played a film with a collection of negative and positive clips as stress stimulator. The clips are separated by displaying 5 seconds blank screen in-between the clips in order to neutralize the participants' emotion (state of mind) before displaying the next movie. Therefor, in the ground truth data, we assigned all the frames as stressed/unstressed when the label of the film is stressed/unstressed. At the end of the experiment, participants were asked to fill a questionnaire survey for the validation of the experiment.

For classifying the extracted features using SVM, 60% of the samples from each modality were selected for training, and the rest for testing. Since the stress is a temporal process, besides considering the SVM scores directly, we have considered applying some temporal post-processing technique which consider a kind of history for the current frame to be involved in the decision making about the current frame to be classified as stress or not. For this purpose, we have simply looked into mean and median filters. In other words, to decide whether or not the current frame is of a stressful situation, besides looking into the SVM score of the current frame, we apply once a mean and once a median filter to the SVM scores of the frames located within a neighborhood of the current frame.

Table 12.1 shows the results obtained by SVMs for each modality without any post-processing (the second column), with mean and median filters applied to the results of the SVM (third and fourth columns, respectively), and after fusing the post-processed (mean-filtered) results of both modalities (fifth column).

Tabel 12.2: Comparing improvement of the results in each step of post processing filters and fusing

Modal	SVM	Median	Moving Average	Fusion
RGB	60 %	60 %	62 %	89 %
Thermal	82 %	84 %	86 %	

Figure 12.7 shows the results of the proposed system aganist the ground thruth after fusing the scores coming from both modalities.

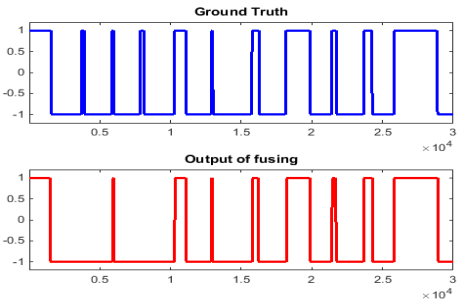


Figure 12.7: Comparing the ground truth with final result after fusing the modalalities

Table 12.2 shows the results of comparison of the modality fusion of the proposed system against those of [3]. It can be seen from this figure that the proposed system outperforms Sharma et. al's approach [3] by more than 4% accuracy.

Tabel 12.2: Comparing the proposed system against the state-of-the-art system of [3]. Bars number I to V represent, respectively, (VLBP+TLBP) with SVM classifier, (VLBP+TLBP) with Genetic Algorithm SVM classifier (GASVM), (VLBP+THDTP) with SVM classi

methods	I	II	III	IV	V
Accuracy	61 %	79 %	76 %	85 %	89 %

12.5. **Conclusion**

Stress recognition using computer vision techniques is of great importance as it does not need any contact with users, which is unavoidable in traditional methods. To this end, in this paper we proposed a system that uses facial images of different modalities, including RGB and thermal to make a decision if a user in a current frame is in

a stressful situation or not. From RGB and thermal modalities, LBP and temperature of super-pixels have been used as features that are fed to a SVM classifier. The SVM results of the two different modalities are then combined using a score level fusion. The experimental results showed that the purposed fusion results in a system that outperform the state-of-the-art stress recognition system.

12.6. **References**

[1] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network," in Computer Vision and Pattern Recognition Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on. IEEE, 2005, pp. 70–70.

[2] F. Bousefsaf, C. Maaoui, and A. Pruski, "Remote assessment of the heart rate variability to detect mental stress," in Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on. IEEE, 2013, pp. 348–351.

[3] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, "Thermal spatio-temporal data for stress recognition," EURASIP Journal on Image and Video Processing, vol. 2014, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2014-28

[4] S. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. Schramek, "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition," Brain and Cognition, vol. 65, no. 3, pp. 209 – 237, 2007. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S02782626070003-22

[5] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She, "Detection of psychological stress using a hyperspectral imaging technique," Affective Computing, IEEE Transactions on, vol. 5, no. 4, pp. 391–405, 2014.

[6] D. Metaxas, S. Venkataraman, and C. Vogler, "Image-based stress recognition using a model-based dynamic face tracking system," in Computational Science-ICCS 2004. Springer, 2004, pp. 813–821.

[7] D. Dinges, E. McGlinchey, S. Venkataraman, and D. Metaxas, "Optical computer recognition of behavioral stress in space flight," Habitation International Journal for Human Support Research, vol. 10, no. 3/4, p. 233, 2006.

[8] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, "Optical computer recognition of facial expressions associated with stress induced by performance demands," Aviation, space, and environmental medicine, vol. 76, no. Supplement 1, pp. B172–B182, 2005.

[9] H. Gao, A. Yuce, and J. P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 5961–5965.

[10] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci Bitti et al., "Universals and cultural differences in the judgments of facial expressions of emotion." Journal of personality and social psychology, vol. 53, no. 4, p. 712, 1987.

[11] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," in 13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey, 2005.

[12] W. S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 3515–3522.

[13] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 3430–3437.

[14] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Improved pulse detection from head motions using dct," in International Conference on Computer Vision Theory and Applications, 2014.

[15] J. Fei and I. Pavlidis, "Thermistor at a distance: unobtrusive measurement of breathing," Biomedical Engineering, IEEE Transactions on, vol. 57, no. 4, pp. 988–998, 2010.

[16] J. Fei, Z. Zhu, and I. Pavlidis, "Imaging breathing rate in the co 2 absorption band," in Engineering in Medicine and Biology Society, 2005. IEEE EMBS 2005. 27th Annual International Conference of the. IEEE, 2005, pp. 700–705.

[17] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Contactless measurement of muscles fatigue by tracking facial feature points in a video," in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4181–4185.

[18] I. Pavlidis and J. Levine, "Thermal image analysis for polygraph testing," Engineering in Medicine and Biology Magazine, IEEE, vol. 21, no. 6, , 2002, pp. 56–64.

[19] I. Pavlidis, N. L. Eberhardt, and J. A. Levine, "Human behaviour: Seeing through the face of deception," Nature, vol. 415, no. 6867, 2002, pp. 35–35.

[20] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, "Perinasal imaging of physiological stress and its affective potential," Affective Computing, IEEE Transactions on, vol. 3, no. 3, 2012 pp. 366–378.

[21] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 6, 2007, pp. 915–928.

[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[23] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in Proc. Forum Bildverarbeitung, 2012, pp. 1–12.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 11, pp. 2274–2282, 2012.

[25] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 12, pp. 2290–2297, 2009.

[26] M. Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 2097–2104.

[27] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," Trans. on PAMI, vol. 31, no. 12, pp. 2209–2297, 2009.

[28] M. Pietikainen, ¨ A. Hadid, G. Zhao, and T. Ahonen, "Local binary patterns for still images," in Computer Vision Using Local Binary Patterns. Springer, 2011, pp. 13–47.

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

[30] Template matching using correlation coefficients. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/28590-template-matching-using-correlation-coefficients

[31] A. Hadid, M. Pietikainen, ¨ and T. Ahonen, "A discriminative feature space for detecting and recognizing faces," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2. IEEE, 2004, pp. II–797.

[32] X. Feng, A. Hadid, and M. Pietikainen, ¨ "A coarse-to-fine classification scheme for facial expression recognition," in Image Analysis and Recognition. Springer, 2004, pp. 668–675.

[33] M. Taini, G. Zhao, S. Z. Li, and M. Pietikainen, ¨ "Facial expression recognition from near-infrared video sequences," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008, pp. 1–4.

[34] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in Proceedings of the 5th international conference on Multimodal interfaces. ACM, 2003, pp. 258–264.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-064-7

AALBORG UNIVERSITY PRESS