Video-based infant discomfort detection

Video-based infant discomfort detection

Y. Sun

Video-based infant discomfort detection

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen op
woensdag 7 juli 2021 om 16:00 uur

door

Yue Sun

geboren te Tianjin, China

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. S.M. Heemstra de Groot
1e promotor:	prof.dr.ir. P.H.N. de With
co-promotor:	prof.dr. R.M. Aarts
leden:	prof.dr.ir. S.J.A. Van Huffel (Katholieke Univ. Leuven)
	prof.dr. C.G.M. Snoek (Universiteit van Amsterdam)
	prof.dr. J.P.W. Pluim
	prof.dr. C. Shan (Shandong Univ. Science & Technology)
	dr.ir. C. van Pul (Máxima Medisch Centrum Veldhoven)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

To my parents

Video-based infant discomfort detection

Y. Sun

Cover design: Xi Chen and Yue Sun
Printed by: ProefschiftMaken

ISBN 978-90-386-5314-3
NUR-code 959

Summary

Video-based infant discomfort detection

Neonates have a neurobiological vulnerability to pain, due to their lower pain threshold, sensitization from repeated pain, and immature systems for maintaining homeostasis. Frequent and/or recurrent pain and discomfort can lead to abnormal brain development, yielding long-term adverse neurodevelopmental outcomes. Hospitalized infants receive special care, where their vital signs are continuously monitored. Unfortunately, only a few systems have been developed for detecting infant stress and discomfort. Monitoring by healthcare professionals invokes high costs, and is time-consuming and subjective in the assessment. Because of these limitations, infants are currently only observed during short intervals for a few times a day, which likely leaves many discomfort moments unnoticed. To aid clinical staff in detecting discomfort/stress status of infants for timely and appropriate treatments, the research in this thesis investigates an automated system for the computer-aided detection of infant discomfort using video monitoring.

The first part of this thesis covered by Chapter 2 and Chapter 3, introduces infant discomfort detection based on facial expression analysis of single static video frames. The initial steps of the workflows in both chapters are face detection and the face Region-of-Interest (RoI) normalization. After the face ROIs are identified, Chapter 2 presents conventional machine learning techniques, which involve extracting handcrafted features of geometrical and appearance facial information. In addition to this subject-independent scheme, template matching-based features are further explored for subject-dependent detection. Chapter 3 exploits deep Convolutional Neural Network (CNN) algorithms to address the problem of single-frame analysis. Given the limited available data for training, a pre-trained CNN model is employed, followed by a fine-tuning strategy. The usage of deep learning-based methods substantially increase the

performance achieved by conventional machine learning techniques (AUC increases from 0.87 to 0.96).

The second part of the thesis described by Chapter 4 and Chapter 5, is devoted to the explicit analysis of temporal information for better understanding the infant videos. Chapter 4 discusses a method for automated detection based on analyzing facial and body motion. Motion trajectories are estimated from frame to frame using optical flow. Time- and frequency-domain features are further calculated for the classification. Chapter 5 describes infant body motion by two-dimensional (2D) time-frequency representations characterizing the distribution of signal energy. Log Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCCs), and Spectral Subband Centroid Frequency (SSCF) features are calculated from the one-dimensional (1D) motion signal. Finally, deep CNNs are applied to the 2D images for the binary comfort/discomfort classification. The best results have shown an AUC of 0.985 and an accuracy of 94.2 % when combining 2D feature representations with CNN-based classification.

The third part of the thesis in Chapter 6 exploits 3D CNNs to analyze spatial and temporal information in videos simultaneously. An automated and continuous video-based system for monitoring and detecting discomfort is proposed. The system employs a novel and efficient 3D CNN, which achieves an end-to-end solution without the conventional face detection and tracking steps. Experimental results show that the proposed system achieves an AUC of 0.99, while the overall labeling accuracy is also 0.99.

In the fourth and last part of this thesis, Chapter 7 demonstrates the feasibility of quantitive physiological measurements in infants based on deep learning methods. An automated pipeline is investigated to estimate respiration signals from videos for premature infants in neonatal intensive care units (NICUs). Two flow estimation methods, namely conventional optical flow-based and deep learning-based flow estimation methods, are employed and compared to estimate pixel-based motion vectors between adjacent video frames. The respiratory signal is further extracted via motion factorization. The deep flow-based method reaches the best results, where an overall average cross-correlation coefficient of 0.74 is obtained, together with an average root-mean-squared error of 4.55 bpm.

The work of this thesis substantiates that video-based artificial intelligence (AI) is efficient and effective to facilitate clinical staff in detecting infant discomfort status. The research reveals that developing AI systems is feasible for clinical applications of infant monitoring. The use of video recording methods facilitate accurate and robust clinical information on the infant discomfort

status. The performance evaluation metrics for such systems have been elaborated to provide an objective assessment. The videos used for evaluating the AI systems were recorded from real clinical scenes. The research work and the obtained results offer a promising possibility to deploy the proposed techniques in clinical practice.

Samenvatting

Video-based infant discomfort detection

Neonaten hebben een neurobiologische kwetsbaarheid voor pijn vanwege hun lagere pijngrens, gevoeligheid voor herhaalde pijn en onvolwassen ontwikkeling voor het handhaven van homeostase. Frequente en/of terugkerende pijn en ongemak kan leiden tot afwijkende hersenontwikkeling, hetgeen een problematische neurologische ontwikkeling kan opleveren op de lange termijn. In het ziekenhuis opgenomen kinderen krijgen een speciale zorgbehandeling, waarbij hun vitale lichaamssignalen continu worden bewaakt en geobserveerd. Helaas zijn er maar enkele systemen ontwikkeld voor het detecteren van stress en fysiologisch ongemak bij kinderen. Het monitoren door medische professionals leidt tot hoge kosten, is tijdsintensief en subjectief in de beoordeling. Door deze beperkingen worden kinderen alleen geobserveerd in korte tijdsintervallen op slechts een paar momenten per dag waardoor veel momenten met pijn en/of ongemak niet worden geregistreerd. Om de medische staf te helpen bij het detecteren van dergelijke ongemak/stressmomenten bij baby's en de tijdige en geschikte behandeling hiervan, wordt in dit proefschrift een geautomatiseerd systeem onderzocht voor de computerondersteunde detectie van ongemak/stressmomenten met behulp videobewaking.

Het eerste deel van dit proefschrift, namelijk Hoofdstuk 2 en Hoofdstuk 3, introduceert detectie van ongemak bij baby's op basis van gezichtsuitdrukkingsanalyse van afzonderlijke statische videobeelden. De eerste stappen van de processtappen die zijn beschreven in beide hoofdstukken, zijn gezichtsdetectie en de beeldnormalisatie van het gezichtsgebied ("region of interest"). Nadat de gezichtsgebieden zijn geïdentificeerd, presenteert Hoofdstuk 2 de conventionele machine-learning-technieken, waarbij handmatig geselecteerde kenmerken worden geëxtraheerd van geometrische en uiterlijke gezichtsinformatie. Naast dit persoonsonafhankelijke schema, worden verder sjabloon-

matching gebaseerde eigenschappen onderzocht voor persoonsafhankelijke detectie. Hoofdstuk 3 onderzoekt diepgelaagde algoritmen voor convolutionele neurale netwerken (CNN) om het probleem van beeldgebaseerde analyse op te lossen Gezien de beperkt beschikbare trainingsdata, wordt een vooraf getraind CNN-model gebruikt, gevolgd door een verfijningsstrategie voor het verbeterd leren van het netwerk. De resultaten van de dieplerende netwerken verbeteren de performance van de conventionele machine-learning technieken aanzienlijk (AUC neemt toe van 0,87 naar 0,96).

In het tweede deel van het proefschrift, namelijk Hoofdstuk 4 en Hoofdstuk 5, is de temporele informatie expliciet geanalyseerd om de kindervideo's beter te begrijpen. In Hoofdstuk 4 wordt een methode voor geautomatiseerde detectie ontwikkeld die gebaseerd is op het analyseren van gezichts- en lichaamsbewegingen. Bewegingstrajecten worden geschat van beeld tot beeld met behulp van nauwkeurige bewegingsanalyse van beeldelementen ("optical flow"). Tijds- en frequentiedomeinkenmerken worden berekend voor de classificatie. In Hoofdstuk 5 wordt de lichaamsbeweging van de baby's beschreven door tweedimensionale (2D) tijd-frequentierepresentaties die de distributie van signaalenergie karakteriseren. Kenmerken van Log Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCC's) en Spectral Subband Centroid Frequency (SSCF) worden berekend op basis van het eendimensionale (1D) bewegingssignaal. Tenslotte worden diepe CNN's toegepast op de 2D-beelden voor een binaire classificatie van comfort of ongemak. Wanneer 2D-representaties van de beeldeigenschappen worden gecombineerd met een CNN-gebaseerde classificatie, krijgen de best gemeten resultaten een AUC waarde van 0,985 en een nauwkeurigheid van 94,2 %.

Het derde deel van het proefschrift in Hoofdstuk 6 onderzoekt 3D-CNN's om spatiële en temporele informatie in video's gezamenlijk te onderzoeken. Daartoe wordt een automatisch en continu draaiend videosysteem ontwikkeld voor het bewaken en detecteren van ongemak. Het systeem maakt gebruik van een nieuw efficiënt 3D-CNN netwerk, waarmee een volledig geïntegreerde oplossing kan worden bereikt zonder de conventionele stappen voor gezichtsherkenning en tracking. Experimentele resultaten leveren een AUC waarde van 0,99, waarbij de totale classificatienauwkeurigheid ook 0,99 is.

Het vierde en laatste deel van dit proefschrift in Hoofdstuk 7 demonstreert de haalbaarheid van kwantitatieve fysiologische metingen bij zuigelingen op basis van deep-learningmethoden. Een geautomatiseerd proces wordt onderzocht om ademhalingssignalen te schatten uit video-opnamen van zuigelingen op de neonatale intensive care units (NICU's). Twee methoden voor het schatten van de ademhalingsbeweging, namelijk de conventionele methode op basis

van optical flow en een deep-learningmethode, worden gebruikt en met elkaar vergeleken om pixel-gebaseerde bewegingsvectoren tussen opeenvolgende videobeelden te schatten. Het ademhalingssignaal wordt verder geëxtraheerd via bewegingsfactorisatie. De deep-learningmethode bereikt de beste resultaten, waarbij een totale gemiddelde kruiscorrelatiecoëfficiënt van 0,74 en een gemiddelde RMSE fout van 4,55 bpm wordt verkregen.

Het onderzoek in dit proefschrift bevestigt dat video-gebaseerde kunstmatige intelligentie (KI) efficiënt en effectief is om klinisch personeel te helpen bij het detecteren van ongemak bij baby's. Het onderzoek toont aan dat het ontwikkelen van KI-systemen haalbaar is voor klinische toepassingen van babybewaking. Met gebruik van video-opnamemethoden is het mogelijk om nauwkeurige en bruikbare klinische informatie over de ongemaksstatus van een zuigeling te verkrijgen. De prestatie- en evaluatiemetriek voor dergelijke systemen zijn uitgewerkt om een objectieve beoordeling te kunnen geven. De video's die zijn gebruikt voor het evalueren van de KI-systemen, zijn opgenomen vanuit echte klinische scènes. De verkregen resultaten bieden een veelbelovende mogelijkheid om de voorgestelde systemen in de klinische praktijk toe te passen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Clinical background of infant discomfort

The birth of a baby is one of the most joyful events in the history of a family, provided that all is well with the baby and the mother in the initial period after giving birth. However, the pregnancy period or the early stages of life of the baby can be complicated, caused by health issues for both the baby and the mother. Such issues can seriously hamper the healthy start for an infant and the involved parents. When complications arise, typically infants are born in the hospital and kept for several days or even longer periods for monitoring and treatment, if they are required. The primary objective in the hospitals is to correctly diagnose the occurring problems, solve those as far as possible, and then help the baby to get recovered and become strong enough to be released from the hospital to their normal lives.

When serious situations happen to the infants (e.g. infants can be extremely weak) after being born, the current solution in the hospitals is that the infants are kept in Neonatal Intensive Care Units (NICUs) for continuous monitoring. The infants are regularly visited and checked by doctors, and special feeding/tests are also being performed. There are various types of infants being monitored in NICUs, which can be categorized as conditions, including 1) weak infants with diseases, and 2) infants without diseases, but just intrinsically premature born. Regardless of the types of infants, pain and discomfort are important aspects for infant well-being and also for starting a healthy development. Monitoring of the well-being of neonates is an important topic for

clinicians, parents and the infants themselves. It is critical for both short-term and long-term development of infants, and it has been found that it influences their brain development. Controlling pain in the newborn period of infants is beneficial for improving physiological, behavioral, and hormonal outcomes.

Recent findings show that frequent pain or discomfort in newborn infants, who are at a time of physiological immaturity and undergo rapid brain development, can cause complications, such as delay in cognitive and motor development [41]. Cumulative pain/stress experienced at an early-life stage significantly contributes to abnormal brain development, which can yield long-term adverse neurodevelopmental outcomes [24] [108] [173] [11] [121] [100]. Significant and long-lasting consequences following pain/stress in the newborn periods can change the central nervous system and responsiveness of the neuroendocrine and immune systems, which can lead to stress many years later, even at maturity [42] [103]. Furthermore, for infants born preterm, neonatal pain-related stress is associated with alterations in both early and later developmental outcomes [161]. Neurobiological vulnerability to pain in newborn infants is well established, due to their lower pain threshold, sensitization from repeated pain, and immature systems for maintaining homeostasis [34] [35]. Therefore, in the Neonatal Intensive Care Unit (NICU), continuous discomfort or pain assessment for newborn infants is highly desired, since it helps caregivers to understand the severity of infant situations and develop appropriate treatments. Reliable discomfort detection can play a crucial role in appropriate and timely treatment in pediatric clinics. In the meantime, it also contributes greatly to the well-being of the infants for a better early-life quality.

Self-reporting is currently considered to be the gold standard for pain assessment among patients, and is the most reliable indicator of the existence and intensity of acute pain and discomfort [99]. Assessment scales that rely on patient self-reporting are commonly used to measure the intensity of pain. However, infants cannot report verbally their pain or discomfort and must rely on healthcare professionals to recognize their behavioral or physiologic signs suggesting pain and discomfort [61]. Because of this lack of communication, the value of discomfort detection is even further increased.

1.2 Current measurements in clinical setting

At present, discomfort and pain monitoring for infants is performed manually and visually by healthcare professionals, who check both the behavioral parameters (e.g. facial expression, body movement and crying sound) and/or the

physiological parameters (e.g. vital signs of the infants, like breathing, heart rate, body temperature) of infants.

1.2.1 Vital signs

Currently, vital signs of NICU infants are continuously monitored by sensors and transmitters in combination with a single central monitor (see Figure 1.1). Vital signs include heart rate, respiration rate, blood oxygen saturation, and body temperature, which are measured by the sensors and electrodes attached to infant skin. However, this approach of measuring parameters can damage the vulnerable skin of neonates and cause infections, which adds an extra burden. Moreover, the wires can also interfere with the clinical staff and parents when taking care of the infants. These issues initiate the objective for a friendly, contactless measurement of vital signs. Moreover, the alarm of a professional monitoring system is only triggered when the monitored parameters exceed the boundaries of normal ranges. When infants experience discomfort moments, such as feeling hungry, needing a diaper change, missing mom, or even disease-induced pain, the system cannot notify clinical staff immediately as the vital signs may stay within the normal range.

1.2.2 Comfort scales

In common practice, there is no universal or standard method to monitor and assess discomfort/pain. Currently, the stress/comfort levels of hospitalized infants are regularly checked by caregivers and clinicians. For pediatric units, multiple pain/comfort-scale forms have been developed, in order to assist healthcare professionals in detecting the level of pain or discomfort.

For term neonates [1], there are well-known discomfort/comfort scale tools, which are also reported in literature. For example, the Comfort Scale [5] considers both observational and physiological factors, while the Neonatal Infant Pain Scale (NIPS) [75] is based on interpreting facial expression, crying, breathing patterns, and upper and lower limb movement, etc. The Objective Pain Scale (OPS) has been utilized with infants and children from birth to three years of age [98]. For preterm neonates, the Premature Infant Pain Profile (PIPP) score [138] is a multi-dimensional measure developed to assess acute pain. Table 1.1 summarizes the above-mentioned validated pain/discomfort tools for neonates.

[1] The word “term" means here non-preterm.

Figure 1.1: – *Visual example of a NICU setup at the Máxima Medical Center, Veldhoven, the Netherlands. The subpicture at the left shows the physiologic parameters measured and shown on the display for monitoring.*

Name of the scale	Incl. phys. measures	Preterm	Term
Comfort Scale	Yes	No	Yes
Objective Pain Scale (OPS)	Yes	Yes	Yes
Neonatal Facial Coding System (NFCS)	Yes	Yes	Yes
Neonatal Infant Pain Scale (NIPS)	Yes	Yes	Yes
Premature Infant Pain Profiles (PIPP)	Yes	Yes	Yes

Table 1.1: – *Validated scales and scoring systems for neonatal pain/discomfort monitoring (phys. means physiology).*

The scale-based assessment is typically performed by caregivers/clinicians by observing infant behavior for 2-3 minutes. However, visual assessments are only scheduled for a few times a day. The intermittent assessment can lead to under/misdiagnosis and therefore delay in treatment or incorrect therapy. Another important aspect of detection is that the current procedure is based on the subjective assessment of personnel, which incurs inter-assessor variation.

Since a continuous discomfort/pain assessment tool is not available, an automated monitoring system that can continuously detect discomfort is highly desired to replace the current intermittent manual observation, such as by a camera-based health monitoring system.

The existing pain scales are valuable for this research work, since they can help and guide the development of video-based automated monitoring systems. The existing scales for assessing infant status can give pre-information about which infant features are important to observe and assess discomfort, for example, with regards to facial appearance, body movement, etc. The existing comfort scales are assessed by a human. In the framework of this thesis, the purpose of developing an automated video-based monitoring system is to offer an assisting artificial care-evaluator. In other words, the system is expected to extend the visual system of humans over a continuous time period and mimic the presence of an observing caregiver.

The followed approach is that for our work, we are inspired by the aforementioned comfort scales, to adopt important features from them if useful, but we are not bounded by them in the sense that sometimes technology can offer more possibilities than existing human-based monitoring. It is expected that the facial features of the infants can provide important information indicating

the infant's discomfort status, since parts of these scales are based on facial expressions. However, it is of interest to also consider additional parameters that are important for pain/discomfort assessment, such as movement behavior patterns, and also the physiological parameters, e.g., heart rate and respiratory rate of the infant.

1.3 Techniques for human monitoring

In the past several years, there has been an increased interest in the development of learning systems for understanding human behavior related to pain/distress. Advanced methods for recognizing emotion or behavior have been mostly developed for adults. However, such methods for adults can be leveraged for infant monitoring. The existing behavior monitoring systems can be classified into two categories: 1) physiological-based and 2) behavioral-based analysis.

Various approaches have been developed to extract physiological parameters including Heart Rate (HR), Respiratory Rate (RR), blood oxygen saturation (SpO2), body temperature, and blood pressure (see Table 1.2). These methods are typically based on contactless or less invasive manners. With extracted physiological indicators, the human-health status can be further assessed. The aforementioned parameters/signs can be measured and used to assess the person's level of physical functioning. Besides the physiology-based methods, behavior-based systems focus on the analysis of facial expression, body motion, and crying sound, which are mostly extracted from video and audio signals.

It is hypothesized that motion behavior induced by discomfort/pain can be measured from video signals, provided that the cameras are of sufficient quality and resolution and the camera setup allows to make informative recording for further analysis. Good techniques are existing for extracting motion information, such as optical flow, which will be introduced later in this thesis. Starting from this view point, when applying optical flow-based motion analysis, it is our intention to reuse the technique of extracting motion information for the quantitative analysis of respiration measurement. However, our study on exploiting physiological parameters for the sake of discomfort detection, will not be exhaustive from the side of using physiological parameters. This thesis just offers the first work on this exploration (e.g. respiration rate measurement) for infant monitoring and we consider that more work will follow afterwards.

Physiological parameters	*Techniques*	*Comments*
Heart rate [104]	Smart video	The system requires the user to place the tip of his/her index finger on the lens of a smart phone camera, while the flash is on.
Respiratory rate [133]	Respiratory sounds	A plurality of respiratory rates are derived from the recorded sounds and then a heuristic is applied to select one of the derived respiratory rates.
Blood oxygen saturation [20]	Reflectance pulse oximeter	The results show that the heart rate and blood oxygen saturation measured by the proposed design are corresponding to the readings of the standard monitor.
Body temperature [21]	Conductive textile wires	Accurate temperature monitoring obtained by a prototype belt.
Blood pressure [58]	High-speed camera	High intra-individual correlation between pulse transit time and blood pressure.

Table 1.2: – *Examples of physiological parameter measurements.*

1.4 Support of technology developments

Nowadays, camera-based video surveillance systems are being used everywhere, notably in the surroundings of our daily living, like traffic surveillance and safety applications. Cameras are installed for different surveillance purposes on the street, in factories, in train stations, and even in nursing houses. Those cameras are often connected with remote monitoring systems via the Internet, where human experts typically observe the videos and recognize a face, detect incidents, track objects, etc.

In the past ten years, lenses of cameras have become smaller and cheaper, which is also fueled by mobile phones and networking, which can support video streaming. The surveillance is also continuously getting smarter in the

analysis of pictures and videos, due to the advances in computer vision and machine learning. Furthermore, mobile video surveillance systems have become popular, and the captured video can be transmitted to its final destination under specified delay constraints when needed.

Surveillance applications have been largely boosted because of the development of Computer Vision (CV), which models and analyzes the semantic content of videos. The CV systems are first developed on the basis of supervised learning, and then can identify the recurring activities within the videos or recognize the abnormal events in the video contents, such as an incident in the traffic, falling events of humans, etc. Besides recognition applications, one of the most interesting areas of research is tracking and identification of objects. The goal of these CV algorithms is to define bounding boxes containing an object of interest from each frame of the video. For example, CV can be used to recognize patterns between human body movement and pose over multiple frames in video footage or real-time video streams. It can be used also to track the pose and movement of multiple team players in a match calculated from both monocular and multi-view sports video datasets. Conventional machine learning has evolved well into localizing and classifying objects of interest in real-time. However, the accuracy and robustness of various computer vision techniques remains a topic of research for further improvement.

In the past few years, because of the advances in algorithms and computation power, Convolutional Neural Networks (CNN)-based deep learning methods have become the latest machine learning technology in many fields including computer vision. Deep learning is replacing conventional computer vision techniques in object detection, action/activity recognition, motion tracking, human pose estimation, and semantic segmentation. The global term for this field of technology is called Artificial intelligence (AI). Within this large field, video- and image-based analysis forms a specific area on its own.

In the healthcare domain, videos can be also leveraged for infant care. A machine learning or deep learning-based solution is desired to be available and reliable for real-time monitoring. The objective of this monitoring system is to assist clinicians in their observation tasks by performing a continuous surveillance and automated alerting when needed. The system can detect baby discomfort or other stressful moments and filter them out from all other surrounding events. In this way, the clinicians are always notified when the infants need care.

To summarize, in the past few years, with the improvement of the camera quality, video processing techniques have been leveraged in different computer vision applications, such as license-plate recognition, pedestrian monitoring,

and speed testing. In the healthcare domain, the application of video monitoring is still in its infancy, but offers promising possibilities. The purpose of this thesis is to leverage from these developments and design computer vision techniques for better healthcare and for benefiting clinicians, parents, and infants.

1.5 Problem statement and research questions

This section decomposes the imposed clinical questions into a series of technical research problems and formulates each problem into underlying research topics.

1.5.1 Problem statement

The purpose of the research in this thesis is to exploit the AI technology for the healthcare surveillance on infants. Here, this section first poses the problem statement and then details it into research questions.

The objective of this thesis is to develop a video-based automated discomfort detection system that is able to monitor and interpret infant discomfort status. This implies advanced video analysis on both facial features and infant behavior components. The analysis techniques should be learned by an automated system in order to come to a knowledgeable decision on the discomfort status of the infant.

Considering the practical factors for the implementation of a clinical application of a surveillance system, important system requirements need to be specified. (1) The system needs to be sufficiently accurate to be accepted by the clinicians (e.g. for clinical trials). (2) The system should operate with a fast execution time to allow real-time monitoring without delaying diagnosis. (3) The cost should be feasible and the camera(s) should be installed at a position(s) rarely interfering with clinical treatments. The proposed system has to be evaluated thoroughly against these requirements. The validation experiments and results will provide an insight into the feasibility of incorporating such an intelligent assisting system for future clinical application.

With the desire of the assisting system and posed system requirements, we further detail the research aspects that need to be addressed for developing such a system.

1.5.2 Research questions

From the above problem statement, a number of specific research questions can be derived to support an in-depth analysis in the field of (preterm) infant monitoring, which are formulated as follows.

RQ1: Facial features for frame-level discomfort detection Many emotion recognition systems have been developed by analyzing facial expressions. However, most of this research focuses on surveillance of adults. The effects of facial information to distinguish infant discomfort from comfort have not been established.

Most of the infant videos were recorded when infants were lying on their backs (supine position). However, they are still free to rotate their heads. Therefore, considering the different sizes of infant faces and various head poses, the system should be robust against variation in the face sizes and head-/face poses among different infants. This leads to the following research questions.

- RQ1.a: Which facial features are relevant and discriminating for characterizing the facial expression for the infant comfort/discomfort detection task?
- RQ1.b: How should the facial features be normalized across different infants, since infants have different levels of facial expressions?

RQ2: Deep learning-based frame-level discomfort detection Deep learning techniques have shown impressive performance in multiple computer vision applications. The performance of deep networks is unknown regarding the task of infant discomfort detection, leading to the following research questions.

- RQ2.a: Can deep learning be used for this infant comfort/discomfort classification task? If so, in what way?
- RQ2.b: Can deep learning be applied on a small dataset or are special actions required to facilitate this?
- RQ2.c: Is deep learning sufficiently robust to environmental settings and changes?

RQ3: Incorporating temporal information - motion analysis Body movement is an important indication for comfort or discomfort when clinical experts visually assess infant status. The use of temporal information extracted from video sequences needs to be validated by machine learning methods. Information embedded in the video sequences enables a meaningful interpretation of the body motion signals. CNNs are capable of automatic and deep learning of features directly from input data. Three-dimensional (3D) CNNs can capture both the object appearance and contextual information, together with motion information. When applying 3D CNNs on the infant videos without explicit detailed steps on face detection/normalization, the performance of the overall system is not known at the time and needs research. This leads to the following questions.

- RQ3.a: Can we leverage the motion information for infant discomfort detection?
- RQ3.b: Can we employ optical flow and handcrafted features for the extracted motion signal to capture the temporal information?
- RQ3.c: In what way can deep learning facilitate the classification of a 1D motion signal?
- RQ3.d: Can we use CNNs to jointly process both spatial and temporal information?
- RQ3.e: How can an additional motion channel give guidance to CNNs to focus on areas related to discomfort?

RQ4: Quantitative physiological signal measurement The tasks mentioned above are all for the binary classification of comfort and discomfort. The feasibility of video-based quantitative physiological measurements providing informative value of respiratory rates, is an important research topic for this thesis. Only using video-based measurement offers clear benefits of contactless measurements on the infants, so that it has a clear preference for further investigation. The related research questions are formulated as follows.

- RQ4.a: How can we quantitatively extract physiological signals from video recordings?
- RQ4.b: Are the results calculated from videos reliable by comparing them with the current standard methods (chest impedance)?

1.6 Contributions of this thesis

This section summarizes the overview of the scientific contributions presented in this thesis. The contributions are described in three perspectives, which are elaborated below.

Contributions to facial feature analysis for discomfort detection: Our research has investigated whether the facial features are discriminative to distinguish infant discomfort from comfort moments.

An automated system is proposed for facial expression recognition for infants during in-hospital care. Moreover, this study proposes to normalize infant faces for effective feature computation. To better take into account variations between each individual, we incorporate landmark-based template matching for enhancing the performance. Furthermore, deep learning is employed to improve the performance by self-learning of the facial features. This system applies a deep learning model based on pre-trained CNNs, followed by a fine-tuned learning process. The performance of the deep-learning model is improved when using the proposed fine-tuning steps, involving pre-training with generic people pictures and dataset balancing combined with twofold cross-validation. Using all refinements, the Area Under the Curve (AUC) then substantially increases from 0.77 to 0.96.

Contributions to video-based AI systems for infant discomfort detection: Going from 2D images to full videos, the thesis provides solutions to perform fully analyzing video data without face detection/tracking. First, a preliminary AI solution is developed to leverage the extraction of the motion information embedded in the video segments. Sequentially, we propose a multi-channel input to the learning network to help draw the attention of the networks to the regions with significant movement related to facial expression changes, where spatial and temporal information are learned jointly. The obtained attention-based learning network is an innovative system implementation for neonatal observation.

Contributions to the quantitative physiological signal measurement from videos: For further supporting the classification of comfort status, we have developed a contactless video-based physiological analytic method, for which an AI system is designed that can quantitatively extract respiration rate for infants. This solution allows healthcare personnel to create more in-depth understanding of the comfort status of infants.

1.7 Thesis outline and corresponding scientific publications

This section presents an outline of the chapters in this thesis and briefly discusses the contributions of each chapter. The corresponding scientific publications from each chapter are also indicated.

Chapter 2 presents the system of image-based facial expression analysis, using conventional machine learning techniques. The system extracts handcrafted features characterizing both geometrical and facial appearance information. Moreover, template matching-based features are introduced for subject-dependent detection.

The contribution of this chapter was published in a journal article in Machine Vision and Applications (2019) [145].

Chapter 3 exploits deep CNN algorithms for the task of image-based analysis. To address the limited available data for training, a pretrained model is utilized, which is followed by training the networks using a public dataset with labeled facial expressions for pain assessment (first fine-tuning). The model is further refined with a self-recorded dataset of infant videos (second fine-tuning).

The contribution of this chapter was published in a journal article in Physiological Measurement (2019) [147].

Chapter 4 investigates an automated and continuous discomfort detection method, based on analyzing motion patterns of preterm infants. For each video segment, the motion matrices are first estimated from adjacent video frames using optical flow. Then, the motion acceleration rate is calculated and 18 time /frequency-domain features are extracted for characterizing motion patterns. A support vector machine (SVM) classifier is then applied to video sequences to recognize infant status to be comfort or discomfort.

The contribution of this chapter was published at the IEEE Int. Conf. of the Engineering in Medicine and Biology Society (EMBC) 2019 [141].

Chapter 5 describes the infant body motion by two-dimensional (2D) time-/frequency-domain representations characterizing the distribution of signal energy. Log Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCCs) and Spectral Subband Centroid Frequency (SSCF) features are calculated from

the one-dimensional (1D) motion signal. Deep CNNs are further applied to the 2D images for the binary classification of comfort or discomfort.

The contribution of this chapter was published at SPIE Medical Imaging in 2020 [140].

Chapter 6 investigates the video characteristics (e.g. intensity images and motion images) and the associated CNN architectures (e.g. 2D and 3D) for infant discomfort detection. The realized improvements of the 3D CNN are based on capturing both the motion and the facial expression information of the infants, and the joint classification of those information parts.

The contribution of this chapter was accepted for publication by the Journal of Quantitative Imaging in Medicine and Surgery in 2021 [139].

Chapter 7 proposes an automated pipeline to estimate respiration signals from videos for premature infants in neonatal intensive care units (NICUs). The conventional optical flow and deep learning-based flow estimation methods, are employed and compared to the estimation of motion information between adjacent video frames. The respiratory signal is further extracted via motion factorization. The feasibility of quantitative physiological measurement in infants based on deep learning methods is demonstrated.

The contribution of this chapter was published in the Journal of Applied Science in 2019 [149].

Chapter 8 summarizes this thesis and discusses the research questions of Chapter 1 and the findings in the thesis. The results of this thesis are finally concluded and a brief outlook is presented regarding video-based infant monitoring applications.

Chapter 2

Handcrafted features for facial expression analysis

2.1 Introduction

The previous chapter has outlined the scope of this thesis by introducing the clinical background and the desire for supportive systems in infant monitoring to timely detect their discomfort status. This chapter starts with describing a system based on conventional methods exploiting handcrafted features and a machine learning classifier for the task of infant discomfort detection. Most machine learning systems share a similar processing chain, where each processing step performs a specific function, giving the components of data input, information extraction, and finally a prediction from the processing analysis. Regarding the second step of information extraction, existing computer vision algorithms are based on the detection and extraction of local features in images/videos. The extraction of so-called handcrafted features involves algorithms that are committed to understand, quantify, and improve features extracted from images/videos. Conventional handcrafted features are manually designed by researchers, so that they are easy to be interpreted and intuitive for human understanding. Handcrafted features can be also task-oriented for a specific application, which requires that researchers should have domain-

knowledge prior to designing any features.

As a sequential step, the designed features are usually combined with a machine learning classifier to perform the classification tasks in supervised learning. For example, Support Vector Machine (SVM) [155] is a popular machine learning classifier, that maximizes the distances from each class to the decision boundary. This approach works effectively and robustly if the margin separation is clear between the classes. SVM is more effective in high-dimensional spaces and is also memory-efficient.

This thesis aims at exploring facial expressions of neonates to detect the comfort status of the neonates. An important part of such detection is to analyze the facial expressions of the infants, in order to come to a decision on their comfort status. This facial expression is explored by designing handcrafted features and comparing the performance between different categories of features.

The objective of this chapter is to analyze infant facial expressions shown in video frames by exploiting feature extraction and classifying those features for estimating the comfort status of the infants. The facial expression is one of the most common indicators of discomfort and pain for infants. The Primal Face of Pain (PFP), shown in Figure 2.1, is a universal facial expression, associated with pain, which is hardwired and presents at birth [126]. Considering the input feature vectors, the final prediction of comfort/discomfort will be made by an SVM classifier. In more detail, the following challenges and requirements of such a discomfort detection system should be addressed.

- *Normalization of head pose:* With respect to the behavior of the infants, our approach relies on the system requirement on infant head pose, which has been described in Chapter 1. Most of the infant videos are recorded when infants are lying on their backs (supine position), but they are still free to rotate their heads. Therefore, considering the different sizes of infant faces and various head poses, the system should be robust against variations in the face size and head/face position among different infants. This means that an extra normalization step is needed prior to feature extraction to ensure proper system operation. Therefore, the first challenge for the design of the system is the normalization step that should handle the variations in size and pose to enable feature extraction.

- *Selection of expression-related features:* As the PFP is an indicator of infant discomfort status, the employed handcrafted features should specifically capture the typical patterns of PFP-related facial information, that differentiates discomfort status from comfort. In order to facilitate this, we distinguish facial expression-related features and texture features from the

face. It is expected that for defining proper handcrafted features, the performance of facial expression-related features and the group of features for describing texture from the face have to be explored and compared.

- *Robustness against inter-infant differences:* Pediatricians have observed that when infants experience discomfort/pain, the levels of face distortion due to pain experience vary greatly among different infants. For a selection of premature infants, the changes of facial expression from comfort to discomfort are hardly visible. Therefore, the system should be able to handle these inter-infant differences for maintaining reliable comfort detection.

Taking the above requirements and challenges into account, our aim is to develop an automated system that classifies infant comfort/discomfort status using the described handcrafted feature groups followed by an SVM. The focus on the classification is then on features describing the PFP, which concentrate on the facial expression-related features.

The remainder of this chapter is organized as follows. In Section 2.2, related work on pain and discomfort detection is described. Section 2.3 elaborates our method, and Section 2.4 explains the experimental results, with some discussions in Section 2.5. Finally, Section 2.6 concludes the chapter.

2.2 Related work

2.2.1 Face detection

Face detection has been one of the most studied topics in the field of computer vision. From the past, the Viola-Jones algorithm [166] is a well-known successful face detector, which can be applied even in real-time applications. The three main components of the Viola-Jones detector include (1) integral image, (2) classifier learning with AdaBoost, and (3) attentional cascade structure. The integral image is also known as a summed-area table, which can efficiently compute the sum of values in a rectangular subset of a grid. The main concept of boosting is to find a single and highly accurate hypothesis by combining multiple simple hypotheses (called “weak" classifiers), each having moderate accuracy. The attentional cascade structure is a critical component in the Viola-Jones algorithm. The mechanism is the construction of small, and thus efficient, boosted classifiers, which results in rejecting most of the negative sub-regions of the objects, while keeping almost all the positive object samples. Therefore,

most of the sub-regions will be rejected at the early stages of the detection process, leading to an extremely efficient procedure. For the object detection, in our case, this task aims at face detection.

SVM has also attracted much attention on the topic of face detection because of its fast computing and good performance [102]. Jee *et al.* [57] proposed a real-time face detection system for personal authentication using color, edge, binary information, and SVM classifiers. The system detects facial region by simultaneously using skin color and edge information. Candidates of eye pairs are further detected by exploiting edge and binary information in the facial region. False detection is prevented by verifying eye and face candidate regions using SVMs.

Shavers *et al.* [132] present a face detection system also based on SVM. The λ-coefficients corresponding to the SVM support vectors are determined from training a set of images. The support vectors are the archetypes for face and non-face images. The SVM algorithm maps the test images (containing both face and non-face images) into a high-dimensional transform space where a hy-

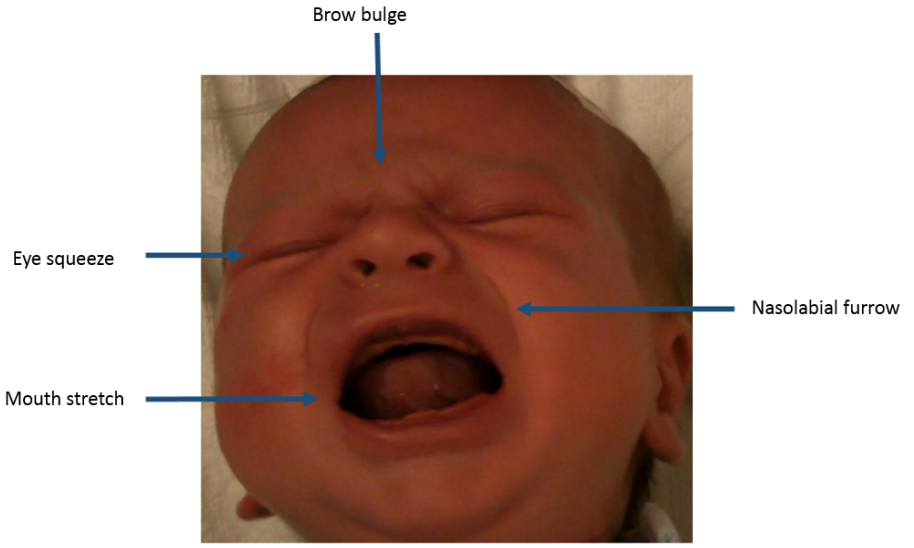


Figure 2.1: – *Primal face of pain, involving brow bulge, eye squeeze, and a horizontally stretched open mouth with deepening of the nasolabial furrow.*

perplane decision function is constructed. The decision function is equidistant between support vector archetypes, in order to create an optimal hyperplane decision function.

For background reading, a large amount of papers on face detection are captured in the review papers [159] [169] on handcrafted features with SVMs. More recent papers on face detection using deep learning will be discussed later in this thesis. However, nearly all the existing methods are developed for adult faces and not infant's faces.

2.2.2 Pain/discomfort detection

A. Metadata parameters on discomfort detection

In the past years, emerging interesting applications [45] have fueled the increased interest in pain assessment. There is a category of discomfort detection methods based on metadata parameters, that assess pain or discomfort based on behavior analysis. Existing approaches of metadata parameters can evaluate infant pain based on crying sound and body motion. Infant crying is a common sign of discomfort, hunger, or pain. For classifying the crying sound, Mima *et al.* [94] presented a method that analyzes baby cries within spectrographic images, and classifies these cries into pain, sleeping, hunger, etc. The obtained overall accuracy of the proposed method was 85%. Up to this point, very few studies on pain assessment based on body movements have been published [183] [185].

B. Physiological parameter-based methods

Various approaches have been devoted to assess pain based on physiological indicators, for instance, vital signs such as Heart Rate (HR), Heart Rate Variability (HRV), Respiratory Rate (RR), blood oxygen saturation (SpO2), body temperature, and blood pressure. The signs can be measured and used to assess the person's level of physical functioning. Lindh *et al.* [81] presented an approach to assess infants pain by frequency-domain analysis of HRV during heel lancing, which showed an increase in low-frequency power in the heel stick response of preterm infants compared to baseline. Acharya *et al.* [2] detected cardiac abnormalities by classifying cardiac rhythms, using an artificial neural network and fuzzy relationships, which achieved an accuracy level of 80-85%. However, vital signs such as HR and RR are currently measured using

techniques including electrocardiograms (ECG) and pulse oximetry, which require contact with the patient's skin. Attaching the sensors to infant skin adds an extra burden to infants, compared to the contactless methods using videos.

C. Facial expression-based analysis

Substantial attention has been paid to facial expressions in adults. Shan *et al.* [131] empirically evaluated facial representation based on statistical local features like Local Binary Patterns (LBPs), for person-independent facial expression recognition and illustrated that LBP features are effective and efficient for facial expression recognition. They achieved a recognition rate of 91.4% for 7-class facial expressions (anger, disgust, fear, joy, sadness, surprise, and neutral) by using Boosted LBP-based SVM with an RBF kernel. Kotsia *et al.* [73] achieved a recognition accuracy of 99.7% for facial expression recognition, using the proposed multi-class SVMs and 95.1% for facial expression recognition based on a set of chosen facial Action Units (AUs). Kharghanian *et al.* [66] applied a hierarchical unsupervised feature learning approach to extract the features needed for pain detection from facial images using a convolutional deep belief network and achieved near 95% for the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Ashraf *et al.* [7] explored an approach for automatically recognizing acute pain with Active Appearance Models (AAM). The shape and appearance components represented by AAM were further decoupled to separate features. Finally, SVM was employed for classification. The method was evaluated on 15,761 frames from the UNBC-Mcmaster shoulder pain database, which showed a frame-level accuracy value of 82.4% for detecting pain frames, and a false positive rate of 30.1%. One of the challenges in this task is to align faces together for more effective feature extraction. Lucey *et al.* [89] showed that the AAM can deal with patient movements and can achieve significant improvements in both the facial AU and pain detection performance. Using the UNBC-McMaster database, the obtained AUC was 0.847 when combining similarity-normalized shape features (SPTS) with similarity-normalized appearance features (SAPP) and canonical normalized appearance features (CAPP). Littlewort *et al.* [82] applied an automated facial expression recognition system to spontaneous facial expressions of pain. A set of 20 detectors from the facial action coding were extracted and passed on to a classification stage, in which a classifier was trained to detect the difference between expressions of real pain and fake pain. The automated system obtained an accuracy of 88% for subject-independent discrimination of real versus fake pain.

Most of the existing methods for automatic pain assessment based on facial expressions focus on adults. However, the methods designed for assessing adult pain may not have similar performance on the infants, because the facial morphology and dynamics vary between infants and adults as reported in [43]. Sikka *et al.* [134] presented a Facial Action Coding System (FACS)-based method to describe children's facial expressions of pain. The model detection of pain versus no-pain achieved an AUC of 0.84–0.94 in both ongoing and transient pain conditions on the videos from 50 children. One challenge of FACS-based methods is the extensive time required for labeling AUs in each video frame. Fotiadou *et al.* [36] proposed an infant discomfort detection system based on the AAM. The system was evaluated in 15 videos of 8 infants, yielding an AUC performance of 0.98. Unfortunately, the AAM mesh has to be initialized for each individual baby by identifying a set of landmarks manually, which makes it difficult to implement in clinical practice.

Summarizing the related work in the context of this chapter, we concentrate on the techniques of extracting facial features and then processing these features with an SVM. The FACS-based method is not considered in our research, because the working scheme requires manual labeling of AUs for each video frame, which increases the complexity of the process and is time-consuming. For the AAM-based approach, the initial step needs manual identification for facial landmarks, which is not attractive as it does not provide a generic solution. Infant discomfort is generally assessed by observing their facial expression. Therefore, the handcrafted features referring to PFP will be designed and utilized to discriminate facial information of comfort and discomfort. Conventional handcrafted features (e.g., LBPs) will also be included in this study because of its proven efficiency for facial expression recognition. Once the handcrafted features are obtained, an SVM classification will be directly applied because SVM would find the optimal margin gap between separating hyperplanes and is also computationally efficient.

2.3 Methods for handcrafted feature-based classification

This section describes the designed method for an automated discomfort detection system based on handcrafted feature extraction aiming at describing the PFP and classifying those features into a comfort/discomfort class. The

proposed method involves three key steps listed below.

- *Face normalization:* The face region is first detected by identifying 68 facial landmarks. This is followed by resizing and rotating the region according to the location of the facial landmarks, to correct the face size and rotation variance. In this way, a normalized face ROI is obtained.
- *Handcrafted feature extraction:* To extract the features characterizing the PFP-related facial information, geometric features calculated from the 68 facial landmarks are leveraged. Appearance features from the facial ROI are also extracted for describing the facial texture. These features are finally supplied to an SVM [16] [38] classifier.
- *Subject-(in-)dependent detection:* A subject-independent detection scheme is proposed. In addition, we further introduce template matching-based features for subject-dependent detection, which compensates the inter-infant variance.

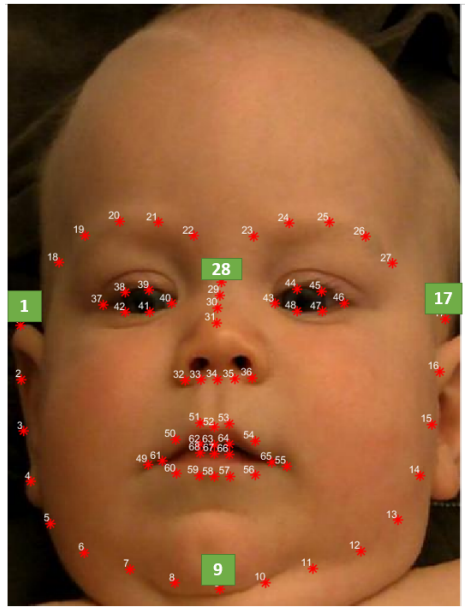
The above properties will be further explained in detail in the following subsections.

2.3.1 Face detection and normalization

The subsection starts with extracting the infant face from each frame of the video. The complete face detection and normalization workflow generates the face ROI as input for the next step of discomfort/comfort classification. Given an input video frame of a face image, 68 facial landmarks are first localized using the Dlib facial landmark detector [67], implemented by Kazemi *et al.* [64]. We utilize the Dlib face landmark detector because it detects the face with a rich number (68) of landmarks [54] and its implementation is also efficient [19]. The 68 landmarks are points on the face such as the eye corners, mouth, along the eyebrows, along the face boundary, and so forth.

Figure 2.2 shows an example of an original face image and a processed version. Subfigure 2.2 (a) indicates the original face, while Subfigure 2.2 (b) depicts the corresponding normalized face ROI and 68 facial landmarks. Once the 68 landmarks are identified, we select the central point between two inner eye-corner points as the point to rotate the image. The image is rotated to the position that the two eye-corner points are along the same pure horizontal line. Thereby, this processing step corrects the rotation variance of faces. Then we select Landmark 1 as the leftmost point, Landmark 17 as the rightmost, and Landmark 9 as the bottommost points to define the left, right, and

(a)



(b)

Figure 2.2: – *Derivation of landmarks of a face image. (a) Example of face image, (b) corresponding normalized face ROI and 68 facial landmarks. The highlighted Landmarks 1, 9, and 17 are used for defining the face ROI. Landmark 28 is employed in Phase 2 for further aligning sets of landmarks from different frames.*

bottom boundaries of the face ROI (see Subfigure 2.2 (b) for the landmarks). We measure the distance between Landmark 9 and the mid point of the two inner eye corners. The top boundary of the ROI is defined to be the horizontal line that has the same distance from the mid inner eye point as Landmark 9. A margin of 20 pixels is added to all the boundaries to cover the whole infant face, and avoid loss of facial information. Finally, all face images are cropped and resized to the size of 500×375 pixels.

2.3.2 Phase 1: Subject-independent discomfort detection

Geometric and appearance features are extracted from the facial ROI for discomfort detection using the SVM classifier.

A. Geometric features

In general, when infants start suffering from discomfort, they tend to squeeze their eyes and stretch their mouths, as is shown in Figure 2.1. In order to extract relevant features, we calculate the areas of eye and mouth. Then we count the number of pixels inside the polygons surrounded by the landmarks of the left eye, right eye, outer lip contour and inner lip contour, respectively (see Figure 2.3). These four area sizes are considered as four geometric features. Figure 2.4 shows the distributions of geometric features for comfort and discomfort cases of our dataset.

B. Appearance features

The appearance of an infant face is changing when experiencing discomfort. For example, as shown in Figure 2.1, brow bulge and nasolabial furrow become apparent in the discomfort faces, which leads to texture changes in the forehead and between the nose and upper lip area. Figure 2.5 compares the face appearance of comfort and discomfort status of one infant. Different texture descriptors have been exploited for facial expression recognition. Two methods that have proven to be effective for facial representation are Histogram of Oriented Gradients (HOG) [28] [167] and Local Binary Patterns (LBP) [101] [4] [130]. As the histogram of gradient descriptor, HOG has been widely utilized to capture edges or local shape information [86]. LBP is invariant to monotonic gray-level changes and is highly discriminative, which makes it suitable for demanding image analysis tasks such as object detection. In this work, HOG and LBP are combined as the feature set to describe faces for discomfort detection. In

detail, for the HOG calculation, the input face is divided into 20×15 equal non-overlapping regions. For each region, the HOG is computed (see Figure 2.6). The final HOG vector features are formed by concatenation of HOG vectors of regions. The length of the HOG feature vector is 5,625 elements. For LBP, each face image is divided into blocks of 100×75 pixels to which the uniform LBP operator [101] is applied, and LBP histograms are extracted and concatenated into a single, spatially enhanced feature histogram (see Figure 2.7). The length of the LBP feature vector is 1,475 elements.

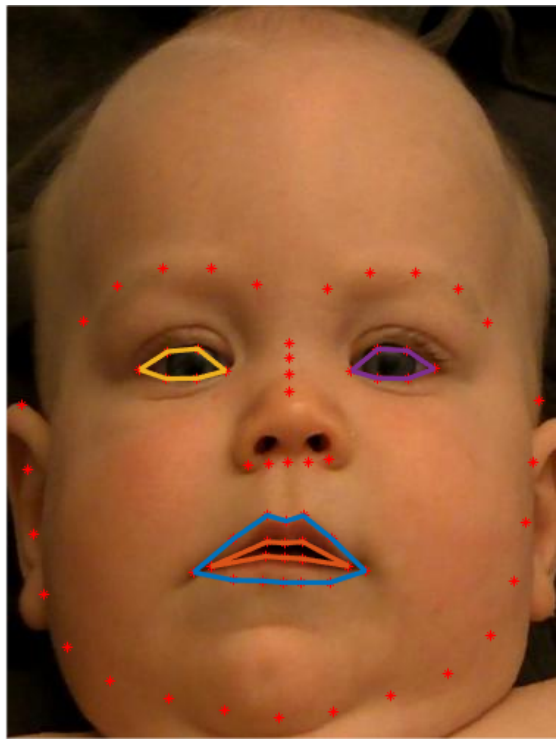
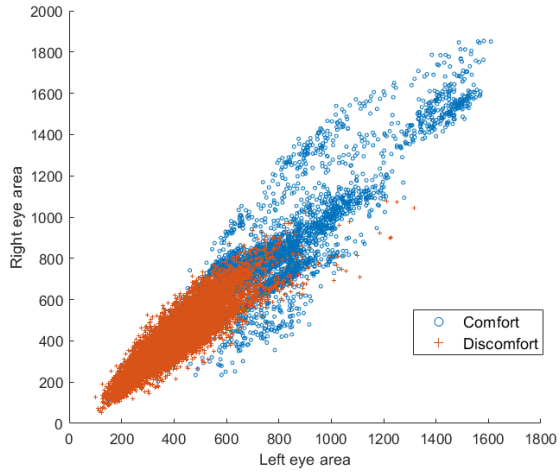
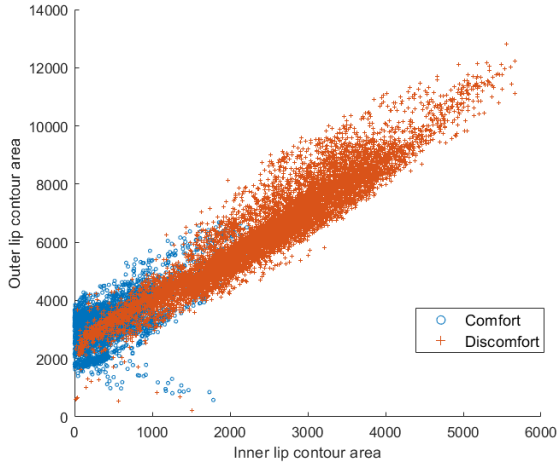


Figure 2.3: – *Polygons outlining facial landmarks. The polygons are defined by landmarks for eye- and mouth-area calculation, where a yellow polygon outlines left eye, purple for right eye, blue for outer lip contour, and orange for inner lip contour.*



(a)



(b)

Figure 2.4: – *Scatter plots of geometric feature values of eye and mouth. Comfort (blue-circled dot) and discomfort (orange plus sign) cases for all infant face images in our dataset. (a) Plot of feature values of left and right eye area. (b) Presentation of the values of the inner and outer lip-contour area.*

C. Classification

The large size of HOG/LBP feature vector limits the number of training samples and increases the computation cost in classification. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [15]. The selection reduces the number of features and removes irrelevant, redundant data and even noisy components. In our work, feature selection is used to reduce the dimensionality of the input feature space and thus enable the subsequent use of classification algorithms. The AUC criterion is chosen to identify relevant significant features, which means features selected by optimizing the AUC.

Finally, we adopt an SVM [16] [38] with a radial basis function (RBF) as the kernel, to recognize facial expressions using the selected features. We employ the SVM implementation in Matlab (Mathworks, Natick, MA, USA) for the binary classification. Leave-one-subject-out cross-validation is used for the experiments. The ROC is plotted to evaluate the performance with the value

Figure 2.5: – *Example of a comfort (left) and a discomfort (right) case from an infant, which shows the face appearance difference between comfort and discomfort status.*

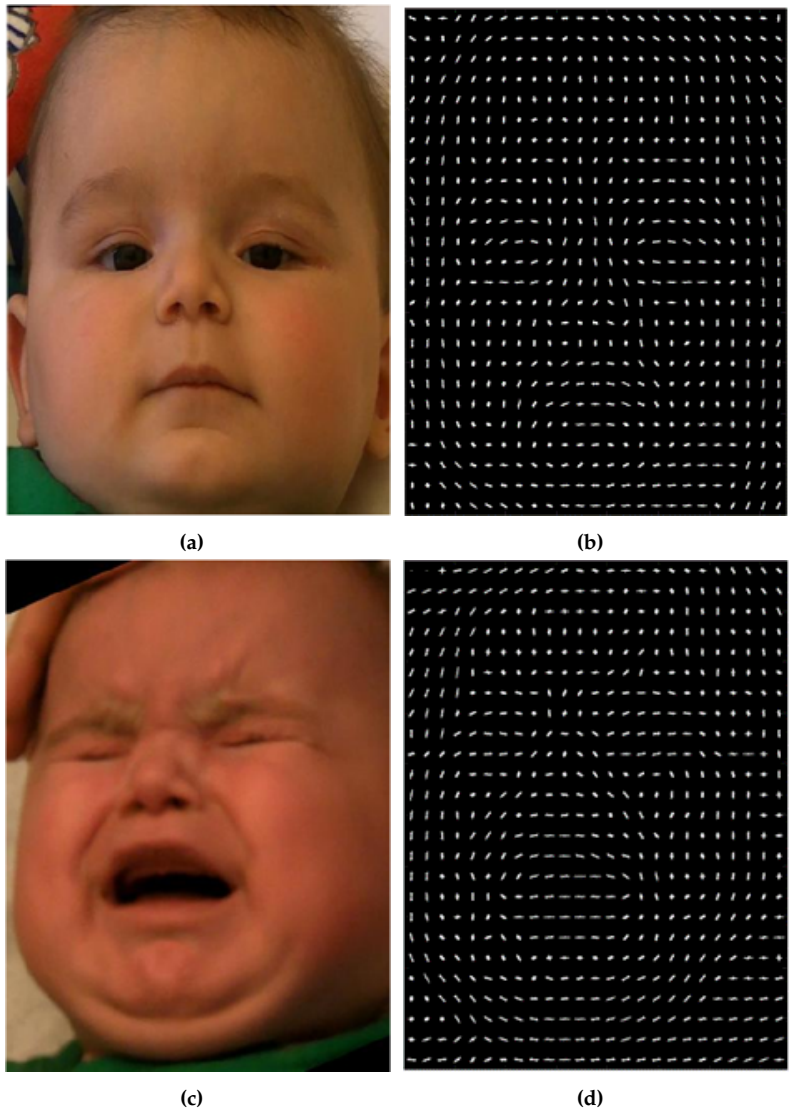


Figure 2.6: – *Examples of HOG (8 orientations) processing. Subfigures (a) and (c) are examples of comfort and discomfort face images, respectively. Subfigures (b) and (d) are the corresponding HOG of Subfigures (a) and (c).*

of the AUC. The labeling error rate is also measured and reported as an additional metric. Moreover, the quality of the classification is measured from a confusion matrix, which records correctly and incorrectly recognized faces for each class.

2.3.3 Phase 2: Subject-dependent discomfort detection

Phase 2 is proposed to incorporate infant-dependent features in addition to the independent features of Phase 1. In clinical practice, comfort moments of infants occur more often than discomfort, which means they are easier to capture. As a result, we obtain annotated comfort moments of infants and use the faces as predefined parameterized templates for expression comparison. These expressions are then compared with the input frame in terms of similarity.

To define the origin of the landmarks, all facial landmarks are further translated in such a way that Landmark 28 (the topmost point outlining nose) becomes the origin of each landmark point-set. We calculate the Euclidean distance between the coordinates of eye brow-, eye-, nose-, mouth-related facial landmarks of every template image and the corresponding coordinates of the input face image.

For each infant, we use 50 templates. The template images are randomly

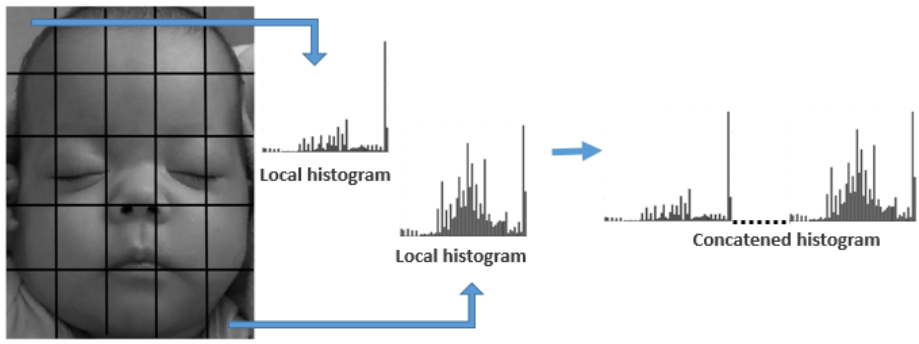


Figure 2.7: – *Processing of the infant image patches for LBP calculation. The infant face image* (500×375 *pixels) is divided into* 5×5 *sub-regions from which LBP histograms are extracted and concatenated into a single, spatially enhanced feature histogram.*

selected from the comfort face images of each infant to serve as an input for creating a reference set (see Figure 2.8 for the reference set and each input image). These landmark point-set includes in total 51 points of eye brow-, eye-, nose-, mouth-related facial landmarks, which are further used for similarity calculation. *The template* is now defined as the set of two-dimensional coordinates of the 51 facial landmarks of the face image. The Euclidean distance is then computed between the coordinates of the 51 facial landmarks of the input image and each template of the 50 face images. We pick the mean, median, maximum, minimum, 30^{th} percentile and 80^{th} percentile of the 50 obtained values of the Euclidean distance as a set of numerical features. The algorithm block diagram of the complete system is depicted in Figure 2.9. As can be observed from the large metablocks, the division into Phase 1 (drawn at the top) and Phase 2 (at the bottom) can be readily noticed. Each phase contains a feature extraction block, where Phase 1 concentrates on the facial features, while Phase 2 adds the distance calculation for template matching and determines statistical parameters of the distance. These parameters will serve as input to the SVM classification (depicted at the right).

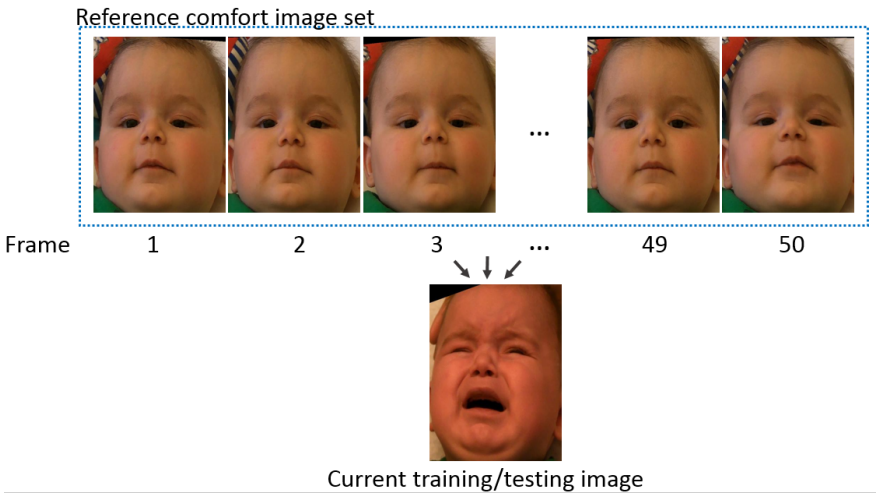


Figure 2.8: – *Visualization of the template matching. The top row shows example frames from the 50 templates representing comfort of the same infant in Phase 2: subject-dependent discomfort detection. The bottom image is an illustration of the input training/testing image.*

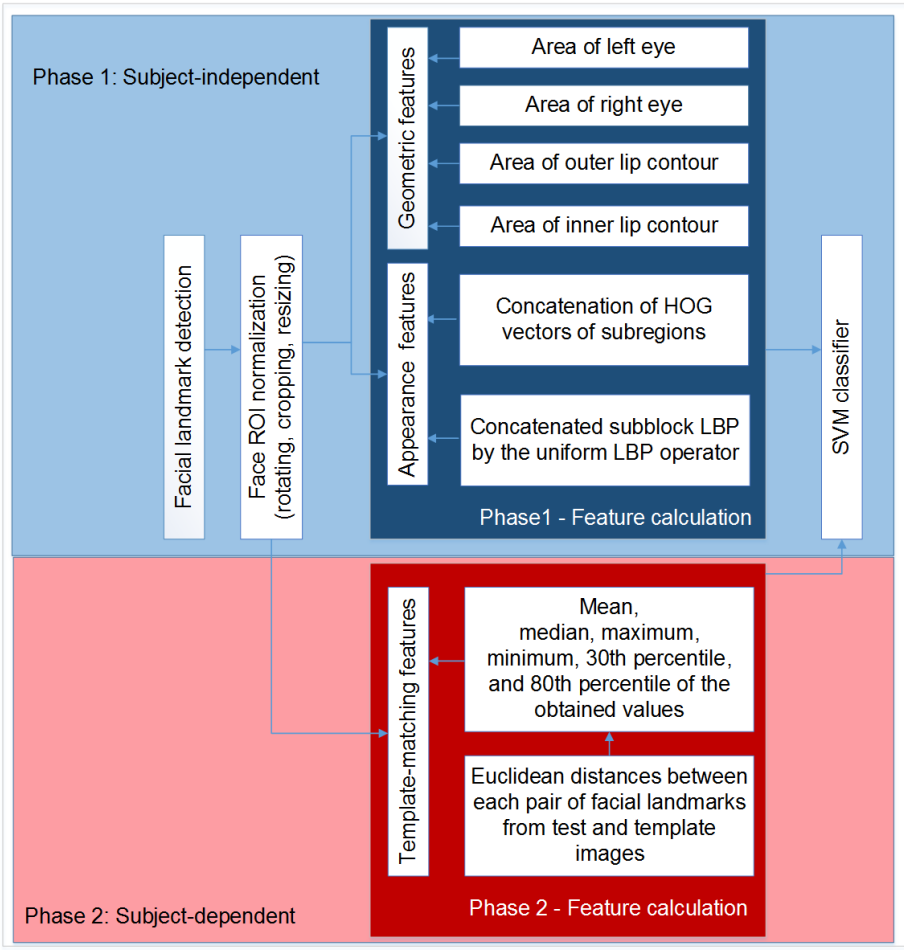


Figure 2.9: – *Block diagram of the proposed algorithm for the classification of infant comfort/discomfort.*

2.4 Experimental results

2.4.1 Materials.

The study was conducted with videos recorded at the Maxima Medical Center in Veldhoven, The Netherlands, by a handheld high-definition camera (Sanyo Xacti VPC-FH1BK). For all infants in the database, written consent was obtained from at least one of the parents. Twenty-two infants were recorded in total. Twenty infant faces were captured when they were experiencing stressful moments, including treatment moments and special occasions, like the clinical treatment of a heel prick, placing an intravenous (IV) line, venipuncture, vaccination or post-operative pain, and discomfort moments of a diaper change, feeling hungry or crying for attention. For 10 out of the 20 infants, the relaxed comfort state of resting or sleeping was also recorded. There were 2 infants only having their relaxed moments recorded. Thus, the image frames contain 1-2 emotions per subject. The number of infants regarding the recorded status of comfort/discomfort is summarized in Table 2.1. The duration of the video segments varies from less than one minute to several minutes.

Infant status	Number of videos
Comfort only	2
Discomfort only	10
Exhibiting both	10

Table 2.1: – *Dataset summarization.*

The age of the 22 recorded infants ranged from 2 days to 13 months old. Three of the infants were born premature, and under 37 weeks at the time of recording. Example of video frames for all the infants in the dataset are shown in Figure 2.10. The frames of the 3 premature infants are outlined by yellow dotted rectangles. The resolution of each video frame is 1920×1080 pixels, while the frame rate is 30 fps. The videos were recorded under uncontrolled, regular hospital lighting conditions. The labels of comfort/discomfort for each frame were annotated according to the consensus of two clinical experts.

We have extracted video segments of which infants are in supine position. Finally, a total of 16,378 frames are obtained, on which facial landmarks are detected. From all of the frames, 6,075 present comfort, and the rest 10,303 are discomfort frames.

Figure 2.10: – *Examples of frames in the database. Comfort frames are highlighted with a green box, and discomfort frames are indicated in a red box. Top 4 rows show the 10 infants with both comfort and discomfort moments recorded, where the 1^{st} and 3^{rd} rows are the discomfort frames, and the 2^{nd} and 4^{th} rows are the comfort frames. The 5^{th} and 6^{th} rows are the 10 infants with only discomfort moments recorded. The two pictures at the bottom show the infants with only comfort moments. The three premature infants are outlined by yellow-dotted rectangles.*

2.4.2 Results for Phase 1

The subject-independent method (Phase 1) is evaluated by using the 22-infant dataset. Leave-one-out cross-validation is performed for testing evaluation. Table 2.2 summarizes the results of the AUCs and labeling error rates for each feature type and for the combination of all features. When all features are

Feature category	**AUC**	**Labeling error rate**
Geometric features	0.85	0.22
HOG	0.69	0.29
LBP	0.71	0.28
All features	**0.87**	**0.15**

Table 2.2: – *Performances of various feature categories in terms of classification error rates and AUCs.*

combined, the average of the labeling error rates for all infants is 0.15. The confusion matrix based on the selected 40 best features is shown in Figure 2.11.

Figure 2.11: – *Confusion matrix when all features are combined. The numbers below the percentages are the number of frames. Among all the comfort frames, 73.0% are correctly detected. Among all the discomfort frames, 83.1% are correctly detected.*

The accuracy values for the three premature infants are 0.97, 0.61, and 0.79.

Figure 2.12 shows examples of misclassified faces. Figure 2.12 (a) depicts a false negative case, where the image is a discomfort frame. However, our system treats it as comfort because of the opened eyes and mildly opened mouth in the face. Figure 2.12 (b) portrays a false positive example. In this case, our system classifies this frame as discomfort, since the mouth size is large and the eye size is small. However, the infant is actually yawning at this moment and feeling comfortable.

(a) **(b)**

Figure 2.12: – *Examples of misclassified frames. (a) False negative frame, and (b) false positive frame.*

2.4.3 Results for Phase 2

The performance of the Phase-2 subject-dependent method is evaluated by conducting experiments on the 10-baby dataset, of which both comfort and discomfort moments are present. For this dataset, the workflow of Phase 1 is first executed and then Phase 2 afterwards, to obtain the infant comfort/discomfort classification. The experiment for Phase 1 (without subject-dependent methods) is based on the selected features with the best AUC. After the feature

selection in Phase 1, we combine the selected features of Phase 1 with template matching features, to evaluate the efficacy of Phase 2. Fig 2.13 shows the ROCs for Phase 1 (without template matching features) and Phase 2 (with template matching features), where the AUC increases from 0.89 to 0.97. Fig 2.14 plots the accuracy values per infant when separately applying subject-independent features, template matching, and the combined features. The overall accuracy of Phase 1 for the 10-infant dataset is 0.79, and for Phase 2 it is 0.95, which forms a significant increase of accuracy of approximately 20%.

The highest AUC is achieved when combining Phase-2 subject-dependent features with selected subject-independent features of Phase 1. This proves that the features are sufficiently complementary and that combining them is useful. With our proposed landmark-based template matching, the AUC is increased to 0.97, resulting in an overall accuracy of 0.95. The detection accuracies for comfort and discomfort frames are 73.0% and 83.1%, respectively. The average accuracy of the 3 premature infants in the dataset is 0.79, which highlights that our system is also interesting for serving as premature infant discomfort detection.

2.5 Discussion

Overall results: Automated classification on videos of infants helps to detect their discomfort. The proposed system combines two phases of subject-independent and subject-dependent methods, using geometric and appearance features and template matching features, respectively. The obtained AUC of 0.97 is considered promising for clinical practice, but the practical implementation aspects such as head-pose variations, occlusions by physicians and nurses, require some further discussion.

Decision-support system: The high detection accuracy of the proposed system indicates that the computer system has potential to be used as an alert system for notifying doctors and/or nurses about the status of the infants. However, the alert system serves as an advise, since the medical staff can still make a final assessment after a quick inspection of the system result.

Suitable for real-time application: For the entire dataset of 22 infants, we have extracted 7,104 features per frame. After feature selection, the AUC ranges from 0.84 to 0.87. After empirical experiments, it was found that the best feature count is 40 features, which should be extracted from each frame, in order to obtain the high AUC values as previously indicated. This amount of features leads to a feasible vector when using e.g., one number for each feature, which

is suitable for a practical real-time video-based application, given the current status of computing platforms.

Template acquisition: For the 10-baby dataset with both comfort and discomfort moments, we have shown the benefits of subject-dependent features, using landmark-based template matching. Since the reference landmarks are infant-specific and extracted from a normalized comfort status, the extracted features are very effective for discriminating the infant status. However, the disadvantage of using this type of features is that we need an additional initialization step that records the comfort status of each infant in real practice. Such an initialization recording clearly hampers usage in clinical practice.

Labeling error rate: The proposed system can reach a labeling error rate of only 0.05. In the case that a discomfort expression is misclassified as a com-

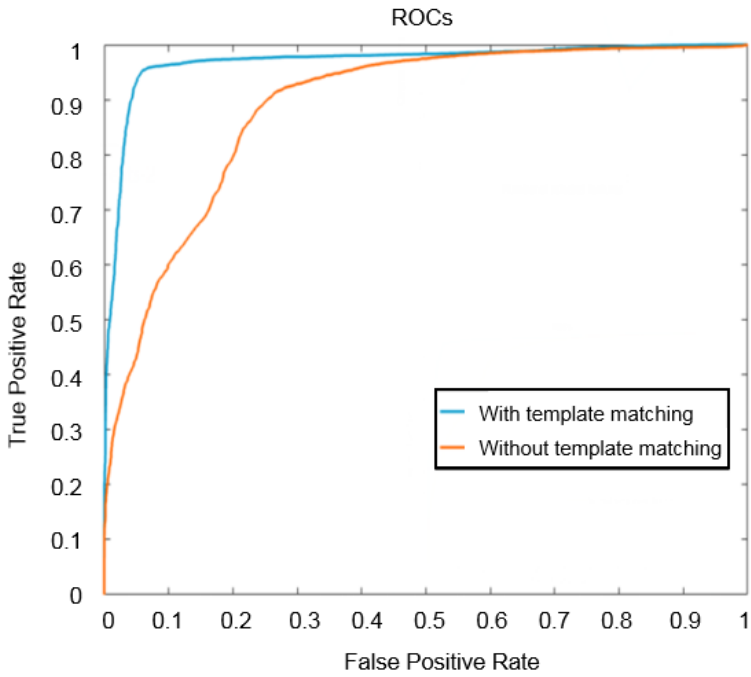


Figure 2.13: – *ROC curves for classification with and without template matching features, where the AUC is 0.97 for the method with template matching, and 0.89 without template matching.*

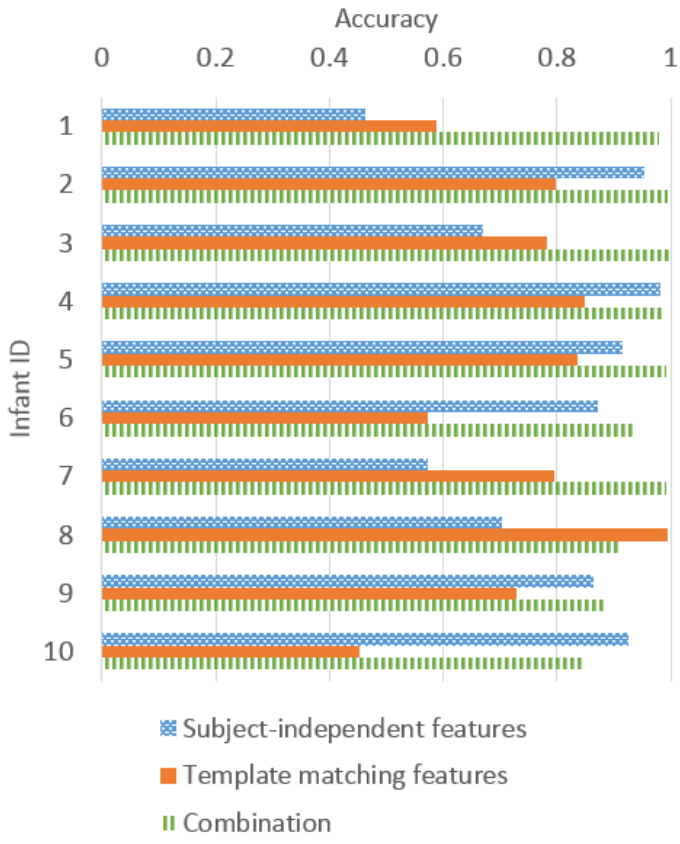


Figure 2.14: – *Obtained accuracies for each infant when applying subject-independent features, template matching features, and the combination of both. The average accuracy values of all the 10 infants and feature categories are 0.79 for subject-independent features (Phase 1), 0.74 using template matching features, and 0.95 when combining all features (Phase 2).*

fort expression by the automated system, there is often no face feature that is linked to the status. For example, the face itself may be also not in a suitable position. The proposed system makes decisions only at frame level. However, it should be noted that annotation is done by medical experts looking over a time interval to the infant, so that they annotate with the context information extracted from the temporal domain. This context information is not exploited in our proposed system. Additionally, there are border cases in the infant behavior. For example, when an infant is yawning (or aiming to) with opened mouth, our system may classify a comfort expression as a discomfort status by mistake. Such cases need further evaluation/study.

Sensitivity and specificity: The proposed automated system is selective. The area under the ROC curve is very high (0.97). From the ROC curve, it can be observed that we can keep the sensitivity of detecting discomfort status of our computer system to be almost 100%, while the specificity is 80%. It means that our system can identify about 80% comfort frames without missing any discomfort frames. The fraction of remaining frames is so small that the healthcare professionals can decide on the status of these remaining frames. However, this requires interaction of the staff members with the machine learning system. This option is still under investigation to evaluate its feasibility.

Video source: For the occurrence of face occlusion or head turning, it is not feasible to detect the face and therefore analyze the face expression when using our system. It should be noticed that the used recordings could be made under nearly ideal conditions, since we were allowed to capture the infant faces in many ways. However, in practical clinical conditions, e.g., during treatments, this may not be possible without interrupts, because interventions of the nurses and doctors cause occlusions and therefore interrupt in the comfort status tracking. This phenomenon leads to drops in the detection. In the future, we would like to add features based on body motion analysis to make our system robust, especially against facial occlusions.

2.6 Conclusions

In this chapter, we have proposed a video-based automated system that can differentiate discomfort of infants from comfort status by analyzing handcrafted facial features. The system aims at alerting pain and discomfort, with the aim to improve the long-term developmental outcomes of infants. The experimental results are promising, so that it has potential for clinical usage.

Our contribution to this field is in three ways. First, this is one of the first au-

tomated systems for facial expression recognition for infants subject to hospital care. Second, we propose to normalize the faces for effective feature computation. Third, we incorporate landmark-based template matching, involving Euclidean distance to boost our system performance. This point discriminates the specific infant properties by landmark-based template matching (subject-dependent phase).

The method has been evaluated using a dataset consisting of videos from 22 infants. Both Phase 1 (subject-independent) and Phase 2 (subject-dependent) feature extraction procedures have been evaluated for the sake of investigating the effectiveness of subject-independent/dependent features. Experimental results show an AUC of 0.87 for the subject-independent Phase 1, and 0.97 for the subject-dependent Phase 2.

One limitation of our study is the small size of the dataset. Collecting videos from infants at the pediatrics department in the hospital is rather restricted because of the ethical and legal issues. In the meantime, not many infants within our target group are available for recording. Moreover, some of the captured recordings are not usable due to interruptions from nurses or parents, face occlusion, etc. These aspects effectively limit the dataset even further. This explains why more data is involved for the studies in the upcoming chapters.

Another limitation of this work is that features used are computed from a static frame. The temporal dynamics of facial expressions are not taken into account. In future work, we will analyze facial expression changes over time and also consider dynamic features based on body motion analysis. Furthermore, we may add features extracted from vital signs, such as heart rate and respiration rate, and evaluate those in combination with our visual features for improved robustness and reliability. To further improve the accuracy of our system, we will explore different settings (parameters) of HOG and LBP features in the future. Regarding geometric features, it is possible to further exploit the rich information from landmarks, such as the distance between the eye and eyebrow, or the distance between upper and lower lips, etc.

The aforementioned limitations in data is a recurring theme for further research as well. Because of the success of deep learning, this thesis will address this development as well. Deep learning methods have shown superior performance recently in many computer vision applications. The continuous increase of computation power enables experimentation with deeper architectures of networks, resulting in broader perception of deep learning. In the next chapter, we will investigate deep learning-based methods for our problem, albeit with limited data availability, so that we start with retraining/refining concepts.

Chapter 3

2D CNN-based facial expression analysis

3.1 Introduction

The previous chapter has discussed the systematic design for the automated detection of infant discomfort based on conventional handcrafted features followed by an SVM classifier. As described and explored in Chapter 2, facial expression appears to be one of the behavior indicators of discomfort or pain. As shown in Figure 3.1, the PFP is an intuitive and universal facial expression associated with pain [126]. Automated detection of discomfort (or pain) in videos by analyzing the facial expression of infants, is a potential solution for continuous monitoring. At the end of Chapter 2, future work on the deep learning direction is suggested. For this reason, this chapter concentrates on exploiting the performance of deep learning-based methods for the same task of discomfort detection based on facial expression analysis.

Convolutional neural networks (CNNs) were introduced by Lecun *et al.* [76]. Each layer in the network can react to different levels of data and when the layers are stacked together, a new representation is created. Recently, deep learning has achieved the state-of-the-art recognition accuracy and has outperformed the results of existing conventional methods for many computer vision systems. CNNs have been successfully applied to various computer vision problems such as image classification, object detection, pattern recognition, etc. CNNs have also been increasingly leveraged to learn discriminative rep-

resentations for automated facial expression recognition, which showed good results.

In this chapter, we propose an automated discomfort detection method for infants by analyzing their facial expressions with deep CNN algorithms. In more detail, the following challenges and requirements of such a discomfort detection system should be addressed.

- *Performance comparable with conventional methods:* It should be investigated whether it is possible to apply a data-driven approach, which potentially could outperform conventional handcrafted features combined with SVM-based methods.
- *Small dataset learning:* It is difficult to collect facial expression video data of infants due to privacy limitations and ethical issues. We have managed to collect a dataset of 55 videos from 24 infants. However, to train a deep CNN algorithm, a large dataset is required for a good generalization ability of the network. In contrast, a small dataset easily leads to overfitting. The proposed system should handle this problem of a small dataset for learning.
- *Robustness under variable imaging conditions:* The system should be robust under different face poses and environmental parameters (e.g., various lighting conditions). A validation test using different face sizes and/or different face rotation angles should be carried out.

This chapter aims to develop an automated CNN-based system to classify the infant comfort/discomfort status, despite the use of a small dataset. This challenge is addressed by pre-learning on a large generic dataset, after which tuning of the network is performed with the small specific dataset. The robustness aspect is then addressed during the validation.

The chapter is divided as follows. In Section 3.2, related work on pain/discomfort detection and CNN-based classification is described. Section 3.3 elaborates the proposed method, and Section 3.4 explains the experimental results. Finally, Section 3.5 discusses the results and Section 3.6 draws conclusions.

3.2 Related work

3.2.1 Conventional facial expression recognition

The amount of studies on automated pain or discomfort detection for infants in videos is quite limited [134] [37] [146] [176]. Sikka *et al.* [134] proposed a

Facial Action Coding System (FACS) and associated methods to describe the facial expressions of children with pain. Fotiadou *et al.* [37] presented a discomfort detection system utilizing the active appearance model (AAM) and a Support Vector Machines (SVM) classifier. The system achieved an AUC of 0.98 by evaluating in 15 videos from 8 infants. However, for each baby, various landmarks need to be placed manually for initializing the AAM mesh. Sun *et al.* [146] designed appearance and geometric features to describe infant faces for discomfort detection, and achieved an AUC of 0.87. However, this system was based on conventional techniques with handcrafted features and an SVM classifier.

In the past, significant attention was paid to facial expression recognition of adults [129] [122]. Kotsia and Pitas [73] proposed two methods for facial expression recognition: (1) estimating geometrical displacement of certain selected Candide grid nodes, which was followed by a multi-class SVM system, and (2) an approach based on Facial Action Units (FAUs). The recognition accuracy of 99.7% and 95.1% was achieved when using the multi-class SVM and FAU-based method, respectively. Shan *et al.* [131] presented a comprehensive empirical study of facial expression recognition, based on Local Binary Patterns (LBPs) and illustrated that LBP features perform stably and robustly over a useful range of low-resolution facial images.

Different machine learning methods, such as Adaboost [128], have been fur-

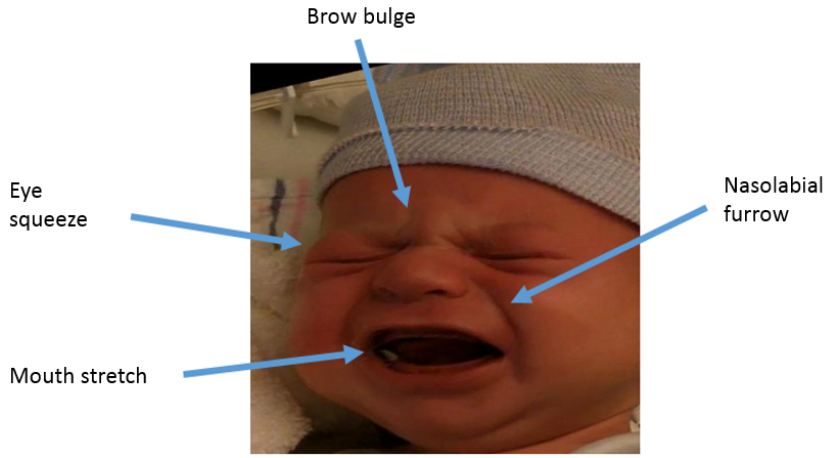


Figure 3.1: – *Primal face of pain (PFP).*

ther exploited for facial expression classification. Neshov and Manolova [97] used a supervised descent method and scale-invariant feature transform, which yielded a high recognition rate (more than 95.7%). A hierarchical unsupervised feature learning approach was employed by Kharghanian *et al.* [66] to extract the features detecting pain from facial images, based on a convolutional deep belief network. The AUC of the Receiver Operating Characteristic (ROC) was near 95%.

Lucey *et al.* [89] showed that the AAM-based system can overcome the facial deformation and head motion and yields significant improvements in both the FAU and pain detection. Ashraf *et al.* [7] explored various face representations derived from AAMs for detecting pain from faces, and demonstrated that decoupling a face into separate non-rigid shape and appearance components offered a significant performance improvement.

Hammal *et al.* [46] applied a set of Log-Normal filters, consisting of 7 frequencies and 15 orientations to extract 9,216 features for pain estimation and the statistical analysis – using the F_1 metric (the harmonic mean of the precision and recall)– for each level of pain intensity, ranging from 91% to 96%. Littlewort *et al.* [82] proposed an automated facial expression recognition system to differentiate real pain from fake pain. A 20-channel output of facial action detectors from the FACS was passed to a classification stage to determine the labels. An accuracy of 88% was obtained for the subject-independent discrimination of real versus fake pain. Hazelhoff *et al.* [49] developed a prototype of an automated video surveillance system by analyzing facial expressions. The method first localized eye, eyebrow and mouth regions, which was followed by employing a hierarchical classifier to discriminate between different behavioral states of sleep, awake and cry. With this system, an accuracy of 95% was achieved. A similar method [47] was applied on a dataset of Neonatal Intensive Care Unit (NICU), which resulted in an accuracy of 88%. Zhao *et al.* [188] proposed a novel Set-to-Set (S2S) distance measure, to calculate the similarity between two sets with the aim to improve the recognition accuracy for faces with real-world challenges, as compared to traditional feature-average pooling and score-average pooling. However, this method still depends on the effectiveness of the extracted features. Ding *et al.* [30] proposed a deep confidence network (DECODE) for robust training. The method adopted an effective evaluation module based on a probabilistic confidence measure. This module assigned small training weights to suspicious samples, to suppress the influence of noisy data. The weighted training data was also used to update the weights after each iteration. Evaluation of this method was carried out on several datasets, where the effectiveness of DECODE was demonstrated.

3.2.2 CNN-based methods

Recently, CNN has become a powerful tool for automated two- and three-dimensional image classification. Raghuvanshi *et al.* [120] classified images of human faces into discrete emotion categories using CNNs and experimented with different architectures and methods, such as fractional max-pooling and fine-tuning, ultimately achieving an accuracy of 48% in a seven-group classification task. Lopes [85] proposed a simple solution for facial expression recognition that uses a combination of a CNN and a specific image pre-processing step for extracting only expression-specific features from a face image. The proposed method achieved competitive results when compared with other facial expression recognition methods.

Wang *et al.* [168] proposed a network that fine-tuned a state-of-the-art CNN face verification network, using a regularized regression loss and additional data with expression labels, which achieved state-of-the-art performance. Shin *et al.* [53] used transfer learning to address the problem of CNN training on limited labeled medical data and domain knowledge. They learned a codebook from 15 million images from ImageNet in an unsupervised learning fashion, which encodes the fundamental features of those images without expert knowledge. A weighting vector for each image in their experiment was obtained from the codebook and was supplied into SVM classifiers for supervised learning. They concluded that it is promising to use transfer representation learning for analyzing medical data. Christodoulidis *et al.* [23] employed pre-trained networks on six public texture datasets and further fine-tuned the network architecture on lung-tissue data. The resulting conversational features were fused with the original knowledge, which was fed back to the network in compressed form. Their results showed that the proposed method improved the performance by 2 % in terms of accuracy, compared to the same network without using transfer learning. In [157], Tajbakhsh *et al.* conducted experiments for addressing the research question whether pre-trained deep CNNs with sufficient fine-tuning eliminate the need for training a deep CNN from scratch. The authors concluded that deeply fine-tuned CNNs are useful for analyzing medical images and they performed as well as fully trained CNNs. Given the limited training data, the fine-tuned CNNs even outperformed the fully trained networks. The experience gained from one subject can be transferred to other subjects. For CNNs, the parameters trained on one dataset can be reused by a new dataset. Usually, transfer learning is used for training a base network and then its first n layers are copied to the first n layers of a new network. The remaining layers

of the new network are then initialized randomly and trained according to the new task [178]. Elements of this approach have been adopted in modified form into our method described below.

Summarizing the related work in the context of this chapter, we concentrate on CNN-based methods using pre-trained networks. In order to address the problem of limited available data for learning, we utilize transfer learning by implementing several training steps based on various sources of image data.

3.3 Methods for CNNs using fine-tuning

The proposed method encompasses both a special training procedure and the fine-tuning of the CNN, which are discussed subsequently. The entire workflow for training the CNN is depicted in Figure 3.2. More specifically, we adopt a pre-trained model, which is followed by training the networks, using a public face-image dataset [89] with labeled facial expressions for pain assessment (first fine-tuning). The networks are then further fine-tuned with our dataset of infants (second fine-tuning). The method description is divided into three subsections, i.e. pre-processing, CNN model, and transfer learning.

3.3.1 Pre-processing

To be able to classify facial expressions, we first need to locate the face area within video images. The face Region of Interest (ROI) is generated based on the workflow of face detection and normalization. The selected ROI is passed on to the next step as input for the discomfort/comfort classification. Given an input video frame, 68 facial landmarks are first localized using the Dlib face landmark detector [67], which is an implementation of a detection method by

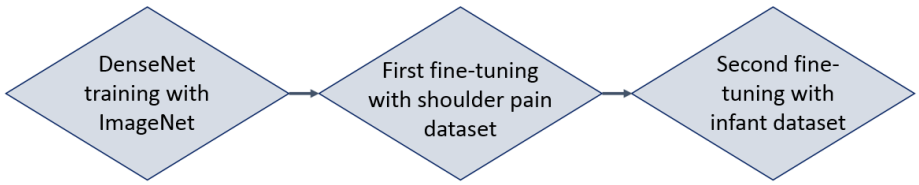


Figure 3.2: – *Workflow of the proposed discomfort detection method using CNNs with fine-tuning steps.*

Kazemi *et al.* [64]. The 68 landmarks include points on the face such as the corners of the eyes, mouth, along with the eyebrows, and the boundary of the face. Once the 68 landmarks are identified, the middle point between two inner eye-corner points is used as a reference point to rotate the image. The image is then rotated to the position where the line connecting the two eye-corner points is horizontal. Thus, this step minimizes the in-plane rotation variance of the face. We select Landmark 1 as the leftmost point, Landmark 17 as the rightmost, and Landmark 9 as the bottom-most points to define the left, right, and bottom boundaries of the face ROI. For the top boundary, we choose the horizontal line that has a vertical distance from the inner eye-corner center that is identical to the distance to Landmark 9 (thereby making this eye-corner center point of the middle of the face region). A margin of 20 pixels is added to all boundaries to cover the whole face and avoid any loss of facial information. Finally, all face images are cropped and then resized to 224×224 pixels using bilinear interpolation, in order to adapt to the required input image size for CNNs. Figure 3.3 exemplifies an original face image with a corresponding normalized face ROI and the detected 68 facial landmarks.

3.3.2 CNN model

Inspired by the performance of CNN models [51, 56, 74, 137], this section concentrates on exploring the use of a CNN model as a solution for the posed problem converting the landmark description into a facial detection. The CNN usually contains several pairs of convolutional layers and a max-pooling layer for non-linear downsampling, followed by fully connected layers. Compared to the conventional facial expression systems based on handcrafted features [146], a CNN model has the advantage of self-learning. The parameters in a CNN are automatically learned from the data and researchers do not need to design complicated features as input.

In this work, the adopted base model is DenseNet [55], which is a network architecture where each layer is directly connected to every other layer in a feedforward fashion within each dense block. For each layer, the feature maps of all preceding layers are treated as separate inputs, whereas its own feature maps are passed on as inputs to all subsequent layers. DenseNet has the advantage of alleviating the vanishing-gradient problem [55], strengthening feature propagation, and reducing the number of parameters. In the referred work, this connectivity pattern yields state-of-the-art classification performance on CIFAR10/100 (with or without data augmentation) and SVHN [55]. On the large-scale ILSVRC 2012 (ImageNet) dataset, DenseNet achieves a similar per-

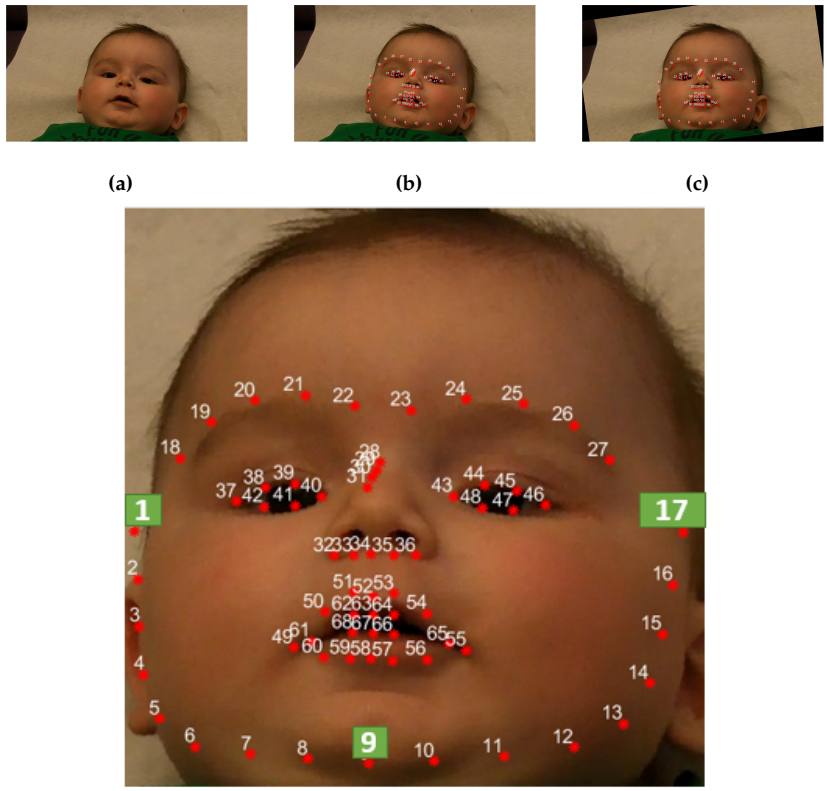


Figure 3.3: – *Example of face ROI detection and normalization. (a) Sample original face image, (b) the original face image with identified 68 facial landmarks, (c) in-plane head rotation corrected, and (d) the final normalized face ROI with the corresponding 68 landmarks, of which Landmarks 1, 9 and 17 are highlighted.*

formance as ResNet, while using less than half of the number of parameters and roughly half of the number of floating-point operations (Flops). Figure 3.4 shows the CNN architecture of the proposed work. The number 121 corresponds to the number of layers with trainable weights (excluded batch normalization layer), i.e. convolutional layers and fully connected layers. The additional 5 layers include the initial 7×7 convolutional layer, 3 transitional layers, and a fully connected layer. Between network blocks, the processing steps are convolution and pooling, which consist of a batch normalization layer and an 1×1 convolutional layer followed by a 2×2 average pooling layer. The Dense block is identical to the one introduced in [55], except for the last output layer, which is modified according to the binary discomfort/comfort classification task because the number of classes is different. This specific CNN model is chosen because of its computation efficiency and to achieve the goal of a real-time, or semi-real-time application. Compared to other existing networks, such as ResNets [50] and Inception networks [156], DenseNet concatenates feature maps learned by different layers, which increases variation in the input of subsequent layers and improves efficiency. Therefore, the simpler structure with only 4 dense blocks is sufficiently efficient for the learning procedure (see Figure 3.4 for the network structure).

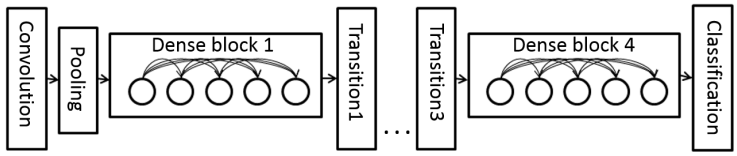


Figure 3.4: – *Structure of the DenseNet 121 network used in the proposed system.*

3.3.3 Transfer learning

In the proposed work, a significant challenge is that a very limited number of infant videos are available. This is a common problem in the medical field, due to privacy issues and the boundaries on expenses for medical equipment. One effective solution is to use transfer learning [105] [158], to address the problem of limited availability of labeled data. For the purpose of transfer learning, we can preserve all pre-trained layers prior to the last output layer and connect these layers to a final layer for the considered classification problem. To train the networks for the new dataset, we desire that the parameters from the

Figure 3.5: – *Facial expressions of six pain levels from a woman in the Shoulder-Pain dataset. For subfigures (a) through (f), the pain-intensity score increases from 0 to 5.*

fully connected layers of the network are updated or optimized. The alternative choice is to fine-tune more layers, or even the whole set of pre-trained network layers. It is also possible to keep the first convolutional layer fixed, since this layer is often used for edge extraction [127], which is common for generic image processing problems. Therefore, as discussed earlier, since not all parameters are re-trained or trained from scratch, transfer learning is beneficial to problems with a small labeled dataset. In this work, to fully exploit the learning power of CNNs, we start with a pre-trained network, but we have chosen to fine-tune *all* parameters of the pre-trained DenseNet model. First, we obtain a DenseNet model trained on ImageNet data [55] as the pre-trained network. The first fine-tuning is based on a public dataset, which is followed by a second-tuning step based on our own data of infants. The ImageNet dataset is also first resampled to 256×256 pixels using bilinear interpolation and then the patch of 224×224 in the center is cropped.

First fine-tuning step Considering the limited number of samples in our dataset, instead of directly fine-tuning the DenseNet model using our own data, we first fine-tune the DenseNet model on a public dataset, called the Shoulder-Pain dataset [89]. This dataset contains 200 videos of 25 sub-

jects (48,398 frames in total) and is widely used for benchmarking the pain-intensity estimation. For each frame, discrete pain intensities (0-15) according to Prkachin and Solomon [113] are provided by the database creators. Similar to previous work [168] [192] [123] [195], we quantify the original pain intensities within the range of [0; 15] to be in the range of [0; 5] for the purpose of data balancing. The pain intensities are discretized into 6 pain levels as follows: 0 (none), 1 (mild), 2 (discomforting), 3 (distressing), 4-5 (intense), and 6-15 (excruciating). The data balancing is performed in order to avoid overfitting of the test methods on the majority classes. Figure 3.5 shows 6 levels of the facial expression (status) from a woman in the Shoulder-Pain dataset. In the pre-trained DenseNet 121 model, we replace the original 1000 output nodes by 6 output nodes to represent the 6 classes. The data are randomly split into a training dataset (70 %) and validation dataset (30 %). The validation dataset is used to select the best classifier, where the loss function achieves the minimum. To obtain the final label or class of each sample, we assign the label to the corresponding node in the last layer that gives the highest likelihood value. For training, the number of epochs (m) is limited to 50. Furthermore, in order to augment the number of training samples, images are randomly flipped horizontally, rotated within -10 degrees to $+10$ degrees, and translated within a distance of 20 pixels. Finally, we use the Adam algorithm [68] to optimize our loss function.

Second fine-tuning step After the first fine-tuning step, a reasonably trained model for facial expression recognition is obtained. In this step, continuing with the previous model, we replace the 6 output nodes with 2 nodes, in order to match the classes to our problem statement, i.e. automated detection of discomfort and comfort. The parameters for the networks, training, and data augmentation are the same as with the first fine-tuning step (see Fig. 3.6 for the data examples, and Fig. 3.7 shows the corresponding pre-processed images).

To obtain an unbiased evaluation of the classification performance, a twofold cross-validation is employed. Specifically, the input dataset is randomly divided into two equal parts at patient level, where one part is left out for testing, and the other part is split again for training (70 %) and validation (30 %), to avoid bias. All the parameters are updated during validation. The classifier with the lowest loss based on the validation set is chosen as the best classifier and is used for testing. Such a procedure is repeated two times with a different dataset part used for testing. We have pooled and evaluated the results from both parts to obtain various measurements on performance.

3.4 Experimental results

This section contains a description of the following experiments. First, we evaluate the model using our infant data directly without any pre-training in terms of accuracy and confusion matrix. Second, the first fine-tuning step is illuminated using the Shoulder-pain dataset by testing on the 6-class pain-intensity levels. The accuracy and error metrics are provided. Third, the final results after the second fine-tuning step are presented in the form of accuracy, confusion matrix and ROC curves.

3.4.1 Captured video materials

The study was conducted with videos recorded at the Máxima Medical Center (MMC) in Veldhoven, The Netherlands, by a handheld high-definition camera (Xacti VPC-FH1BK). For all infants in the database, written consent was obtained from at least one of the parents. Data from 24 infants were collected in total. The faces were recorded when they were experiencing stressful moments including clinical treatment of heel prick, placing an intravenous (IV) line, venipuncture, vaccination, post-operative pain, and discomfort moments of the diaper change, feeling hungry or crying for attention. For 10 out of the 24 infants, the relaxed comfort state of resting or sleeping was also recorded. For 4 infants, only their comfort moments were recorded. Thus, the image frames contain 1 to 2 emotions per subject. The number of infants regarding the recorded status of comfort/discomfort is summarized in Table 3.1. The duration of the video segments varies from less than one minute to several minutes. The age of the 24 recorded infants ranges from 2 days up to 13 months

Infant status	**No. of videos**	**Infant status**	**No. of frames**
Comfort only	4	Comfort	6,534
Discomfort only	10	Discomfort	10,303
Exhibiting both	10	Total	16,837

Table 3.1: – *Dataset summarization.*

old. Three of the infants were born premature, and under 37 weeks at the time of recording. Examples of video frames in the dataset are shown in Figure 3.6. The resolution of each video frame is 1920×1080 pixels, and the frame rate is 30 frames per second (fps). The videos were recorded under uncontrolled light-

Figure 3.6: – *Examples of frames in the database of this study. Comfort frames are highlighted by green boxes, while discomfort cases are indicated by red. The top four rows show the ten infants having both comfort and discomfort moments recorded. The 2^{nd} and 4^{th} rows are the comfort frames. The 1^{st} and 3^{rd} rows are for discomfort frames. The ten infants with only discomfort moments recorded are shown in the 5^{th} and 6^{th} rows. The four pictures in the bottom row are the infants with only comfort moments recorded. Three premature infants are outlined using yellow dotted rectangles.*

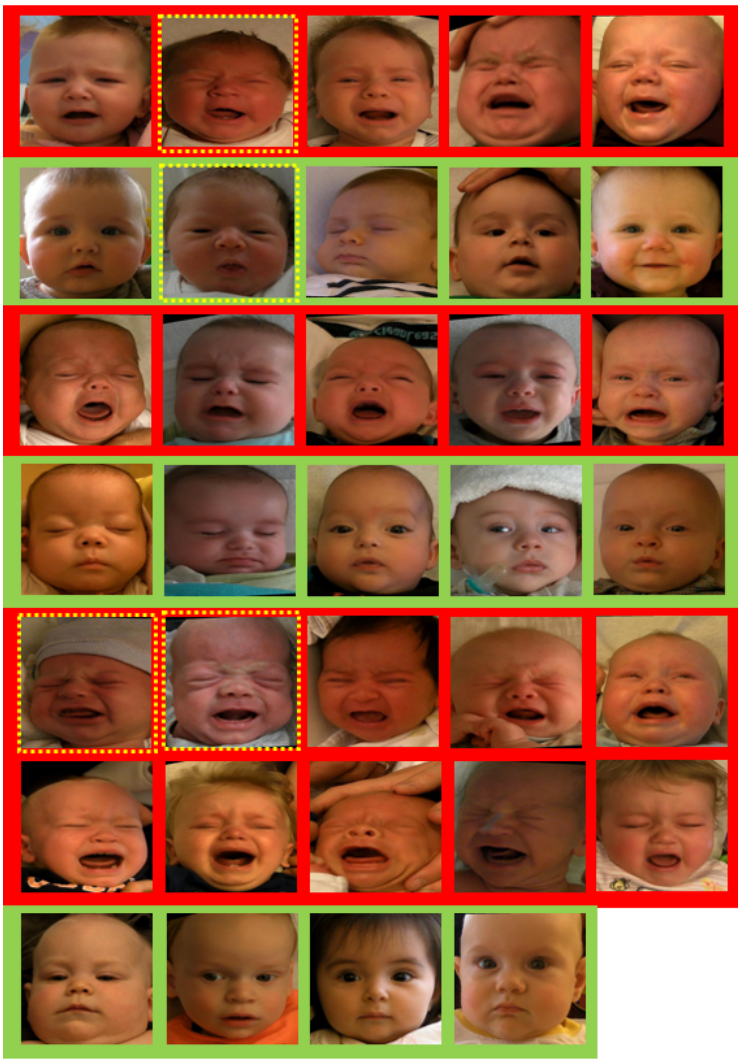


Figure 3.7: – *Pre-processed (pre-network) infant images of the frames from the previous figure and shown in the order corresponding to Figure 3.6. Comfort frames are highlighted by green boxes, while discomfort cases are indicated by red. Three premature infants are outlined using yellow dotted rectangles.*

ing conditions and can be characterized by hospital office lighting. The labels of comfort/discomfort for each frame are annotated according to the consensus of 2 clinical experts. We have extracted video segments where the infants are in supine position. Finally, a total of 16,837 frames are obtained, on which facial landmarks are detected. From all of the frames, 6,534 present comfort, and the remaining 10,303 are discomfort frames. There are more discomfort samples in our dataset than comfort samples, since data collection in the hospital has focused on recording the discomfort moments of the infants.

3.4.2 Results without any pre-training

To show the effectiveness of pre-training with transfer learning, we have first performed the experiments by directly using our data to train a DenseNet classifier from scratch. The data augmentation procedure for the training data is the same as for training from scratch, the first and the second fine-tuning step. Results without any pre-training are directly trained from the model from scratch without using any ImageNet data. A weighted loss function is used according to the size of each class in the training data, in order to account for the nature of data imbalance. Given training samples from the entire 24 infants, a twofold cross-validation is employed for evaluation. The twofold cross-validation is the same as the one described in Section 3.3.3 "Transfer learning" for the second fine-tuning step. Figure 3.8 shows the normalized confusion matrix, based on the cross-validation without any pre-training. The obtained accuracy of all validated images is 52.4 % and the accuracy values for the two classes of comfort and discomfort are 87 % and 30 %, respectively. Since the main focus in our application is to detect discomfort moments of infants in time, the low detection rate for discomfort moments is not acceptable.

3.4.3 Results from the first fine-tuning step

For the first fine-tuning step, the Shoulder-Pain dataset is used. The Shoulder-Pain dataset has been randomly split into a training dataset (70 %) and validation dataset (30 %). The classification accuracies are calculated based on the performance on the validation dataset, which shows that the overall accuracy is 85 % and the accuracy values for the six classes of facial expressions are 86 %, 81 %, 79 %, 78 %, 86 %, and 98 %, respectively. We have compared the performance of our model with existing methods [191] [194] on the Shoulder-Pain dataset, by calculating the Mean Absolute Error (MAE) as deviation from the ground-truth labels, Mean Squared Error (MSE) and the Pearson Correlation

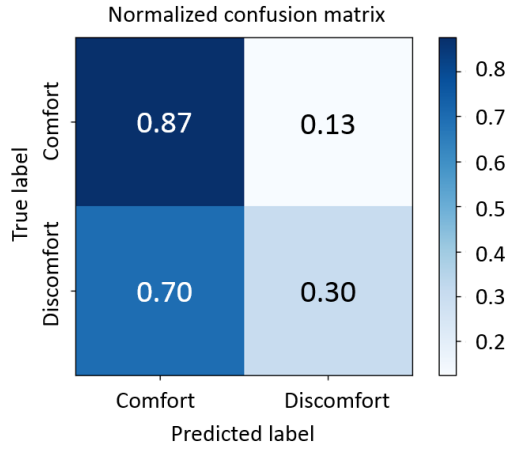


Figure 3.8: – *Normalized confusion matrix of the model without any pre-training.*

Coefficient (PCC) for the 6-level pain-intensity classification (See Table 3.2). Comparing to the two existing methods, the proposed method performs best by achieving the lowest MAE of 0.451 and highest PCC of 0.643.

Method	MAE	MSE	PCC
Proposed transfer-learning model	0.451	0.950	0.643
Ordinal information based regression [191]	1.025	N/A	0.600
Recurrent convolutional neural network regression [194]	0.810	N/A	0.601

Table 3.2: – *Performance comparison on Shoulder-Pain dataset.*

3.4.4 Results from the second fine-tuning step

Figure 3.9 shows the normalized confusion matrix of the method when applying only the second fine-tuning step (omitting the first fine-tuning) on our infant data, based on cross-validation. The accuracy of all validation images is 81 % and the accuracy of the two classes of infant status are 92 %

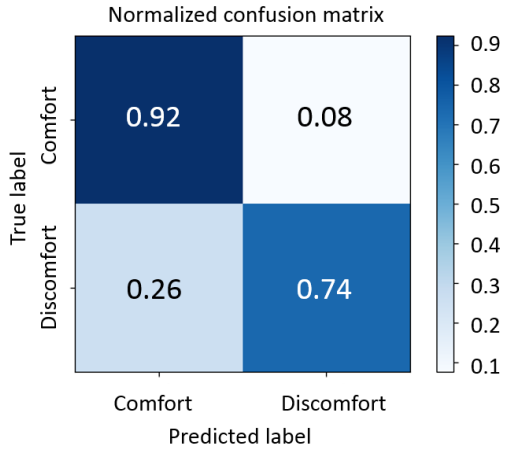


Figure 3.9: – *Normalized confusion matrix of the method, with DenseNet pre-trained on ImageNet data and using the second fine-tuning step on our data (i.e., without the first fine-tuning step).*

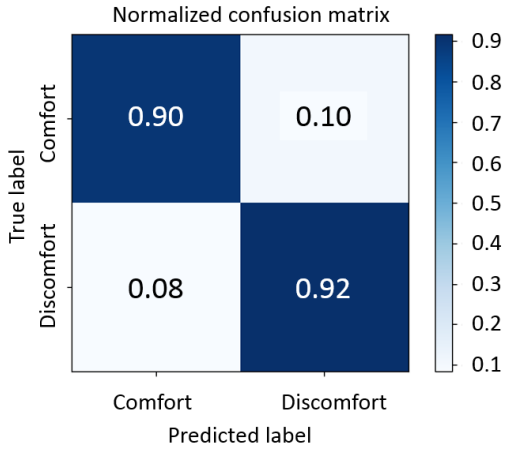


Figure 3.10: – *Normalized confusion matrix of the proposed method with two fine-tuning steps on our infant data.*

and 74 %. For comparison, Figure 3.10 portrays the normalized confusion matrix of the method including the two fine-tuning steps on our data based on cross-validation. It can be readily observed that the results are significantly improved by applying both refining steps. The accuracy of all validation images is 91 % and the accuracy of the two classes of infant status are 90 % and 92 %, respectively. The average accuracy of the three premature infants in the dataset is 85 %, which shows that our system is also interesting to be considered for premature infant discomfort detection.

We have also computed the ROC curve on our infant data without and with the first fine-tuning step (see Figure 3.12). The area under the ROC curve (AUC) increases from 0.93 to 0.96. From the ROC curve, it can be observed that 78 % of the normal status can be safely eliminated by our system, while the sensitivity is preserved at a very high level. Figure 3.11 shows examples

(a) **(b)**

Figure 3.11: – *Facial examples of discomfort expressions being misclassified. (a) Discomfort status is misclassified as comfort. (b) Comfort case is misclassified as discomfort.*

of misclassified cases. Figure 3.11 (a) indicates a discomfort case that is misclassified as a normal case by the automated system. In this case, there is no significant image feature that is linked to the status. The face itself is also not captured in the frontal view. However, it should be noted that the images are annotated within a video of a time period, where the annotator has temporal information about the infant status. Figure 3.11 (b) indicates a comfort case that is misclassified as discomfort status by our automated system. In this case, the pacifier in front of the face probably confuses the proposed network. Table 3.3 summarizes different performances including the classification accuracy and

Training method	ACC on AC	ACC on CC	ACC on DC	AUC	95% CI of AUC
Without any pre-training	52%	87%	30%	0.767	0.759-0.774
Without the 1st fine-tuning	81%	92%	74%	0.934	0.930-0.939
With all steps included	**91%**	**90%**	**92%**	**0.960**	**0.958-0.962**
Handcrafted features + SVM [146]	79%	73%	83%	0.874	0.869-0.879

Table 3.3: – *Measured classification performance. The classification accuracy (ACC) of all cases (AC), classification accuracy of comfort cases (CC), classification accuracy of discomfort cases (DC) and the AUCs with corresponding 95 % confidence intervals (CIs) of different training schemes/methods are summarized.*

AUCs of training from scratch, training without our strategic first fine-tuning step, and training with all steps included. The table also shows the result from a conventional approach [146], which is based on handcrafted features combining geometric features and appearance features with an SVM classifier. For the AUC values, the bootstrapping (resampling) approach [31] (1000 bootstrap samples) was used to calculate the 95 % confidence intervals (CIs).

Figure 3.12 depicts different ROC curves without any pre-training, without the first fine-tuning step and with all the pre-training and fine-tuning steps included. The AUC values are 0.77, 0.93 and 0.96, respectively. Figure 3.13 shows the loss on the validation data of the three training schemes. The proposed method quickly reduces the loss during training. It can also be noticed that the losses from all three settings are not decreasing after the first 30 epochs, which means that 50 epochs are sufficient for training in this case.

Deep learning CNNs are typically referred to as a black box for classification, while it is difficult to track which features are important for a specific classification task. To understand where deep learning concentrates upon, the Class Activation Map (CAM) [193] can be computed by a weighted sum of the feature maps of the last convolutional layer. CAM can be used to indicate whether the deep learning networks focus on a critical facial area or help in understanding which regions in the face are relevant to the discomfort detection problem. It can be observed that the highlighted regions in activation maps

Original test set	Zoom factor from 0.8 to 1.2	Rotation from -45 degrees to +45	Contrast from 0.8 to 1.2
AUC	0.96	0.96	0.95
ACC	0.91	0.90	0.91

Table 3.4: – *Performance of the proposed proposed method on different imaging factors.*

are quite often the mouth and eye areas of the face, as shown in Figure 3.14.

We have investigated the performance when randomly zooming in and out from 0.8 up to 1.2, rotating the face from -45 to +45 degrees and changing the contrast from 0.8 to 1.2. The results are shown in Table 3.4, which demonstrate that our system is robust to different video conditions.

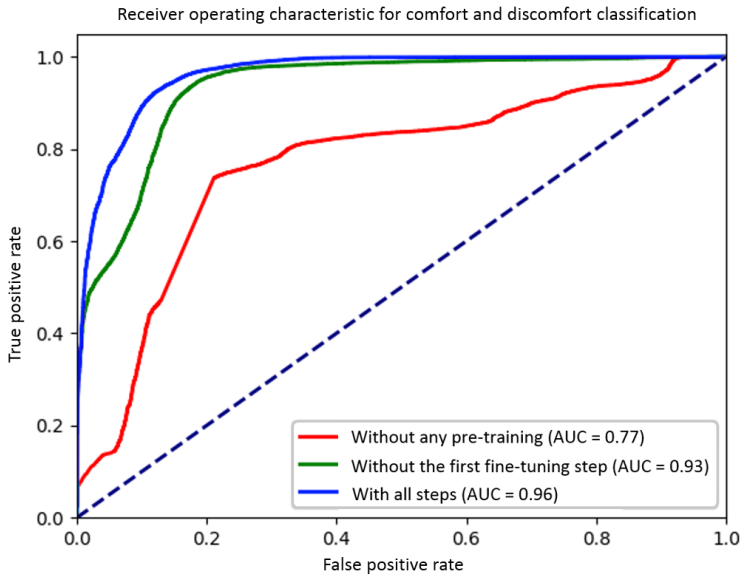


Figure 3.12: – *ROC curves of the proposed method without any pre-training, without the first fine-tuning step, and training with all steps included.*

3.4.5 Video segment classification

We have also included an experiment for segment-based video analysis (5 seconds per segment, i.e. 150 frames) and classification by fusing temporal information. Then, we have computed the mean, maximum and minimum of likelihoods of all frames and obtained the following AUC values: 0.984 (95 % CI: 0.967-0.993), 0.980 (95 % CI: 0.958-0.992) and 0.966 (95 % CI: 0.939-0.982), respectively.

3.5 Discussion

The proposed method has been applied for infant-independent discomfort detection on our infant dataset. Compared to the existing method using handcrafted features [146], the AUC of the proposed method also increases 10 % on the same dataset.

Benefit of pre-training: The advantages of using an intermediate strategic

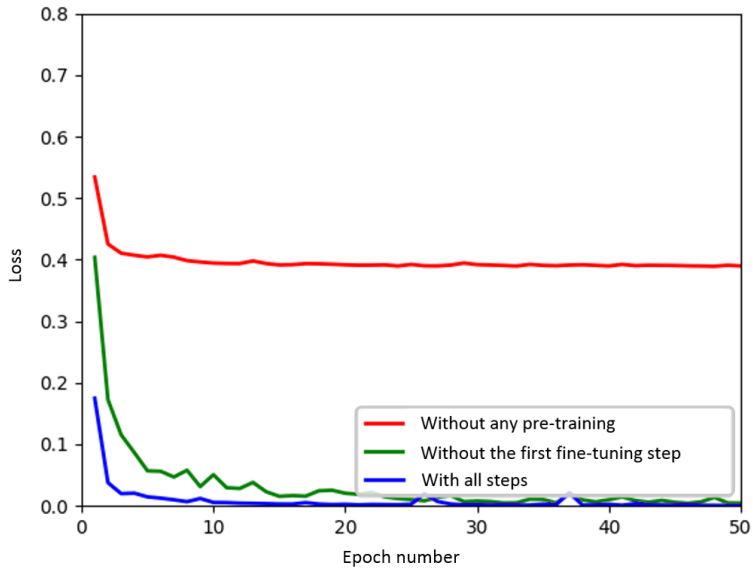


Figure 3.13: – *Loss on the validation data during training epochs without any pre-training, without the first fine-tuning step and training with all steps included.*

step of sequential fine-tuning compared to directly fine-tuning on the pre-trained model have been elucidated. To explain the improvement, our hypothesis is that the size of the applied infant dataset is rather small, which is not suitable for fine-tuning a very large set of parameters of the complete networks at the beginning. For this reason, we first pre-train with a relatively large and similar dataset, in order to pre-learn the networks. In our case, the Shoulder-Pain dataset [89] is quite appropriate, since it is a labeled facial expression dataset of videos from adults. Although the accuracy in the comfort class drops after the first fine-tuning step, the overall accuracy and the accuracy in the discomfort class increase substantially. In this application, it is more im-

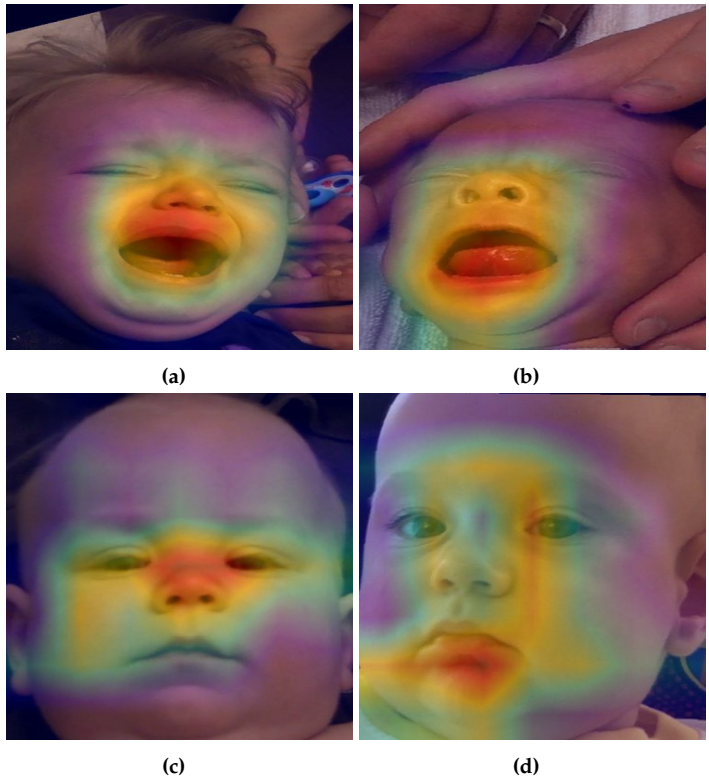


Figure 3.14: – *Examples of discomfort (top row) and comfort (bottom row) faces superimposed by the activation maps. The map highlights the discriminative regions.*

portant to aim at a sensitive system with the purpose to detect more discomfort frames, at the cost of a limited increase in false positive rate. This approach is further elaborated in the next discussion point. The pre-training followed by the fine-tuning steps are the main cause of the high performance of the monitoring system, which makes the overall solution less dependent on the applied type of network.

Sensitivity and clinical application: The proposed automated system is very selective. The area under the ROC curve is rather high (0.96). From the ROC curve, it can be observed that it is possible to keep the sensitivity of detect-

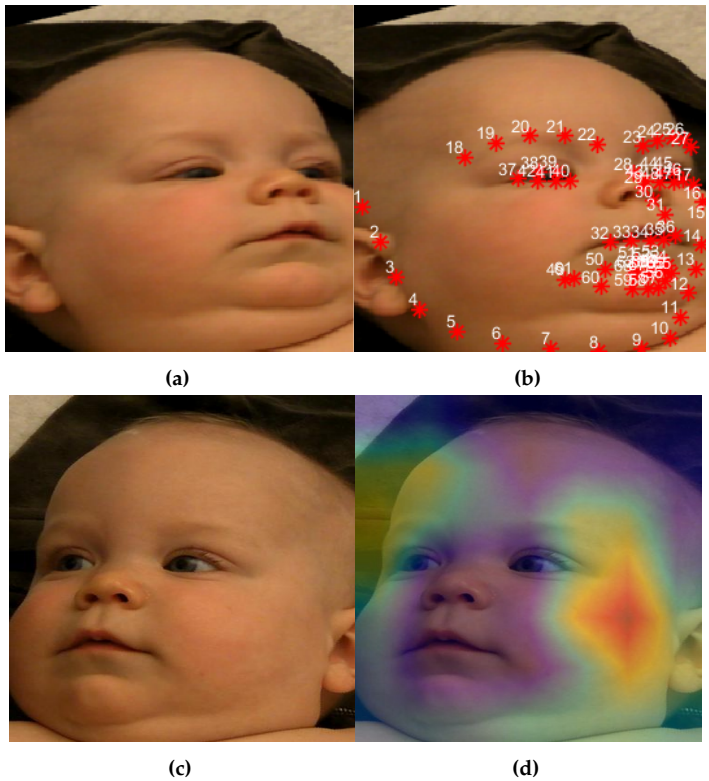


Figure 3.15: – *Examples of false alarms. (a) Failure case from the traditional method, (b) the same result with superimposed corresponding landmarks. (c) Failure case from the proposed deep learning method and (d) original with its superimposed heatmap.*

ing discomfort status of our vision analysis system to be close to unity, while the specificity is 0.78. This means that our system can identify 78 % comfort frames without virtually missing any discomfort frames. For the remaining frames, the clinical system can make decisions on the basis of additional measurements, e.g., comfort scales and manual assessment by nurses. In clinical practice, healthcare professionals expect a discomfort detection system that is sensitive to discomfort moments, while producing false alarms as little as possible. However, the required AUC for such a clinical application is not explicitly known from existing literature when exploiting an automated detection system for this task. In current clinical practice, most hospitals use manual assessment by health professionals. However, one possible solution would be that an observer study is conducted in the future to compare the performance of our system with that of experienced medical staff for further validation. Then, the AUC and accuracy by the proposed automated analysis system should compare favorably to medical staff on the average, while it is also interesting on the requirement of availability of staff.

Comparison with handcrafted features: In our previous work [146], extracting geometric features based on facial landmarks are investigated for discomfort detection. When infants start suffering from discomfort, they tend to squeeze their eyes and stretch their mouths. In order to extract relevant features, the areas of eyes and mouth are calculated by counting the number of pixels inside the polygons surrounded by the landmarks of the eyes and lips. The geometric features achieved an AUC of 0.85 and an accuracy value of 0.78, which is considerably lower than the metrics obtained with the deep learning-based method. This finding is clearly in favor with the proposed approach using CNN-based deep learning.

Failure cases: Both the conventional [146] and the deep learning method are making mistakes occasionally. Figure 3.15 shows a typical example where the conventional method fails and another example where the deep learning method misclassifies comfort as discomfort. The main weakness of conventional features is that it is based on landmark detection with limited contextual information. Consequently, misplacement of landmarks will affect the accuracy of classification. For the deep learning-based method, the robustness relies on the availability of sufficient data, but it learns the complete facial model of the discomfort status. This makes the approach of CNN-based learning suited for the upcoming future, where more data will gradually become available.

Activation maps: We have also computed activation maps for the CNNs, which enables to visualize the focused region in a given image, thereby highlighting the discriminative object parts detected by the CNN. In the studied

cases, eyes and mouth are quite often highlighted, which confirm regions, for which handcrafted features have been extracted in previous studies. This effectively means that the CNN is learning the correct and informative features, which are also used in the clinical pain scoring.

Execution time: The proposed deep learning model has been implemented using the Keras framework. Regarding computation time, the average execution time per frame is 0.013 seconds (i.e. 76 fps) using a GTX-980 GPU in the computing system. This execution speed is sufficient for a real-time application and the applied GPU is a standard version only, which makes the application readily feasible in practice.

3.6 Conclusions

We have developed an automated visual diagnosis system for the classification of discomfort and comfort status of infants in videos. This system applies a deep learning model, based on several stages of pre-trained DenseNet. Using the strategic fine-tuning steps, the proposed model in combination with twofold cross-validation, obtained an overall accuracy of 91 % on a dataset of 6,534 comfort and 10,303 discomfort video frames from 24 infants. The obtained detection accuracy for comfort and discomfort frames are 90 % and 92 %, respectively. By fusing individual frame results, the AUC is further improved from 0.96 to 0.98. This indicates that the visual diagnosis system can be potentially used as an alert system to notify the physicians and nurses on the comfort status of the infants. The medical experts can then combine the decision of the system together with their own assessments. Furthermore, it has been shown that the performance of the deep learning model is improved when using the proposed strategic fine-tuning steps, involving pre-training with generic people pictures and dataset balancing, combined with twofold cross-validation. Using all refinements, the AUC is then substantially increased from 0.77 to an impressive value of 0.96. Although only the DenseNet has been applied and investigated in this work, the use of other state-of-the-art networks is expected to deliver similar consistent results, again confirming the benefits of the proposed fine-tuning steps.

With video capturing, the sound is typically also recorded. Another direction is to capture and analyze the associated sound information for the discomfort detection, since it is easy to distinguish an infant that is feeling discomfort, when the detection system discovers that the infant is crying. To make deep learning applicable, we have extracted 16,837 frames from

24 infants. To further enhance the performance in the future, it is important to recruit more infants for creating a larger database.

When looking to future developments for research, it is logical to consider video-based processing for discomfort detection. The proposed deep learning system is measuring in static video pictures rather than video signals, so that the temporal features such as face/body movement are not yet considered in the algorithm. Additionally, in clinical practice, infant faces are not always visible, which especially holds for preterm infants. The faces of preterm infants in the Neonatal Intensive Care Unit (NICU) are often occluded by nasal cannulas, oxygen masks, or feeding tubes. Therefore, only relying on facial information analysis is not sufficient.

The next chapter will discuss the use of temporal information, which can help in better understanding of the videos by analyzing the motion information contained in the infant movements. It is expected that when infants are in discomfort, the motion pattern will be more expressive than in the comfort status.

Chapter 4

Temporal information: Motion-based discomfort analysis

4.1 Introduction

The previous two chapters have investigated machine learning-based methods for detecting discomfort moments of infants by understanding their facial expressions at video-frame level. The systems have been evaluated using infants from a large range of ages in the field of pediatrics. However, the faces of infants are not always well visible, especially for premature infants in the Neonatal Intensive Care Units (NICUs). The faces of preterm infants in the NICU are often occluded by nasal cannulas, oxygen masks, or feeding tubes. Therefore, discomfort detection fully relying on a facial expression recognition method is not practical. A motion-based method could potentially assist to solve the situations that faces are not well presented, which is also an influencing limitation of the work performed in Chapters 2 and 3. In this chapter, we have constrained the investigation target group to be only preterm infants.

When considering a motion-based discomfort detection system for preterm infants, several important issues should be addressed.

- *Motion estimation:* Body movement is an important indicator for comfort or discomfort when clinical experts visually assess the infant status. It

needs to be investigated whether the temporal information in the video sequences can be extracted by machine learning methods.

- *Leverage of motion signal:* First, motion signals of infants should be extracted from video sequences. The proposed method should extract features that characterize behavior, which are derived from motion signals. The calculated motion-related features can be further leveraged for infant discomfort detection. The key question is that how that should be implemented, given the fact that those features will be already extracted by machine learning methods.

- *Influence of preterm infants:* The faces of NICU infants are always occluded by feeding tubes and/or breathing masks. The proposed method should address this specific group, of which the motion behavior will be different from older infants. The typical pain stimulus caused by a heel prick will be employed as a reference for discomfort status.

Taking into account the requirements and aspects mentioned above, this chapter develops an automated discomfort detection system, which distinguishes discomfort status by analyzing motion patterns. The motion information is first estimated from the videos, which is followed by extracting statistical and spectral features from the motion signals as input for the comfort/discomfort classification.

This chapter is setup as follows. In Section 4.2, the clinical background and related work on pain and discomfort detection is described. Section 4.3 elaborates on the proposed method, while Section 4.4 explains the experimental results, followed by some discussions and conclusions in Section 4.5.

4.2 Related work

4.2.1 Clinical background

Preterm birth is defined as the case when infants are born before 37 completed weeks of gestation. Global prevalence estimation of preterm birth is 9.6 %, influencing approximately 12.9 million infants in 2005 worldwide [10]. Based on statistics for 184 countries, the global average preterm birth rate in 2010 was 11.1 %, giving a worldwide total of 14.9 million infants [14]. Pain/discomfort in infants has received considerable attention from researchers. Early pain experiences of preterm infants have long-term effects on their development,

which include alterations in sensory processing and delay in neurological development. Cumulative pain-related discomfort can lead to long-term perseverance of central nervous system changes and similarly, long-term changes in responsiveness of the neuroendocrine and similar responsiveness issues of the immune systems to stress at maturity [103, 108]. In adults, self-reporting is regarded as the gold standard of pain assessment measurement among patients, since it provides the most reliable indication of pain [93]. However, preterm neonates are not capable of interpreting pain or discomfort in a manner comparable to that of adults. For these reasons, monitoring is required in order to detect pain/discomfort immediately when infants start suffering, which allows caregivers to perform appropriate treatments.

Several pain/comfort scales have been developed to assist healthcare professionals in assessing the pain or discomfort levels of an infant [5, 98]. Each scale is scored by healthcare professionals after observing infants for a few minutes. However, infants are only assessed a few times a day ("spot measurement") without continuous monitoring, which may leave many discomfort moments unnoticed. In this regard, an automated discomfort/pain-assessment method is needed. The objective of this chapter is to develop an automated video-based discomfort detection system for infants, where motion information is integrated into the decision making.

4.2.2 Technical developments

In the past several years, there has been an increasing interest in human stress/pain assessment [45]. Various approaches have been developed to assess pain, based on physiological indicators, for instance vital signs such as heart rate (HR), heart-rate variability (HRV), respiratory rate (RR), oxygen saturation (SpO2), body temperature, and blood pressure. These signs can be measured and used to assess the physical functioning level of a person. Acharya *et al.* [2] detected cardiac abnormalities by classifying cardiac rhythms using an artificial neural network and fuzzy relationships, which achieved an accuracy level of 80-85 %. However, vital signs such as HR and RR are currently measured by electrocardiograms (ECG) and photoplethysmography (PPG), which require contact with the patient's skin. Attaching the sensors to infant skin adds an extra burden to infants compared to a contactless method, for instance, using remote video monitoring.

Besides the physiology-based approaches, there is another category of methods that assess pain/discomfort, which is based on behavior analysis. Existing behavioral-based approaches to evaluate infant pain can be based

on facial expression and crying sound [182] [36]. The crying of infants is a common sign of discomfort, hunger, or pain. For classifying crying sound, Mima *et al.* [94] presented a method that analyzes baby cries in spectrography, and classifies them as cries due to pain, sleeping, hunger, etc. The obtained overall accuracy of the proposed method was 85 %.

Significant attention was paid to facial expressions in adults. Shan *et al.* [131] empirically evaluated facial representation based on statistical local features, local binary patterns (LBPs) for person-independent facial expression recognition, and illustrated that LBP features are effective and efficient for facial expression recognition. Kotsia *et al.* [73] achieved a recognition accuracy of 99.7 % for facial expression recognition, using the proposed multi-class SVMs and obtained 95.1 % for facial expression recognition, based on a set of chosen Facial Action Units (FAUs).

At present, very few studies reported pain assessment for premature infants based on body movements. Cattani *et al.* characterize clonic seizures and apneas by the presence or absence of periodic movements of parts of the body [18]. However, the method is intensity-based (in the paper called luminance) and thus is affected by the lighting condition.

The existing literature does not show research work on normal behavior of infants in relation to comfort and pain. The work of Cattani *et al.* investigates abnormal behavior using body movement analysis, whereas Kotsia *et al.* look to the facial expression without considering the constraints of the NICU settings. The facial expression of premature infants in the NICU is typically occluded by breathing masks/feeding tubes, so that the assessment should not be fully based on a facial expression method for discomfort monitoring. A good alternative is video-based body motion analysis.

The following work proposes a video-based automated system for detecting discomfort moments of premature newborns in the NICUs by analyzing motion information. Because this information is derived from the behavior of preterm infants, specific attention has to be given to the accuracy of the motion analysis.

4.3 Methods for motion-based analysis

To detect the body motion of the infants, we employ optical flow to estimate pixel-based motion vectors between frames, which is followed by feature extraction for discomfort/comfort classification. The choice for optical flow is

motivated by its capability to capture the motion of individual body parts more accurately [33]. This is required because preterm infants have a less developed muscle system compared to older infants and adults.

4.3.1 Study design and population

The study was conducted with videos recorded at the Máxima Medical Center in Veldhoven, the Netherlands, by a fixed-position high-definition camera (uEye UI-222x) filming the infant's face and upper body. Figure 4.1 illustrates the video acquisition system. Since the experimental procedure (heel lance) is part of regular neonatal care, the ethical committee of the Máxima Medical Center provided a waiver [N17.178] for this study. For all infants, written consent was obtained from the parents.

The heel lance procedure is a well-known pain stimulus and is part of regular care for collecting blood samples to monitor glucose, bilirubin, etc. In this work, it serves as a recurring stimulus to study the infant's response to pain. The video recording started approximately 10 minutes prior to the heel lance procedure. When the heel lance procedure was finished, the video recording continued for an observation time to return to baseline (10 minutes). We define the baseline as a period when there is no observable discomfort motion pattern/facial expression. The start and end time of the heel lance procedure was simultaneously noted by a research assistant.

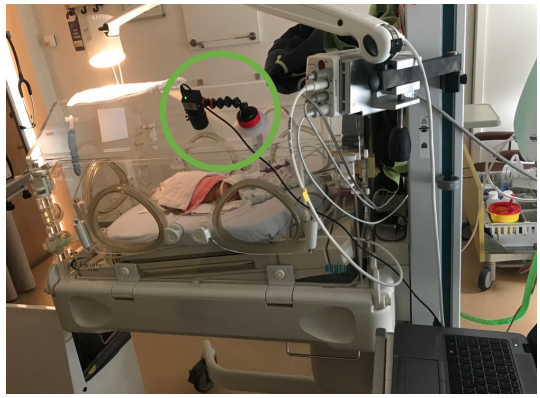


Figure 4.1: – *Video acquisition system and setup in a NICU. The camera is outlined by a green circle.*

Discomfort and comfort video segments were annotated by a researcher according to the timeline relative to the heel prick intervention. Discomfort video segments were labeled from the start point of the heel prick to several minutes after the heel prick was done, based on the researcher's observation. "Comfort" video segments were labeled from the baseline prior to the prick and from the moment onwards when the infant returned to baseline after the heel prick. Each video segment contains only one state (comfort or discomfort). All the moments that show interruption or occlusion from caregivers in the videos are excluded. The video segments lasting less than 10 seconds are removed from consideration. Data of eleven infants with an average gestational age of 31 weeks are collected for the experiments. This results in totally 99 discomfort (2,738 seconds in total) and 84 comfort (3,429 seconds in total) video segments for 17 heel prick events. The duration of each video segment varies from less than one minute to several minutes with an overall median length of 21.2 seconds (interquartile range [IQR] 12.8 - 39 seconds).

4.3.2 Motion estimation

Pixel-based motion vectors are first calculated for each video frame, with respect to the previous frame, using the optical flow proposed by Farnebäck *et al.* [33]. The Farnebäck method models the neighborhoods of each pixel by quadratic polynomials to calculate the optical flow, where the involved optimization of the algorithm is performed at a neighborhood level rather than pixel level. In our study, we compute a motion matrix $\mathbf{M}$ of size $N \times L$, where N is the number of pixels in a video frame and L is the total length of a video segment in frames. The optical flow provides motion derivatives for each row of the matrix that represents the velocity magnitude of a pixel's trajectory.

We accumulate the magnitude values of all motion vectors for each frame. Hence, all the summed magnitude values comprise a one-dimensional (1D) velocity-estimating signal $\mathbf{V}$ (size $1 \times L$) for each video segment. The motion acceleration rate $\mathbf{A}$ (size $1 \times L$) is further estimated by taking the first derivative of the velocity-estimating signal $\mathbf{V}$. Figure 4.2(a) and Figure 4.2(b) show examples of the motion acceleration rate, extracted from a comfort and a discomfort video segment, respectively.

4.3.3 Feature extraction

For each video segment, features are extracted from the 1D signal vector of the motion acceleration rate $\mathbf{A}$, capturing the motion of all pixels in the image.

We calculate two groups of features, which have a statistical and spectral content, and are shown in Table 4.1. In total, 18 features are computed: mean, median, root mean square (see Eq. (4.1)), a group of 3 features from the autocorrelation function, and a group of 12 spectral peak features, which are discussed below. The first statistical parameters are a regular choice, while the autocorrelation is used to identify the major changes in the signal caused by different comfort/discomfort status of infants. The spectral peak features are for analyzing the frequency-domain patterns.

A. Mean, median of the motion acceleration rate A, root mean square
This involves the computation of the mean and median values of the motion acceleration rate vector **A**, by averaging all elements of the vector or computing the median of all elements.

The root mean square (RMS) of the motion acceleration rate elements from

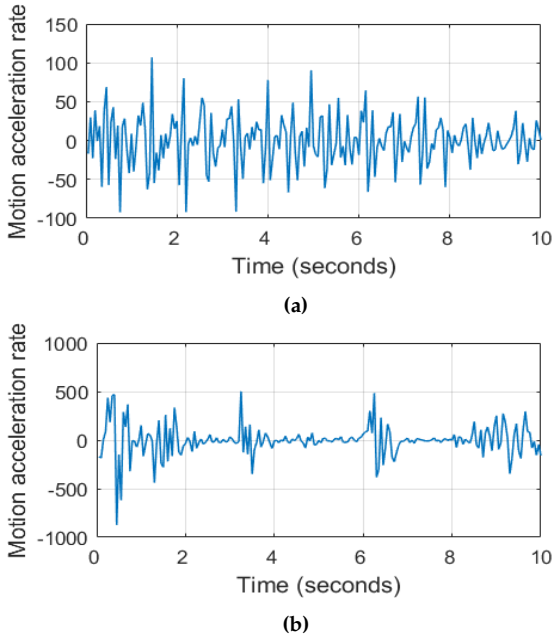


Figure 4.2: – *Examples of extracted motion acceleration rate of (a) comfort, and (b) discomfort moments, where different motion patterns are visible in terms of movement intensity and periodicity.*

A is utilized as an input to a classifier for infant status recognition. The RMS value is calculated according to:

$$RMS = \sqrt{\frac{1}{L}\sum_{i=1}^{L} {x_i}^2}, \tag{4.1}$$

where x_i is an individual acceleration instance and L denotes the total number of frames of the video segment.

B. Autocorrelation features

The autocorrelation characterizes different types of motion signals by indicating motion intensity and periodicity. The autocorrelation of motion acceleration rate **A** is then calculated, which is formulated as

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{L} (\mathbf{A}_t - \mathbf{A}_{\text{avg}}) \cdot (\mathbf{A}_{t-k} - \mathbf{A}_{\text{avg}})}{\sum_{t=1}^{L} (\mathbf{A}_t - \mathbf{A}_{\text{avg}})^2}, \tag{4.2}$$

where $\mathbf{A}_{t-k}$ is the motion acceleration rate shifted over k frames, and $\mathbf{A}_{\text{avg}}$ is the average of the motion acceleration rate. The numerator of Eq. (4.2) is essentially the covariance between the original acceleration rate and the k-frame shifted data. The denominator is the sum of the squared deviations of the original acceleration rate.

The peak height of the autocorrelation function at zero-th lag (k=0) (overall energy) is employed as an individual descriptive feature. The other two features are the height and location at the first peak of the autocorrelation, respectively, which identify the dominant cyclic variation in motion.

C. Spectral peak features

We further estimate the power spectral density of the motion acceleration rate **A** using Welch's method [48] with a rectangular window, having the length equal to the total number of frames for each video segment. From the derived spectrum, positions and power levels of the highest 6 peaks are taken as 12 spectral peak features.

Feature category	Indication	CMVAL	DMVAL
Mean	Motion intensity	0.015	-0.275
Median	Motion intensity	-0.091	0.296
RMS	Motion intensity	17.1	237
Autocorr. - height at 0th lag	Motion intensity	$2.16 \cdot 10^5$	$2.45 \cdot 10^7$
Autocorr. - 1st peak location	Motion periodicity	0.2	0.2
Autocorr. - 1st peak height	Motion intensity	$1.20 \cdot 10^4$	$3.34 \cdot 10^5$
Spectrum - positions of highest 6 peaks	Motion frequency	3.63	1.88
		4.67	3.07
		5.41	3.93
		6.14	5.16
		6.86	6.22
		7.59	7.28
Spectrum - power levels of highest 6 peaks	Motion level at each frequency	115	$2.21 \cdot 10^4$
		150	$2.30 \cdot 10^4$
		150	$2.54 \cdot 10^4$
		121	$2.68 \cdot 10^4$
		119	$2.66 \cdot 10^4$
		129	$2.26 \cdot 10^4$

Table 4.1: – *Feature categories with their corresponding indication and median values for all comfort (CMVAL) and discomfort cases (DMVAL).*

4.3.4 Classification

Finally, we have adopted a support vector machine (SVM) classifier on video segments to recognize infant status of comfort or discomfort, using the 18 extracted features. A linear kernel is employed, since a sigmoid-based SVM does not offer higher performance in our experiments and only increases the complexity and execution time. We have used the SVM implementation of Matlab (Mathworks, Natick, MA, USA) for the two-class classification. Leave-one-infant-out cross-validation is used for the experiments. The receiver operating characteristic (ROC) is plotted to evaluate the performance with the value of

the area under the curve (AUC). The classification accuracy is also measured and reported as an evaluation metric.

4.4 Experimental results

Using the recorded infant dataset, leave-one-infant-out cross-validation is performed for evaluating the proposed method. The classification accuracy is summarized in Table 4.2, and the ROC curves are plotted in Fig. 4.3. From

Feature category	Accuracy rate	AUC
Mean	0.50	0.72
Median	0.39	0.53
Root Mean Square (RMS)	0.81	0.91
Autocorrelation	0.68	0.84
Spectrum	0.75	0.91
Combined	**0.86**	**0.94**

Table 4.2: – *Performance measures for classification, including classification accuracies and AUCs of different features.*

the light-blue curve based on all features, it can be observed that the system can detect 85 % of discomfort video segments at the cost of only 10 % of false alarms. When all features are combined, the average accuracy for all infants is 0.86 (see Table 4.2). When applying each category of features individually, the RMS feature achieves the best accuracy and AUC. The AUC is significantly improved when combining all features, compared to the case that all are used except the RMS feature ($p = 0.006$ using bootstrapping). Although the median shows a low performance as a standalone feature, we have empirically found that it still contributes to the overall performance when combined with other features.

4.5 Discussion and conclusions

Performance: We have extracted only 18 features from each video segment. In our Matlab implementation, the computation time is approximately 0.4 seconds per frame on a regular computer with one E5-1650 CPU (3.60 GHz), which

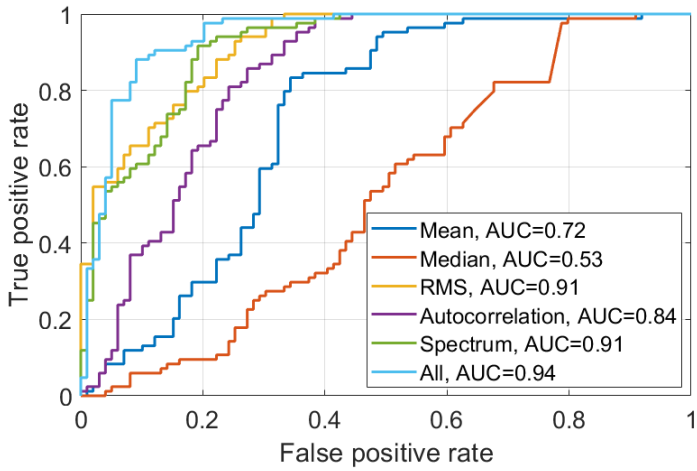


Figure 4.3: – *ROC curves for classification of comfort/discomfort, using each individual category of features and when combining all features together.*

can be further optimized by implementing the code in C++ for a real-time application. In future, a further feature selection may be attractive to come to a good performance trade-off. The comparison with previously developed methods is considered as part of future work. An important published system from [184] is based on the same topic, but has employed different features and lacks the detailed motion analysis that we have proposed. Since we have added this motion analysis, we clearly outperform the results of this work from literature [184].

Cases with occlusions and critical illness: The experiments have shown that including autocorrelation and spectral features yields an enhanced performance in detecting comfort and discomfort status of infants. However, in clinical practice, video interruptions may occur by care-handling the infant and, thereby lead to occlusion of the infant's face and/or body. Moreover, for critically ill infants, no visual response to pain stimulus can be observed. For such cases and further work, we may have to add features extracted from the monitored vital signs, such as heart rate, respiration rate, and blood oxygen saturation [165] [135] [25].

Sensitivity-specificity trade-off: For technically optimizing system performance, further lowering the amount of false positives would be valuable, thereby improving the specificity of the system. However, in real clinical

practice, a majority of these false positives are segments with irregular movements during comfort moments (e.g. stretching an arm). With limited effort of human assistance (e.g. visual check), these false positives may be easily corrected. This implies that the nursing personnel has some understanding of the system usage.

As a conclusion, for the purpose of anticipating on pain and discomfort in premature infants, we have proposed an automated video-based system that can differentiate discomfort of infants from comfort status by analyzing motion patterns. The recognition process starts with the acquisition of the motion signals, which are subsequently estimated by optical flow. For each video segment, a vector of 18 features is extracted from the motion signals, containing characteristics of both the time and frequency domain, describing the motion trajectories. These feature vectors are then jointly classified by an SVM. An AUC of 0.94 is achieved, which is promising for clinical practice. The highest AUC is obtained when combining all proposed features, which proves that the features are contributing and complementary. The proposed system shows that motion-related features can differentiate discomfort status from comfort. Because of incorporating autocorrelation and spectral features that can capture and describe periodic motion patterns, our system can recognize discomfort body motion and regular periodic movements (e.g. respiration). With the high sensitivity of 85 %, the trade-off of the system is only 10 % detecting false positives.

The next chapter will further explore the 1D motion signals using a deep learning approach, which aims to better analyze the motion information. Conventional feature extraction is typically outperformed by deep learning approaches for visual analysis. This motivates that further experiments with deep learning for this purpose is a logical follow-up step.

Chapter 5

Applying deep learning on 2D representations embedding time-frequency information

5.1 Introduction

The previous chapter has started assessing the infant comfort status by analyzing the movement patterns of their bodies from videos. The proposed system first extracts the motion signal using optical flow algorithms and then combines the designed handcrafted features using an SVM classifier for prediction. However, the feature extraction step may lose informative descriptions of the signals, and features are sensitive to parameter-tuning of the SVM. Inspired by the recent success of deep learning, a data-driven approach is desired, in which the machine learning system replaces partly the feature extraction and the comfort classification.

However, developing a deep-learning system for characterizing motion signals has to address the following challenges.

- *Motion representation:* Most mainstream deep learning CNNs are based on 2D image inputs. Therefore, one possible solution could be to convert the 1D motion signal to 2D representations as inputs for the CNNs.
- *Effective learning:* Generally, CNNs require a large amount of data for

training. However, in this study, because of the limited amount of infants, dealing with small data remains challenging.

- *Optimizing the inputs and networks:* As mentioned previously, when the size of the dataset is small, having an informative/descriptive input representation available is especially important for training CNNs.

- *Selection of representative input features:* Among different 2D feature representations, it is important to investigate the performance of each individual feature and select the most effective ones for comfort classification.

Taking into account the above-mentioned challenges and concerns, we aim to reuse an already existing concept to convert audio signals into spectral representations for descriptive analysis. In our case, we would convert then the 1D motion signals to a number of feature representations, which are then explored for further analysis. The detailed processing will be further explained later in this chapter.

As a subsequent step, we employ state-of-the-art CNNs on 2D representations, in which the motion information is embedded. The CNNs with/without pre-training are investigated and the models further transfer the learning on our dataset, since pre-training and transfer learning are effective techniques for small-dataset learning. We have conducted and elaborated experiments to select the best combination between representations and CNNs.

The purpose of this chapter is essentially an extension of the previous chapter in the sense that the motion signal is described in an alternative way and incorporated in a learning network. This explains why we omit the related work section and provide only some background on the conceptual changes.

The remainder of this chapter is organized as follows. First, Section 5.2.2 and Section 5.2.3 revisit the methodology of motion extraction and introduce the conversion of each motion-signal segment into an image representation. Second, in Section 5.2.4, state-of-the-art deep learning networks are exploited for the image classification task, followed by analyzing the effectiveness of the combination of representations and networks in Section 5.3. Section 5.4 discusses the advantages and disadvantages of this approach. Finally, this chapter is concluded in Section 5.5.

5.2 Methods for 2D representations of motion and classification

5.2.1 Study design

In our work, Heel Prick (HP) is a well-known recurring pain event, which was used as a stimulus to study the infant response to pain. The study was conducted at the Máxima Medical Center in Veldhoven, the Netherlands. As the experimental procedure (HP) is part of regular neonatal care, the ethical committee of the Máxima Medical Center provided a waiver [N17.178] for this study. For all infants, written consent was obtained from the parents for the experiments. A camera (uEye UI-222x, IDS imaging, Germany) was used to record the infant face and upper body in a fixed position.

Videos segments were manually annotated for comfort/discomfort by a medical doctor along the HP procedure. The comfort/discomfort video segments were labeled by visual observation, according to the timeline relative to the HP intervention. Comfort video segments were annotated from the baseline prior to the prick and the time period when each infant returned to baseline status after the HP. Discomfort video segments were annotated from the starting point of the heel prick to several minutes after the heel prick was finished. Each video segment is associated with only one infant state (comfort or discomfort).

Eleven infants with an average gestational age of 31 weeks (range $27^{+1\,\text{day}}$–$38^{+5\,\text{days}}$ weeks) were recorded. In total, we obtained 99 discomfort (2,738 seconds) and 84 comfort (3,429 seconds) video segments from 17 HP events.

5.2.2 Motion estimation

First, we employ an optical flow algorithm to estimate pixel-based motion vectors between adjacent video frames, to extract body motion of the infants for each video segment. The utilized optical flow method proposed by Farnebäck *et al.* [33] enables capturing the motion from individual body parts. Dense optical flow is estimated by modeling the neighborhoods of each pixel using quadratic polynomials, and the optimization is done at pixel-cluster level rather than individual pixel level. We accumulate the magnitude values of all motion vectors for each video frame. Hence, all the summed magnitude values comprise a One-Dimensional (1D) signal vector of motion velocity magnitudes for each video segment, containing multiple frames. A segment typically lasts

about 10 seconds. We further estimate the motion acceleration rate by taking the first derivative of the velocity magnitude signal. The motion acceleration rate serves as a key descriptor for further analysis.

5.2.3 Image representation

Following motion estimation, each 1D signal (motion acceleration rate) is clipped to 10-sec. long segments for further processing. One important step for the classification task is to identify the primary information, characterized by the discomfort motion pattern, while discarding other details that carry background noise, random movements, etc. The shape of the 1D signal manifests itself in the envelope of the short-time power spectrum [164]. For this reason, we concentrate on extracting features from the envelope shape to represent motion signals indicating the type of behavior. Three methods of feature extraction are investigated and compared for data representation, namely: Log Mel-spectrogram (LMSpec), Mel-Frequency Cepstral Coefficients (MFCCs), and Spectral Subband Centroid Frequency (SSCF). These methods have been found to be effective in extracting and combining the frequency and magnitude information from the power spectrum [12, 39, 119, 124].

MFCCs are features widely used to represent characteristics of signals in voice recognition. [84] We first execute overlapping sliding windows over the input signal segment, and then compute the Fourier transform over each window. A Mel-filterbank is further applied, and the energies within each filter are accumulated. The Mel-spectrogram is therefore obtained as a graphic image describing the energy content as a function of Mel scales (output of filters) versus time windows. We take the logarithm (log) of the filterbank energies and employ the derived Log Mel-spectrogram as our first type of image representation (see Fig. 5.1 (b) and Fig. 5.2 (b)).

Following the Log Mel-spectrogram calculation, a Discrete Cosine Transform (DCT) is applied on the log-filterbank energies, resulting in 12 MFCC values (DCT coefficients) per sliding window. The total energy per sliding window is also included as a feature. As a result, 13 MFCC feature values are obtained in total, where the last one is the total energy (sum of the initial 12 MFCC values). Concatenating these features leads to a time-frequency representation that can be visualized as a heat map, which is our second feature-image representation, namely MFCCs. In total, each heat map consists of time windows represented on the horizontal axis, and 13 MFCC filterbanks represented on the vertical axis (see Fig. 5.1 (c) and Fig. 5.2 (c)).

Furthermore, we compute the SSCF from the 1D signal segments. The

SSCF represents the centroid frequency in each subband, whereas in MFCC features, the power spectrum in a given subband is distributed and the detailed frequency information is not readily available. Therefore, we consider that the SSCF provides supplementary information to MFCCs (see Fig. 5.1 (d) and Fig. 5.2 (d)).

5.2.4 Classification of the 2D representations

The results of processing the image representation allow each 10-second instance of the motion signal data to be processed as an image, so that energy values over time can be visualized as a heat map.

The derived heat maps are further classified using deep convolutional neural networks. A ResNet (see Fig. 5.3) is used as our classification model. He *et al.* [50] proposed ResNet, which is a network architecture where the layers are explicitly reformulated as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. It was shown that ResNet is easier to optimize, and can gain accuracy from considerably increased depth in the network on the ILSVRC 2015 classification task. To alleviate the limitation of small dataset size in this study, we employ transfer learning using the pre-trained models. We have employed a ResNet with 18 residual blocks, pre-trained with the ImageNet dataset. Each residual block consists of a section where the input is both processed and bypassed and compared at the end to measure the residual, which is kept identical to the original ResNet publication.

5.2.5 Evaluation

The proposed method is evaluated by performing leave-one-infant-out cross-validation to obtain an unbiased label for each video segment. The training set is further split into a real training set (70 %) and a validation set (30 %) on patient level. The validation set is used to refine the model from each training epoch. We perform training with the number of epochs equal to 25, a batch size of 16, and employ the Adam optimizer [70] without dropout.

The Receiver Operating Characteristic (ROC) is plotted to evaluate the performance with the value of the Area Under the ROC Curve (AUC). The classification accuracy and confusion matrix are also computed and reported as evaluation metrics.

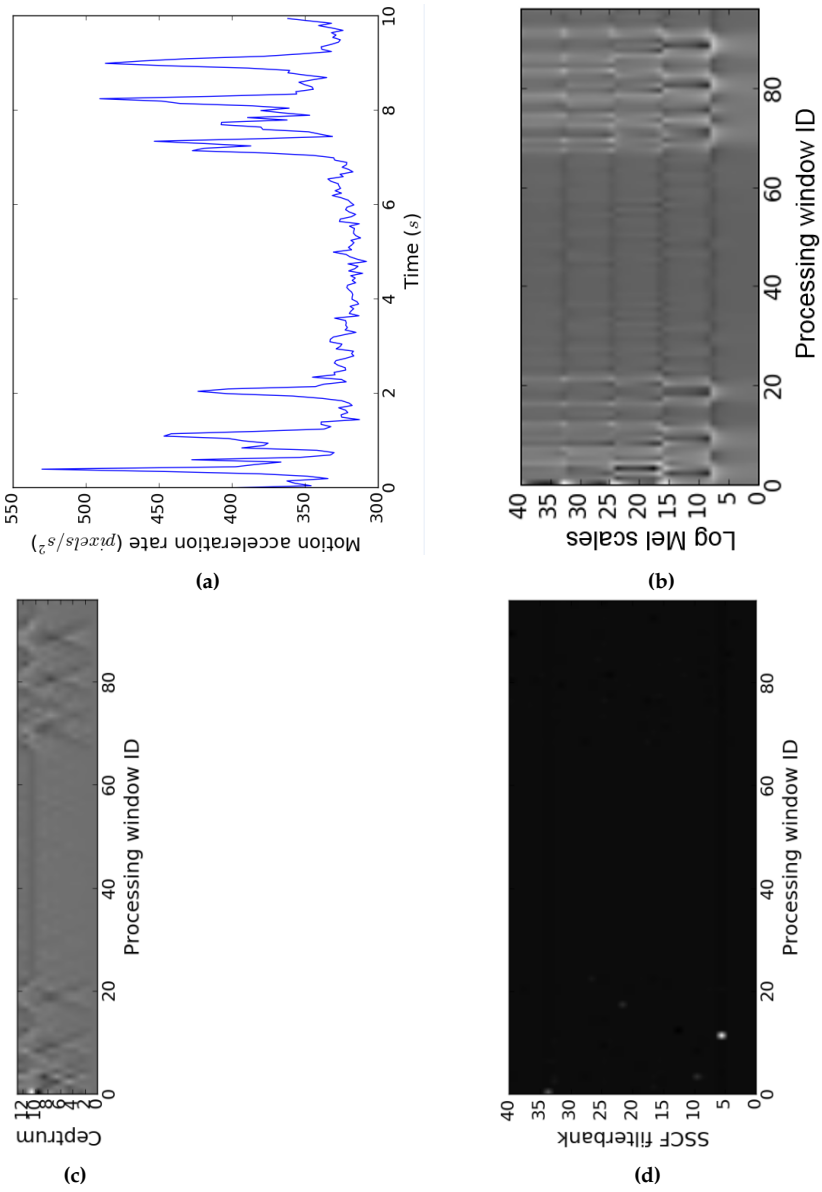


Figure 5.1: – *Example of a discomfort motion segment. (a) Extracted 1D motion signal, which is analyzed further with a sliding window (window size = 500 ms and step size = 100 ms). Feature images for the 10-sec. motion segment are shown in (b) the Log Mel-spectrogram, (c) image of the MFCCs and (d) SSCF visualization. The last three subfigures all use the same window processing.*

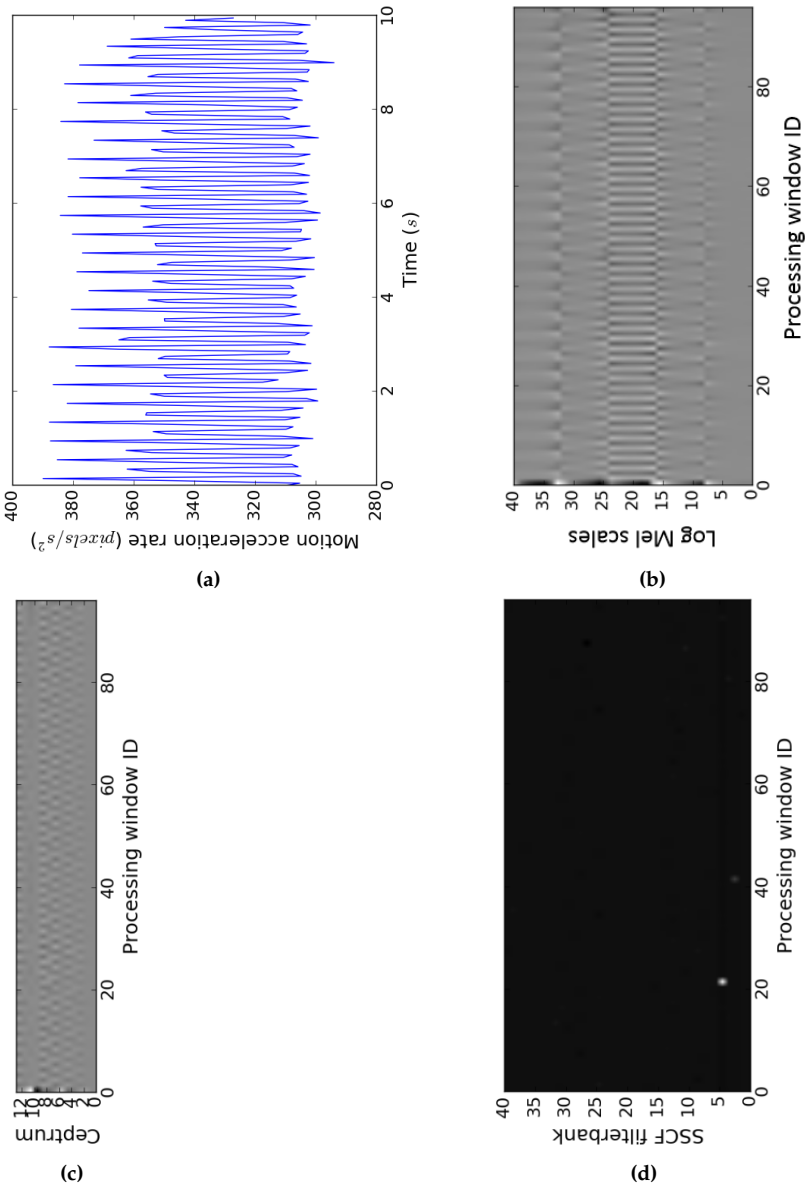


Figure 5.2: – *Example of a comfort motion segment. (a) Extracted 1D motion signal, which is analyzed further with a sliding window (window size = 500 ms and step size = 100 ms). Feature images for the 10-sec. motion segment are shown in (b) the Log Mel-spectrogram, (c) image of the MFCCs and (d) SSCF visualization. The last three subfigures all use the same window processing.*

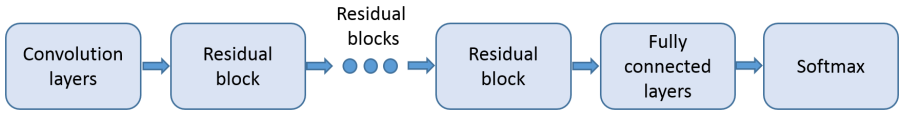


Figure 5.3: – *General architecture of the ResNet. The open part of the residual block chain refers to extra 16 residual blocks.*

5.3 Experimental results

We have conducted experiments with different frameworks of applying ResNet: 1) only fine-tuning the Fully Connected Layers (FCL) of a pre-trained ResNet, 2) fine-tuning all layers of a pre-trained ResNet, and 3) directly training ResNet using our dataset. Combining all features together, the performance of applying different frameworks for the binary classification is shown in Table 5.1. The results from our previous work [141], as also described in Chapter 4 using handcrafted features on the same dataset, are added in the table for reference. Fig. 5.4 shows the normalized confusion matrix for only fine-tuning the fully connected layers. Fig. 5.5 represents the corresponding ROC with the AUC of 0.985. The ROC curve indicates that approximately 90 % comfort video segments can be correctly determined by our automated system without missing any discomfort moments.

The AUC values for applying each type of image representation individually are 0.978 for the LMSpec, 0.961 for MFCCs and 0.677 for SSCF.

Training scheme	Accuracy	AUC
Fine-tuning only FCL	94.2%	0.985
Fine-tuning all layers	93.3%	0.978
Training from scratch	80.8%	0.878
Handcrafted features [141]	86.0%	0.940

Table 5.1: – *Performance of different training schemes in terms of classification accuracy and AUCs.*

5.4 Discussions

Omission of information: In this study, although a deep learning system is proposed, it is still not an end-to-end system for learning the essential information, since we still need to extract motion information from videos and then also have to convert motion signals to 2D representations. More importantly, the extraction of motion certainly lacks additional spatial/contextual information related to diagnosis as this signal only encodes the motion magnitude, whereas the facial expressions and the positions of infant bodies are completely omitted for analysis.

Length of video segments: Another limitation of this study is that we only take video clips of a specific time-window for the analysis, while the optimal duration of the time window is not explored. When choosing such a window duration, a trade-off between computation complexity and information sufficiency should be incorporated.

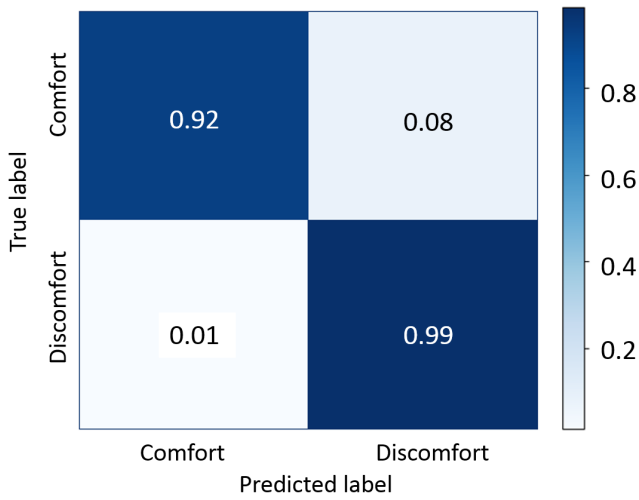


Figure 5.4: – *Normalized confusion matrix of comfort and discomfort classification with only fine-tuning the fully connected layers of ResNet.*

5.5 Conclusions

In this chapter, we have proposed an automated video-based system using deep learning that can differentiate the discomfort status of infants from comfort status by analyzing motion patterns. When infants experience discomfort moments, the system draws the attention of caregivers to the infants. The processing chain includes three steps: 1) 1D signal extraction using optical flow, 2) converting the 1D signal to a feature-image representation, and 3) deep learning classification of the feature images.

Analyzing image representation of motion signals facilitates to use richer frequency-related information in addition to the motion-related time series. This approach has resulted in the following contributions: 1) an algorithm is defined for extracting motion and then converting each segment to time-frequency image representations, 2) a deep learning scheme by Convolutional Neural Networks (CNNs) is employed to classify the feature images, and 3) an automated video-based system is realized for detecting discomfort moments in preterm newborns hospitalized in NICUs.

For each video segment, three image representations characterizing motion

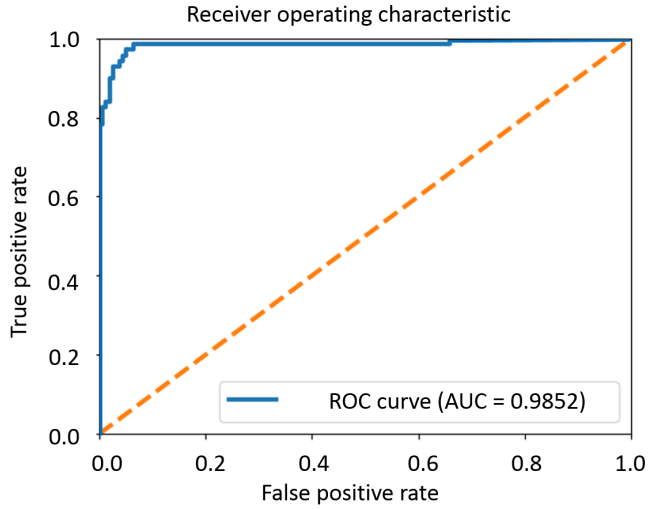


Figure 5.5: – *ROC curve of binary classification of comfort and discomfort with only fine-tuning the fully connected layers of a pre-trained ResNet.*

trajectories are extracted from the motion signals in the time and frequency domain. An AUC of 0.985 is achieved, which is promising for use in clinical practice. The highest AUC is obtained when combining all three image representations, namely the LMSpec, MFCCs, and SSCF, by transfer learning from a pre-trained ResNet, which proves that the different image representations are all contributing and complementary to each other. Although the use of motion-based features is promising, we have also discussed that the system can be improved by adding contextual features from the face and body.

In the future, more infant data will be collected and used for evaluating our system. An end-to-end system will be investigated as a feasibility study. Since 3D CNNs have emerged as a possibility, they could provide opportunities for video analysis without overhead on information conversion. In the next chapter, the usage of 3D CNNs for the comfort classification task will be explored to further determine the infant status of comfort or discomfort.

Chapter 6

Attention-based 3D CNN approach

6.1 Introduction

The previous chapters have analyzed static facial images in order to detect infant discomfort status, and also investigated infant face/body motion information. Chapter 2 has employed the handcrafted features combined with an SVM to analyze the infant facial expressions at an individual frame level. Chapter 3 has continued the task of frame-based facial expression classification by leveraging deep learning-based methods and several steps of fine-tuning strategies. Chapter 4 has started to explore the temporal information from the infant videos and has investigated the embedded motion signals by using optical flow to extract the motion magnitudes. Handcrafted features are designed for the 1D motion-related signal to classify the infant status of comfort/discomfort by an SVM. Chapter 5 has further strengthened the work of Chapter 4 by presenting the motion-related features using 2D spectral feature images. The 2D representations are finally classified by using deep networks.

In this chapter, we are going to fully leverage from video information by simultaneously incorporating both contextual facial expression and motion information. In order to exploit both the spatial contextual information (e.g. appearance) and the temporal information (e.g. motion) in a single optimization framework for discomfort/comfort classification, we are going to explore a 3D-CNN method that can include both spatial and temporal features in one net-

work. This means that the new network with the learned CNN features will jointly exploit spatio-temporal features and thus multiple signal dimensions simultaneously.

We aim to propose a 3D CNN for capturing temporal and spatial changes in infant facial appearance and body and limb positions, which addresses the following challenges.

- *Attention on motion regions:* Body movement is an important indicator for comfort or discomfort when clinical experts visually assess the infant status. The studies in Chapters 4 and 5 have also demonstrated that different motion patterns are visible for comfort and discomfort status. For example, facial expression changes incur motion patterns, while pain can also cause significant motion in the body and limbs. Therefore, the proposed network should pay attention to various regions and types of motion where significant motion occurs.

- *Processing time-window selection:* An optimal length of a time window for processing the video stream has to be investigated. For each video sequence, a temporal sliding window with a fixed length is used as an input clip for the classification network. The optimal length of the sliding window needs to be chosen carefully according to the classification performance and computation complexity.

- *Process spatial and temporal information simultaneously:* Combining facial expression and motion information together is expected to achieve better performance than considering each component individually. As a data-driven approach, the proposed network should be able to simultaneously learn contextual information from both infant face, body and limbs, and also the network architecture should be capable of handling multiple dimensions of the input.

Taking into account the requirements and aspects mentioned above, this chapter develops a novel 3D-CNN architecture. This is followed by investigating different inputs to the network to understand the importance of temporal information (motion) to discomfort classification, and the involvement of additional contextual features from the spatial domain. More specifically, the network has to adapt and learn from multiple inputs, such as 3-channel RGB, 2-channel motion, and 5-channel combination of RGB and motion. For comparison with state-of-the-art systems, a thorough benchmark is performed between 2D CNN and 3D CNN, also between different channel inputs. It is the

objective to eventually arrive at an overall best solution that uses five channels of input for the 3D CNN with a short time duration, namely "multi-channel attention 3D CNN". The CNN training and validation are performed on a real clinical dataset with various infants, which is intended for the study of infant discomfort/comfort.

The remainder of this chapter is as follows. Section 6.2 discusses the existing related work on both conventional and state-of-the-art deep learning-based methods. Section 6.3 elaborates on the proposed methods, while Section 6.4 describes the experimental setup for evaluating the methods. Section 6.5 presents the experimental results. Finally, Section 6.6 re-discusses the proposed method and concludes the chapter.

6.2 Related work

6.2.1 Image-based emotion classification

Studies in emotion recognition have focused on image-based and video-based approaches [29]. Video-based approaches have shown improved recognition performance, since they can exploit temporal features and associate those with emotion changes [32]. For the applications of 2D image processing in infant discomfort detection, extensive research has concentrated on facial expression recognition. Sun *et al.* [147] proposed a sequential fine-tuning strategy to classify 2D images from infant videos and achieved an Area Under the Curve (AUC) value of 0.96. By fusing individual frame results, the AUC was further improved from 0.96 to 0.98. Meng *et al.* proposed Identity-Aware Convolutional Neural Network (IACNN) for facial expression recognition. The results showed an accuracy of 71.3 % when testing on the CK+ dataset [63] [88]. Liu *et al.* [83] upgraded a single CNN to an ensemble of CNNs and the best single subnet achieved 62.4 % accuracy, while the whole model scored 65 % accuracy. However, this ensemble approach may be less suitable for real-time applications. Mollahosseini *et al.* [96] presented a deep neural network architecture for automated facial expression recognition, which consists of two convolutional layers, followed by max-pooling and four inception layers. The inception layers increase the depth and width of the network while keeping the computational budget constant. The work was evaluated on the CMU MultiPIE face database [40] and achieved an accuracy of 94.7 %. Uddin *et al.* [160] leveraged a depth-camera-based solution for efficient facial expression recognition, in which for each pixel in a depth image, eight local direc-

tional strengths are obtained and ranked. Incremental to different 2D CNN approaches, Li *et al.* [78] showed the benefits with data augmentation, including face cropping and rotation. Zhao *et al.* [190] proposed a novel set-to-set (S2S) distance measure to calculate the similarity between two sets, in order to improve the recognition accuracy for faces with real-world challenges. For the S2S distance, the kNN-average pooling is adopted for computing the similarity scores. Luan *et al.* [87] proposed Gabor convolutional networks (GCNs), which utilize Gabor filters as the convolutional filters, such that the robustness of learned features against the orientation and scale changes are reinforced. Zhang *et al.* [186] developed a new representation learning method, named Structure Transfer Machine (STM), which enables the feature learning process to converge at the representation expectation in a probabilistic way.

6.2.2 Video-based emotion analysis

In terms of further exploiting temporal information in a video, less attention has been paid to facial expression recognition. One recent relevant work is presented by Sun *et al.* [141], where the motion acceleration rate and 18 time-domain and frequency-domain features were used to characterize motion patterns, leading to an AUC of 0.94 on an infant dataset. Later, the same authors employed optical flow to estimate body motion across video frames to generate feature images, such as Log Mel-spectrogram, Mel Frequency Cepstral Coefficients, and Spectral Subband Centroid Frequency, which were combined by deep CNNs achieving an AUC value of 0.985. On an adult dataset, Zhao *et al.* [189] investigated learning deep facial expression features from the image and optical flow sequences using a 3D CNN, and obtained an average emotion recognition accuracy ranging from 0.56 to 0.76. Jung *et al.* [62] developed a joint network for facial expression recognition, which includes two networks of (1) the Deep Temporal Appearance Network (DTAN) and (2) the Deep Temporal Geometry Network (DTGN). The CNN-based DTAN is used to extract the temporal appearance feature, while the DTGN is employed for capturing geometric information of facial landmark movements. These two models are further combined to increase recognition performance. The previously discussed work on facial expression and behavioral analysis is still exploratory from nature, and needs to be further strengthened.

Summarizing the existing work, CNNs are good at extracting the spatial features in 2D for facial information analysis. However, video-based approaches for emotion recognition are still immature. The temporal-domain

information can assist the analysis of motion components. Importantly, especially for the application of infant monitoring, motion patterns are highly correlated to infant status. Therefore, adopting a 3D CNN seems to be a good choice, for considering and exploring both the spatial and temporal dimensions.

6.3 Methods for an end-to-end 3D CNN

For developing a suitable method, the following diagram for infant discomfort detection is proposed (shown in Figure 6.1), which compromises the following four steps.

(1) Pre-processing of the input infant videos to remove redundant background.
(2) Optical flow-based motion estimation is applied for estimating body movement between adjacent video frames.
(3) Combination of the informative features resulting from three RGB channels and two motion channels derived from Step (2). Therefore, in total five channels are used as input to the classification network.
(4) Network architecture by 3D CNN is implemented for the binary classification of comfort and discomfort, which embeds a motion-attention module (two motion channels). The 3D CNN extracts discriminative spatio-temporal representations of different infant status, and finally, a decision of comfort/discomfort is assigned to each video segment.

The details of the above four steps in the processing diagram are elaborated in the following subsections.

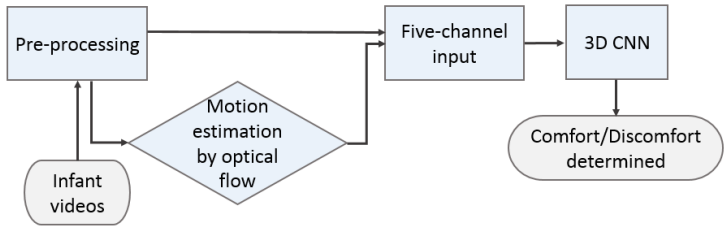


Figure 6.1: – *Processing architecture of the proposed discomfort detection system.*

6.3.1 Pre-processing

All video frames are first cropped to remove superfluous information in the image margins. The original image size of the recorded frame is 720×1280 pixels. The size is decreased to 501×751 pixels after removing pixels along each margin. The image is further down-sampled to 100×150 pixels by nearest-neighbor interpolation [106], since this is the required dimension for the input of the network.

For each down-sampled video frame, we apply Gaussian weighting to suppress the background content that is not relevant for analysis, for example, image-intensity changes caused by caregivers or parents moving around infants for care-handling. In our recording scenario, infants were always located at the central area of the video frames. The Gaussian weighting mask is thereby applied as a weighting function to highlight the central area and suppress the image boundaries. More specifically, the 2D Gaussian mean locations are directed to the center of the image, while the Gaussian standard deviations in the horizontal and vertical directions are empirically set to 40 and 80, respectively.

The original frame rate of the recorded videos is 30 frames per second (fps). To reduce the computation load and memory cost, each video is sub-sampled to half of the original frame rate (i.e. 15 fps) by only keeping the odd-indexed frames and skipping the even-indexed frames. The lower frame rate also increases the pixel movement between re-sampled adjacent frames and enables better motion analysis.

For each video sequence, we use a temporal sliding window with a fixed length of M frames (i.e. containing a time segment of $M/15$ seconds) and a sliding stride of N frames. Consequently, for each window, M frames are used as input clip for the classification network.

6.3.2 3D CNN model

CNNs facilitate the automated and deep learning of feature expressions directly from the input data (e.g. images and videos). However, a number of existing CNNs are only capable of handling 2D inputs due to the inherent network structure. The 2D CNNs are limited to spatial information only, while the temporal information across the video frames is not exploited. The recently proposed 3D CNNs extract features from both the spatial and temporal domains by performing a 3D convolution and 3D pooling. This approach is able to capture both the object appearance and contextual information and the motion information in a single optimization for feature extraction, so that the gen-

erated features resemble the spatial and temporal semantics [59]. 3D CNNs are a generalization of 2D CNNs. As compared to 2D CNNs, challenges for 3D CNNs are the larger memory footprint and higher dimensionality. These exacerbate the intensive computation cost for the network inference. To this end, we propose to use a 3D CNN model that not only considers the spatio-temporal information but also computational efficiency. We use the backbone of the PhysNet - 3D CNN model described in [179]. The input of our network is an image sequence, followed by several convolutional and pooling layers, as shown in Figure 6.2. In the figure, the RGB image is used as the input to illustrate the 3D CNN architecture. However, the color channels of the RGB images are later in this chapter extended to a 5-channel approach for feature extraction.

The first convolutional layer contains a number of kernels with a size of $1 \times 5 \times 5$ pixels, which only convolves the input data for the spatial information in a single frame. The following six convolutional layers with the kernel size of $3 \times 3 \times 3$ convolve both temporal and spatial dimensions. It has been shown that a deep net with small filter sizes like 3×3 outperforms a shallow net with larger filters [136], i.e. small receptive fields of 3×3 convolution kernels with deeper architectures yield better results. The previous consideration motivates our choice for the small size of the convolutional filter $3 \times 3 \times 3$.

The last convolutional layer is used for channel compression, and then space is compressed by max-pooling. Max-pooling is generally favored for classification tasks, since it leads to faster convergence and better generalization, while it also retains the most significant and translation-invariant information from the convolutional layers. Finally, average-pooling is used for compressing temporal information because the weight of each image in the time series is the same. It finally leads to a single confidence value that determines the label of comfort or discomfort.

After the second convolutional layer, the size of each frame is halved, and the number of channels is increased from 3 (RGB) to 64 (number of kernels). For each input video segment, the total number of frames is M and each frame includes three channels (RGB). For each frame, the pixel value in feature maps is the sum of squares from all the channels at the corresponding pixel location. Therefore, each video segment generates a group of feature maps.

The obtained outputs of the feature maps from the second convolutional layer reveal the motion information, which is from multiple adjacent frames in the original input. Figure 6.3 exemplifies the feature maps, which demonstrate that the network can highlight the motion area of the infants, especially the eye-/mouth/nasolabial furrow regions for characterizing their facial expression.

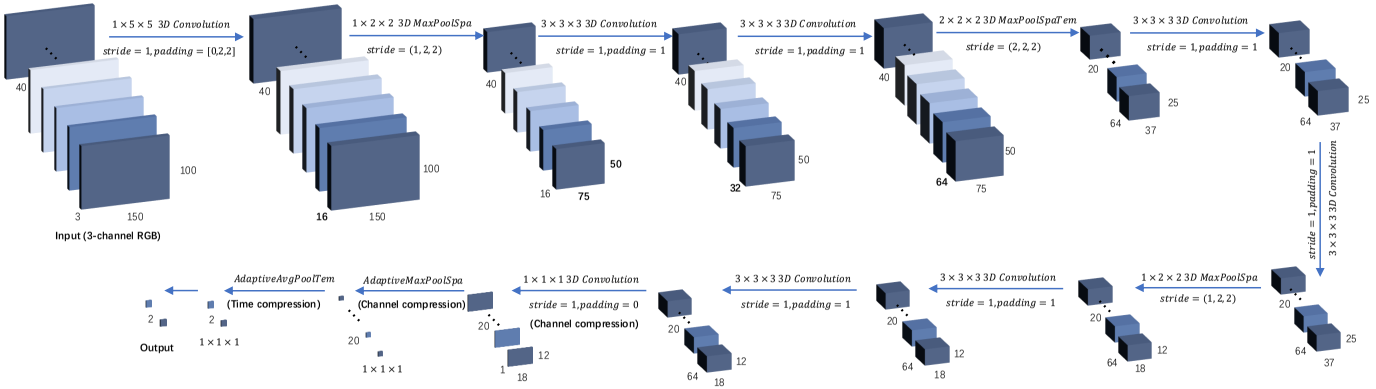


Figure 6.2: – *Architecture of the proposed 3D CNN, which is developed as one deep processing chain.*

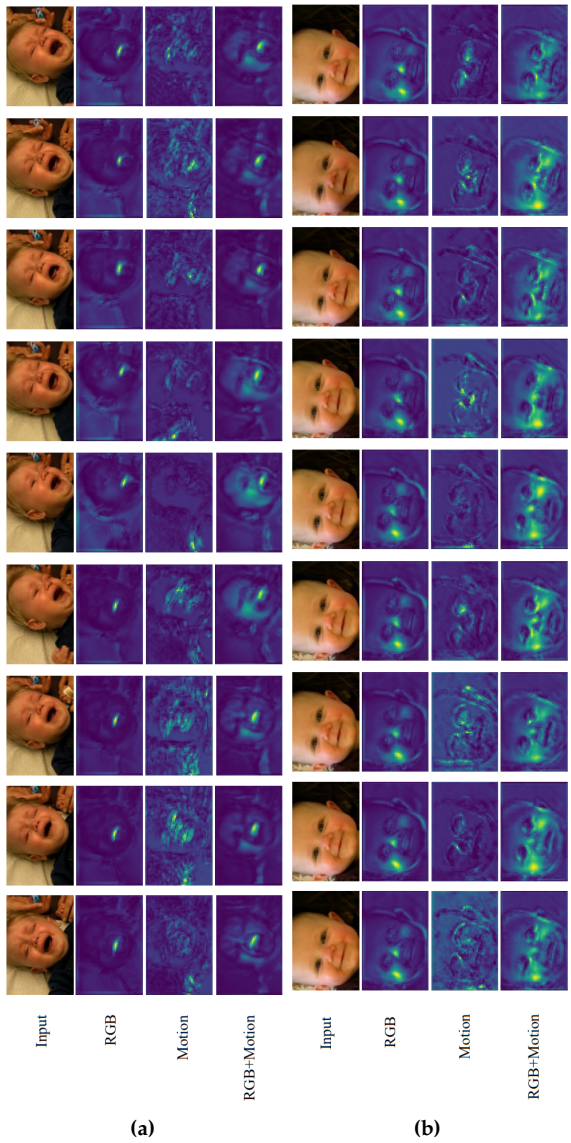


Figure 6.3: – *Visualization of the proposed 3D CNN model. Examples of feature maps for (a) comfortable and (b) discomfortable cases. Time increases within one row from left to right with a sampling interval of 0.2 s. The first row is the original sequence diagram, and the second shows the feature maps when training only on RGB. The third row is for only using optical flow as input for learning, and the bottom row shows the case when combining the five channels (RGB and optical flow). The horizontal axis is down-sampled in time to show larger differences among frames.*

Finally, we disclose some details about network functions and optimization. For learning, the Adam optimizer [69] is employed. Binary Cross-Entropy (BCE) with Logits Loss (BCEWLL) is used as the loss function. BCE is a cross-entropy suitable for binary classification, which is a special case of multi-class classification *softmax_cross_entropy*.

6.3.3 Multi-channel attention model

Motion estimation using optical flow

To estimate the motion of infants, we employ optical flow for calculating the motion vectors at the pixel level. Pixel-based motion vectors are calculated for each video frame between two consecutive frames, using the dense optical flow technique of Farnebäck *et al.* [33]. This technique uses quadratic polynomials to estimate the motion between two consecutive frames. Polynomial expansion is employed to approximate pixel intensities in the neighborhoods of the frames. A pyramid decomposition is used to handle large pixel motions, including distances larger than the neighborhood size. The tracking of motion begins at the lowest resolution level and continues until convergence, therefore effectively operating from coarse to fine level.

Between two consecutive video frames, the optical flow provides motion derivatives. In our study, we compute two motion matrices, which are the velocity magnitudes along the horizontal and vertical directions. These matrices are computed at the pixel level at input resolution. Visual examples are presented in the second and third rows of Figure 6.4 (a) and (b), respectively.

5-Channel attention model

The 3D CNN network needs to focus on the image areas where motion may occur. Therefore, we adopt an attention-based model to guide the network. The dense optical flow highlights the motion areas of infant bodies. After optical flow calculation, we supply the two motion matrices with the horizontal and vertical magnitudes as additional channels to the network input.

Figure 6.3 shows the feature maps when using different training inputs, which are 3-channel RGB, 2-channel optical flow, and 5-channel combined input. It is clear that the optical flow channels highlight the movement from the videos. By leveraging the information from the optical flow channels, the attention from the neural networks concentrates on the facial area in the images. The performance of the fused 5-channel mechanism is further investigated as a benchmark.

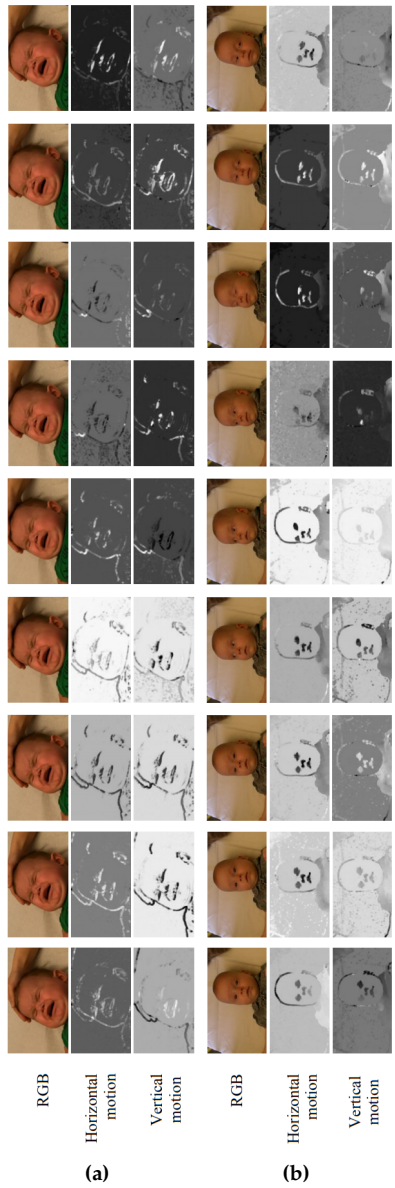


Figure 6.4: – *Examples of optical flow images obtained by Farnebäck's method for two infants with a sampling interval of 0.2 s, using a selection of one out of three frames. This is performed for enlarging the visual differences, as shown in (a) and (b). For each baby, the top row shows the original RGB frames, and the second and third rows are optical flow matrices for horizontal and vertical directions.*

6.4 Experimental Setup

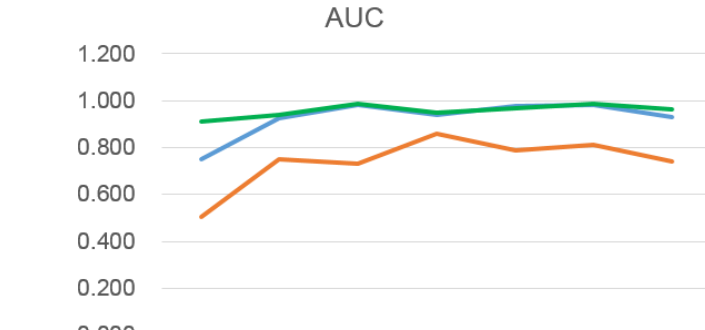
6.4.1 Clinical dataset

The captured videos for the conducted study were recorded at the Máxima Medical Center (MMC) in Veldhoven, The Netherlands, with a handheld high-definition camera (Xacti VPC-FH1BK). For each infant in the database, written consent was obtained from the parents. Videos of 24 infants were recorded. The infants' faces were filmed when experiencing various stress/pain moments, including clinical treatments of heel prick, placing an intravenous (IV) line, venipuncture, vaccination, post-operative pain, and the discomfort moments caused by a diaper change, feeling hungry or crying for attention. For 10 out of 24 infants, the relaxed comfort moments of resting or sleeping were also recorded. In conclusion, for these 10 infants both comfort and discomfort status were presented. For 4 infants only, the comfort moments were captured, while for the remaining 10 infants, only discomfort moments were recorded. Therefore, the video segments contain 1 to 2 types of emotion status per subject. The duration of each video varies from less than one minute to several minutes. The age of the 24 recorded infants ranges between 2 days and 13 months old. Three infants were born prematurely, and were aged under 37 weeks at the time of recording.

The resolution of the recorded video frames is 1920×1080 pixels, and the original frame rate is 30 fps. The videos were recorded under uncontrolled lighting conditions, where typically general office lighting conditions were applied. The label of comfort/discomfort for each frame is manually annotated, based on the consensus of two clinical experts. A total of 55 video segments from 24 infants were selected, where 19 segments present comfort, and the remaining 36 are discomfort cases. There are more discomfort samples in our dataset than comfort samples, since data collection performed by the clinicians in the hospital has focused on the discomfort moments of the infants.

6.4.2 Evaluation metrics

The classification was evaluated by measuring the classification accuracy and confusion matrices. The Receiver Operating Characteristic (ROC) curves are also plotted for visualizing the performance with the corresponding Area Under the ROC Curve (AUC) values.



	2D	1s	2s	3s	4s	5s	6s
RGB	0.748	0.927	0.981	0.939	0.975	0.980	0.929
Optical flow	0.505	0.750	0.730	0.859	0.790	0.813	0.741
5 channels	0.913	0.939	0.988	0.947	0.968	0.987	0.965

Figure 6.5: – *AUCs for using different clip lengths and training schemes.*

6.4.3 Learning protocol

A. Learning schemes

We have exploited different training schemes for investigating the networks, which are listed below.

• *3-Channel RGB*: Training is directly performed on the 3-channel RGB images of our infant dataset, as described in Section 6.3.2.

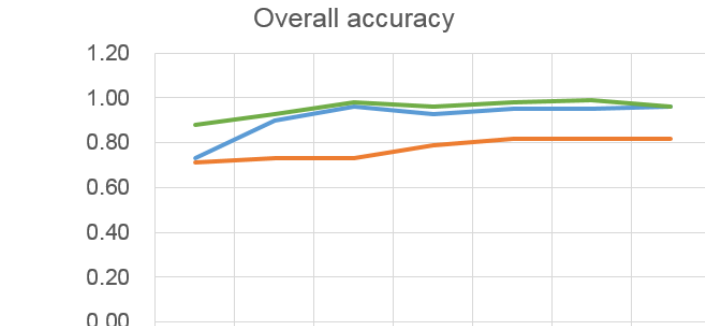
• *2-Channel motion*: Training is based on only 2-channel motion-derivative images estimated by dense optical flow.

• *5-Channel RGB and motion*: The hybrid 5-channel input based on spatio-temporal features has been already described in Section 6.3.3.

For training, we empirically set the number of epochs (m) to 800.

B. Window-size tuning

We have also investigated the effect of the temporal sliding-window length on the performance of the 3D CNN model, in terms of classification AUC, accuracy, and execution time. The window length is tuned from one frame (i.e. the 2D case) to 15 frames (1 s), 30 frames (2 s), ..., 90 frames (6 s). The experiments have been carried out on a GeForce RTX-2080 GPU and an Xeon E5-2680 V2 CPU.



	2D	1s	2s	3s	4s	5s	6s
RGB	0.730	0.900	0.960	0.930	0.950	0.950	0.960
Optical flow	0.710	0.730	0.730	0.790	0.820	0.820	0.820
5 channels	0.880	0.930	0.980	0.960	0.980	0.990	0.960

Figure 6.6: – *Accuracy for using different clip lengths and training schemes.*

6.5 Results

A. Evaluation metrics

• Learning scheme *3-channel RGB*: Figure 6.7 shows the normalized confusion matrix when the training is directly performed on 3-channel RGB images. The obtained overall labeling accuracy is 0.96. The detection accuracy of comfort and discomfort is 0.94 and 0.98, respectively. Figure 6.8 represents the corresponding ROC with the AUC value of 0.98.

•Learning scheme *2-channel motion*: The measured AUC is 0.73 when training the model only on the 2-channel optical flow images.

•Learning scheme *5-channel RGB and motion*: The model achieves the highest AUC of 0.99 when training on the input of five channels (See Figure 6.10). The normalized confusion matrix of fine-tuning using 5 channels is shown in Figure 6.9. The overall accuracy delivers robust and consistent information on reaching the best value of 0.98.

B. Window-size tuning

The results mentioned above are all based on the window length of 15 frames (1 s) and a step size (stride) of 5 frames (around 333 ms). We further explore the influence of changing the window size. Figure 6.5 presents the AUCs for three learning schemes with different window lengths. Figure 6.6 shows the accuracy values during the window-size tuning procedure.

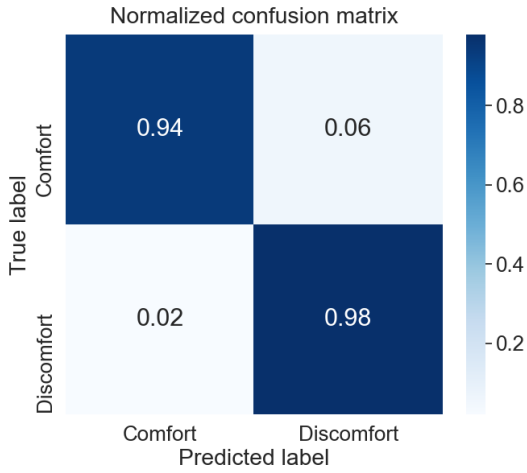


Figure 6.7: – *Normalized confusion matrix when directly training on 3-channel RGB images.*

C. Execution time

Depending on tuning the duration of video clips (window size), the obtained corresponding execution times for testing each video clip are summarized in Table 6.1.

Training method	2D (s)	1s (s)	2s (s)	3s (s)	4s (s)	5s (s)	6s (s)
3-channel RGB	0.064	0.074	0.086	0.097	0.108	0.111	0.128
2-channel motion	0.035	0.073	0.081	0.091	0.099	0.104	0.111
5-ch. RGB+motion	0.064	0.078	0.097	0.109	0.127	0.139	0.159

Table 6.1: – *Execution times of the three training methods for using different lengths of video clips (window size expressed in seconds).*

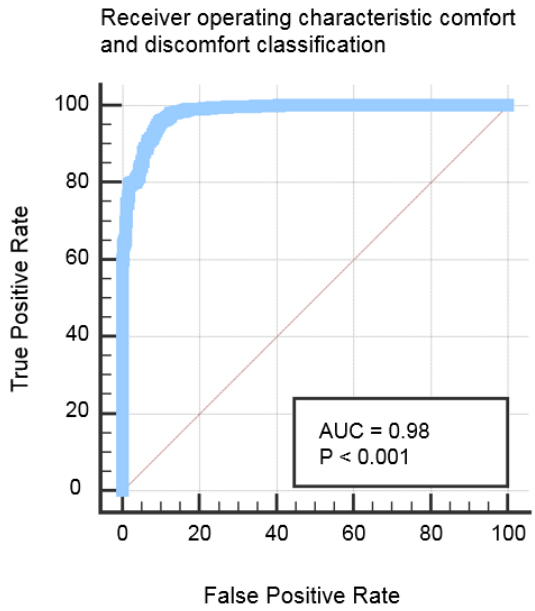


Figure 6.8: – *ROC of the proposed method when directly training on 3-channel RGB images.*

6.6 Discussion and Conclusions

6.6.1 Discussion

Execution time: The best performance is achieved when the sliding-window length is set to 2 seconds (30 frames) and the step size is set to one-third of a second (5 frames). This means that (1) the latency between the start of the system and the first measurement can be as short as 2 seconds, and (2) the monitored infant status can be refreshed every one-third of a second incremented with the required execution time for making a decision on the present clip. The testing execution times shown in Table 6.1 indicate the required processing time for unseen video clips, which is about 0.1 seconds. The experiments are carried out on a GeForce RTX-2080 GPU and an Xeon E5-2680 V2 CPU. The obtained execution times indicate that the infant discomfort detection can be implemented as a real-time clinical application. Compared to 2-channel or 3-channel schemes, the 5-channel scheme requires some extra computing time

(about 0.01 – 0.03 seconds), but this additional time is negligible due to the parallel structure of the proposed solution.

Motion-based attention: When comparing the performance of different training schemes, the AUC and overall accuracy are the highest when applying the learning scheme of 5-channel RGB and motion, which confirms the effectiveness of the 5-channel attention-based network. The two extra channels of optical flow images provide the regions/boundaries with strong motion information, which serve as a beacon that gears the network to focus on the movement. The motion patterns of infants are clinically important for assessing comfort or discomfort. From the feature maps in Figure 6.3, the statement is confirmed that by incorporating the information from the optical flow channels, the attention from the neural networks is dominantly focused on the facial area of the infant and its related motion patterns. However, this discussion could be broadened when the body of the infant would be recorded also. It is known that when infants start from discomfort, their bodies also present different motion patterns compared to comfort moments. Further studies will be verified in the future by collecting more video samples capturing both face and body parts.

Window size tuning: When the sliding-window length is set to one frame, this training/testing procedure becomes 2D processing. Evidently, its performance is significantly reduced, as compared to the 3D processing. For all three training schemes, the AUC and accuracy values improve along with the increasing window size, which appear to be saturating after 2 seconds. This saturation may be caused by the limited valid information from the samples, or constraints by the available training sample size, which should be further investigated by collecting more infant data in the future. The reason for the saturated performance could also be that the discomfort information (facial expression or body motion) involves mostly spontaneous/abrupt changes that typically happen in a very short time interval, i.e. high-frequency temporal information to be captured in a short time window. A long time window may include and highlight low-frequency information (or slow changes like motion drift or handheld camera motion), which are less relevant for comfort/discomfort classification.

Complexity of 3D CNN: Regarding the complexity of the proposed 3D CNN, the total number of the parameters is 514,657, yielding a compact computational model. The execution time for making a decision on an unseen clip is 0.097 seconds using the sliding window of 2 seconds, which results in the best AUC of 0.99. As a conclusion, the complexity of the 3D CNN is lower than expected, while it offers a very high performance.

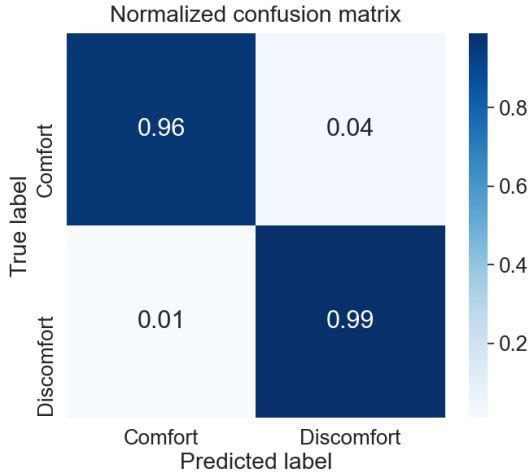


Figure 6.9: – *Normalized confusion matrix for 5-channel training.*

6.6.2 Conclusions

This chapter has proposed a video monitoring system that provides continuous and contactless assessment of discomfort for infants. The system is validated by real clinical infant data from a hospital with expert annotations. In this study, we have investigated the benefit of using optical flow measurement to draw the attention of 3D CNNs. The system aims to alert caregivers/clinicians immediately when infants start suffering from discomfort. The proposed system can monitor infant status continuously by classifying the video frames into either comfort or discomfort, which fills the gaps of the current intermittent manual observation and the human efforts. Moreover, the proposed method also has the potential to be implemented as an infant-care tool for family use on a longer term or remote monitoring critical infants on a professional basis. The system alarm is triggered by the detection of discomfort status, which will notify clinical staff for timely and appropriate treatment. Thus, the system serves to prevent fatal events and eventually improves the early development of infants. Summarizing, this chapter has contributed as follows. To the best of our knowledge, this is the first contribution with a 5-channel spatial-temporal input of the infant behavior to the learning model for discomfort detection. Second, the design of the novel 3D CNN network has a fast processing speed

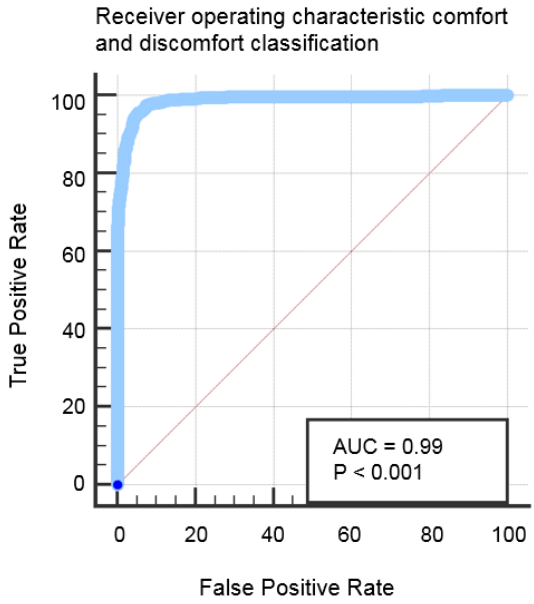


Figure 6.10: – *ROC curve of 5-channel training.*

and high performance, thereby showing that the system suits clinical practice. The system can achieve continuous reliable operation for clinicians.

The proposed system forms a neat end-to-end video processing solution without requiring the conventional front-end steps of face detection and tracking. In general practice, the partial/full-face occlusion is likely to happen, especially during infant physical care. Our method is better suited to handle these challenging moments, as compared to conventional face detection/tracking-based methods.

In the future, the proposed model can be improved to identify discomfort grades by changing the two-class output to multiple classes, or even a regression layer to predict the significance of discomfort. Future work may also include extending the 5-channel input to higher dimensions by fusing information from different sensor modalities, such as the depth information from a 3D sensor (e.g., time-of-flight camera). In addition to the fusion of color-intensity signals and motion signals, it can be beneficial to consider

fusing the contextual information and physiological information for joint classification, since physiological variables (e.g., heart rate, heart-rate variability, respiration rate) can be also measured from the videos.

Up to this point, the research work has deeply investigated the computer-aided diagnosis to detect infant discomfort moments. The next chapter attempts to further exploit the information inside captured videos and develop a broader status report on the well-being of the infant. This approach is based on expanding the quantitative analysis using videos for extracting information on the respiration of the infant. Again, the research aims at a contactless solution for vital sign measurement of the infant as an add-on to comfort/discomfort detection.

Chapter 7

Quantitative measurement - Respiration monitoring for premature neonates in NICU

7.1 Introduction

The previous chapters have focused on the characteristic analysis of infant facial expression and/or motion patterns for classifying infant comfort and discomfort status. This chapter is aiming at further investigating the feasibility of using videos for quantitative physiological signal extraction, which is an extension to the direct video-based comfort/discomfort classification. It is known that specific parameters from a person can be remotely measured from video signals.

Vital signs and the related parameters, such as heart rate, blood pressure, respiratory rate, and body temperature, are physical parameters that can be measured and used to assess physiological state and functioning of a person. Monitoring of vital parameters is a crucial topic in neonatal daily care. Premature infants have an immature respiratory control that predisposes them to apnea/periodic breathing, haemoglobin oxygen desaturation, and bradycardia [22, 125]. Apnea is defined as the status of cessation of respiratory airflow, whereas periodic breathing is characterized by groups of respiratory movements, which are interrupted by small intervals of apnea [109]. Continuous

monitoring of respiration for premature infants is critical to detect abnormalities in breathing and help in developing early treatments to prevent significant hypoxia and central depression from apnea. A long-term continuous monitoring approach of respiration may be used to also assess the sleep stage, which can vary with different clinical conditions [111, 112]. The condition of respiration and its measurement is a key research topic for this chapter.

The existing technical methods for monitoring respiration include nasal thermocouples, spirometers, transthoracic inductance, respiratory-effort belt transducer, piezoelectric transducer, optical sensor (pulse oximetry), strain gauge, impedance plethysmography, and electrocardiogram (ECG). Currently, the respiration of premature infants in a neonatal intensive care unit (NICU) is monitored by bedside monitors. ECG is considered as the standard reference measurement for respiration, since ECG can provide stable and robust monitoring for a NICU. However, the pressure of the contact sensor may also change the local skin perfusion, which can yield not a true measurement as compared to the non-contact sensor. The electrodes may exert pressure to the skin, yielding to tissue compression and vascular insufficiency. As a consequence, applying the ECG electrodes on infant skin for a long time increases the risk of trauma and infections. Removing the adhesives from the immature skin, as part of regular care, can damage the immature skin of preterm infants, as well as cause stress and pain [3, 90]. This technique is impractical for home care, since the sensors have to be placed by skilled caregivers, and wearing the sensors also causes inconvenience for everyday life [8, 92].

A contactless respiration monitoring system is a good alternative to improve infant comfort and safety and when combined with video observation, it also has the potential for monitoring at home. Few works to date have investigated video-based contactless methods for monitoring respiration. This chapter aims to develop and evaluate a contactless respiration measurement method based on video monitoring, which addresses the following challenges.

- *Movement estimation:* The respiration calculation is based on movement estimation. The respiratory motion of infants can be subtle and small in amplitude, which can complicate a video-based measurement. Both conventional and state-of-the-art methods should be investigated and explored for extracting movement signals. Specifically, the performance of deep learning-based methods need to be compared with optical flow techniques for motion measurement.

- *Movement separation:* Since the movement signal from the videos contains not only respiration but also other types of movements. These types

of movement can interfere with the measurement of respiration and act as a noisy component. From the respiratory motion, the respiration rate should be computed and should be only based on respiration-related movement. Therefore, different components of estimated motion signals need to be well analyzed and separated.

- *Respiration signal conversion:* Generic signal processing algorithms need to be employed for processing the respiration signal and enabling reliable detection of the dominant features of the respiration. For example, the peaks of the derived respiration signal need to be detected, so that the respiration rate can be obtained accordingly. The applied signal processing algorithms rely on hyper-parameter tuning, such as the length of the sliding window. Therefore, such parameters should be optimized for performance.

Taking into account the requirements and aspects mentioned above, this chapter develops an approach for non-contact respiration monitoring, where the proposed method is validated on clinical data acquired in a NICU. Two flow-estimation methods are employed to estimate pixel-based motion vectors between adjacent video frames, namely the conventional optical flow method [9] and Deep Flow [171]. The spatial statistics of each derived flow matrix is further analyzed, by applying robust principal component analysis (PCA) [107] to describe respiration information.

The remainder of this chapter is as follows. Section 7.2 discusses the existing related work. Section 7.3 elaborates on the proposed methods. Section 7.4 presents the experimental results with discussions. Finally, Section 7.5 concludes the chapter.

7.2 Related work

A. Work on adults for measuring physiological parameters

In the applications of adults, Deng et al. [27] presented a design and implementation of a novel sleep monitoring system that simultaneously analyzed respiration, head posture, and body posture for adults. For respiration, the region of breathing movement was automatically determined and estimated the intensity, yielding a waveform indicating respiratory rhythms with an accuracy of 96 % in recognizing abnormal breathing. However, the accuracy of

the respiration rate was not provided. Chazal et al. [26] applied a contactless biomotion sensor of Doppler radar for respiration monitoring. Respiratory frequency and magnitude were used for the classification of determining sleep/wake states, which achieved an accuracy of 69 % for the wake and 88 % for sleep state. Gupta et al. [44] estimated heart rate (HR) accurately, using face videos acquired from a low-cost camera for adults. The face video consisting of frontal, profile, or multiple faces, was divided into multiple overlapping fragments to determine HR estimates. The HR estimates were fused using quality-based fusion, which aimed to minimize illumination and face deformations. Prathosh et al. [110] proposed a general framework for estimating a periodic signal, which was applied to derive a computationally inexpensive method for estimating respiratory patterns using two-dimensional cameras, which did not critically depend on the region of interest. Specifically, the patterns were estimated by imaging changes in the reflected light caused by respiration-induced motion. Estimation of the pattern was cast as a blind deconvolution problem and was solved through a method comprising subspace projection and statistical aggregation.

B. Physiological measurements for infants

In terms of the applications for infants, Werth et al. reviewed the unobtrusive measurements for indicating sleep state in preterm infants [172]. Abbas et al. [1] attempted to detect respiration rate of neonates on a real-time basis using infrared thermography. They analyzed the anterior naris (nostrils) temperature profile associated with the inspiration and expiration phases. However, the region of interest (ROI) was assumed to be fixed after initialization. Moreover, the method is not practical for NICU infants, since the faces of premature infants in a NICU are often occluded by feeding tubes and/or breathing masks. In practice, the temperature inside an incubator is continuously monitored and adjusted by caregivers in order to help infants maintain their body temperature in a normal range. However, the controlled environmental temperature might affect the accuracy of the thermography-based respiration monitoring method. Koolen et al. [72] extracted the respiration rate from video data included in polysomnography. They used Eulerian video magnification (EVM) to amplify the respiration movements, which was followed by optical flow to estimate the respiration motion and therefore, obtained a respiration signal. Independent component analysis and principal component analysis were applied to improve signal quality. Finally, the results showed a detection accuracy of 94.1 % for sleeping-stage patients. Antognoli et

al. [6] applied a digital webcam (WeC) and an EVM algorithm to measure HR and respiration rate (RR). The accumulated RGB values of a manually selected ROI were calculated as a single signal, from which the power spectral density was further estimated and used for peak extraction. The evaluation based on data of seven patients yielded a root mean-squared error (RMSE) of 12.2 for the HR and 7.6 for the RR. However, the limitation of a pulse-based respiration extraction is that it extracts the respiratory modulation in blood volume changes, which is both subject-dependent (different respiratory efforts) and measurement-location dependent. The modulation effect varies in different body parts. Moreover, the peak selection from spectrograms was based on common knowledge of normal HR and RR frequency ranges, which is not suitable for infants under clinical conditions of different diseases.

Summarizing the existing work, various sensing techniques including biomotion sensing of Doppler radar, video camera and infrared thermography are leveraged. Automated solutions on remote sensing of respiration or other contactless physiological measures have been developed for adults and infants. In our study, the target population are premature infants in a NICU. We select a video camera instead of other sensors because of its large availability and economical usage, and it is contactless. Therefore, we aim to apply motion extraction and movement decomposition to estimate the respiration signal. Then, this respiration-movement analysis can be combined with overall video observation, so that the video camera serves two purposes simultaneously.

7.3 Methods for respiration estimation

7.3.1 Material

The study was conducted with videos recorded at the Máxima Medical Center in Veldhoven, The Netherlands, by a fixed-position high-definition camera (Camera model IDS uEye monochrome) filming the infant's entire body in the direction of the foot to the head. Figure 7.1 shows an example of a captured video frame. We decided on the recording position of the camera by considering two aspects: (1) the camera position should cause little or no interruption to daily routine care, and (2) the position should offer a good viewpoint for observing vertical movement of the infant chest, where the movement is assumed to be with maximum respiratory motion energy in the vertical direction.

For all infant recordings, written consent was obtained from the parents. The resolution of each video frame was 736 × 480 pixels, while the frame rate

was 8 fps. The videos were recorded under uncontrolled, regular hospital lighting conditions. Parallel with the video capturing, standard chest impedance (ChI) signals were recorded simultaneously as reference standards.

The dataset contains five infants with an average gestational age of 29.6 ± 2.8 weeks (range 27^{+0}–33^{+6} weeks), an average postnatal age of 1.2 ± 0.6 weeks (range 0^{+4}–2^{+1} weeks), and an average weight of 1,555 ± 682.4 g (range 755–2,410 g).

Figure 7.1: – *Example of an acquired video frame from the direction of the foot to the infant head to support vertical motion measurement.*

7.3.2 Motion-based calculation of respiration

From the captured videos, the motion-based respiration is measured at different points in the image to come to a reliable measurement. The different points enable us to create a feature vector of motion values, which will give important information for the machine learning system. For every video frame, the feature values are jointly combined into a matrix, which will be referred to as the motion matrix. The flowchart of the proposed system is shown in Figure 7.2, where motion-matrix estimation is an essential step in the pipeline. We estimate the pattern of apparent motion resulting from the respiration of infants in videos. The apparent motion [1] visible in the videos is not only the real motion from the respiration because humans have various types of motion and the

[1] The apparent motion is the motion visible in the video frames, which can be different from the real motion of the infant chest.

camera is only covering these from a single point of view. Therefore, we consider the measured motion only as an estimate of the real respiratory motion. The motion-matrix estimation can be defined as the distribution of apparent velocities of movement in successive images.

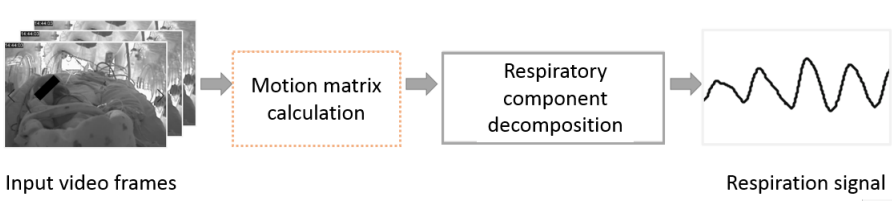


Figure 7.2: – *Flowchart of the proposed video-based respiration monitoring system.*

Two flow estimation methods are employed to estimate pixel motion vectors between adjacent video frames. First, the conventional optical flow method [9] is utilized and evaluated. However, the optical flow is more sensitive to high gradient texture, whereas in our case the infant chest is either bare or covered by a blanket. Thus, in both cases, the chest area, which mostly indicates the respiratory motion, lacks gradient texture. Deep Flow [171] is sensitive to the entire part of moving objects paying less attention to texture information. Therefore, improved performance is expected by using a flow estimation method based on deep learning.

For both methods, the derived motion vectors contain respiration information, induced by the motion of the abdominal wall and chest wall. We have only considered the vertical motion vector, since the respiration-related motion is mainly in the vertical direction in the captured videos. For infants under one year of age, it is recommended by the American Academy of Pediatrics (AAP) that they should be placed to sleep on their backs every time being laid down to sleep. This lowers the risk of the sudden infant death syndrome (SIDS).

The details of both methods are now briefly discussed.

Conventional Optical Flow

The optical flow computation is specified by:

$$\frac{\partial \mathbf{I}}{\partial x} V_x + \frac{\partial \mathbf{I}}{\partial y} V_y + \frac{\partial \mathbf{I}}{\partial t} = 0, \tag{7.1}$$

where **I** is the intensity matrix of a frame in a video, which is locally differentiated, and V_x are V_y are the actual optical flow parameters describing the motion velocity. We have deployed the classic dense optical flow algorithm proposed by Barron et al. [9], where second-order differential equations based on the *Hessian* matrix are used to constrain two-dimensional (2D) velocity. The Barron method creates flow fields with 100% density. However, the conventional optical flow method is limited at estimating motion in poorly textured areas, which lack gradient variation.

Deep Flow

The above conventional optical flow approach only computes image matching at a single scale. However, for complicated situations with complex motions, the conventional approach may not be able to effectively capture the interaction or dependency relationships. Brox and Malik [17] proposed to add the addition of a descriptor matching term in the variational approach, which allows better handling of large displacements. The matching provides guidance in using correspondences from sparse descriptor matching.

In our study, we have applied the method from Weinzaepfel et al. [171]. A descriptor matching algorithm was incorporated and implemented in a CNN, which is based on deep convolutions with six layers, interleaving convolutions and max-pooling. In the proposed framework, dense sampling is applied to efficiently retrieve quasi-dense correspondences, while incorporating a smoothing effect on the descriptors matches. Figure 7.3 shows the framework for Deep Flow estimation.

The local maxima in the response maps correspond to good matches of corresponding local image patches. To obtain dense correspondences between any matched patches (i.e. at local maxima), it suffices to recover the path of response values that generated this maximum. In contrast to most algorithms for optical flow, the Deep Flow method works in a bottom-up fashion. The algorithm starts at a fine level, and then moves up to coarser levels (larger patches), which are constructed as an aggregation of responses of smaller patches.

7.3.3 Respiratory description

The results of flow calculation are captured in the rows of the derived motion matrix **M** (of size $N \times W$, where N denotes the number of pixels in a video frame

and W is the total number of frames in a video[2]), which contain the motion derivatives that represent the velocity magnitudes of the pixel trajectories in the vertical direction.

The spatial statistics of the flow matrix were further analyzed by applying robust PCA [107]. PCA includes the Eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The decomposition task projects the original data onto an orthogonal subspace, where each direction is mutually decorrelated and the most informative data information becomes available in the first several principal components. In our study, instead of considering the originally obtained matrix, the first Eigenvalue of the covariance matrix of flows is analyzed, since the first Eigenvalue dominantly represents the major motion component. Besides, it is beneficial to reduce residual motion noise that has lower energy than the respiratory motion in each video.

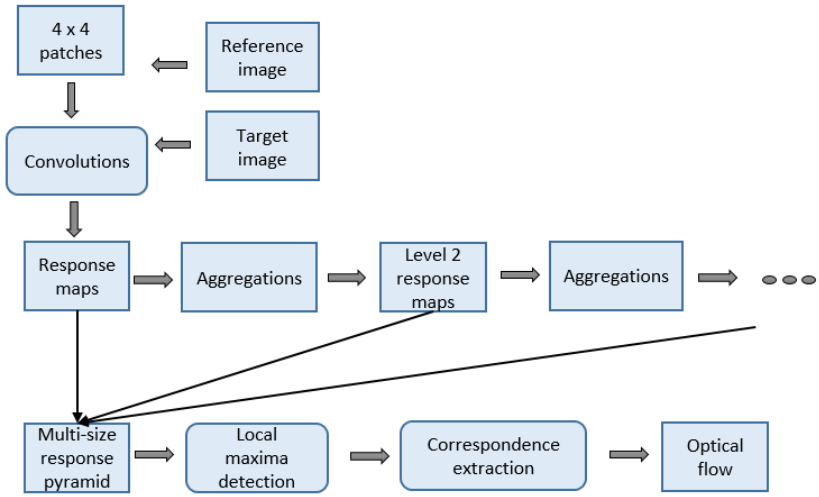


Figure 7.3: – *Diagram of the workflow and CNN-based processing steps in Deep Flow.*

The motion matrix $\mathbf{M}$ is masked by sub-spatiotemporal regions, $\mathbf{m}_i$, where each $\mathbf{m}_i$ is a local mask that stores W consecutive squared blocks from all images in a video. For each region $\mathbf{m}_i$, we can generate the Eigenvectors that

[2]In practice, the length of the video can be partitioned into segments or a sliding window can be used to limit the memory complexity.

satisfy the following general condition, specified by

$$\mathbf{m}_i \cdot \mathbf{D}_i = \lambda \cdot \mathbf{D}_i \quad s.t. \quad \mathrm{Det}(\mathbf{m}_i - \lambda_i \cdot \mathbf{I}) = 0, \tag{7.2}$$

where $\mathrm{Det}(\cdot)$ denotes the matrix determinant, $\mathbf{I}$ represents the identity matrix, and $\mathbf{D}_i$ and λ_i are Eigenvectors and Eigenvalues, respectively.

Since it is difficult to a-priori choose between the two methods, we have evaluated both of them.

7.3.4 Evaluation of optical flow and Deep Flow

In this study, we have employed a sliding window of 120 s with a step size of a unity second to estimate the respiration rate. The respiration rate is calculated by first averaging the time intervals between breathing peaks, followed by converting those numbers to a frequency value, expressed in the unit of breaths per minute (bpm).

It is well known that the respiration signal can be estimated from the chest impedance (ChI) signals. To evaluate the performance of respiration estimation using our video-based algorithms, we have computed the cross-correlation coefficients and the RMSE as measures to compare the respiration rates of the reference breathing signal from the ChI and the extracted respiration signals. The cross-correlation coefficient calculation is initially based on a zero-mean process. Thus, this calculation is only for comparing the respiration rate variance of our estimation and the reference standard. In addition, correlation plots and Bland–Altman plots [13, 71] are also created.

7.4 Experimental results and Discussion

Figure 7.4 shows a visual comparison of the reference breathing signal from the ChI and two extracted respiration signals by the proposed optical flow- and Deep Flow-based methods. The three re-scaled 1D signals from the ChI, optical flow- and Deep Flow-based motion estimations are depicted in Figure 7.5.

Table 7.1 summarizes the estimation of the respiration rate using two different optical flow methods. Table 7.2 shows the root mean-squared errors (RMSE) and cross-correlation (CC) coefficients of the reference breathing signal from the ChI, compared to our optical flow- and Deep Flow-based results. The results obtained from Deep Flow are more accurate than those of the conventional optical flow approach. The average of all cross-correlation coefficients is computed over time for all videos, while the overall average is 0.74

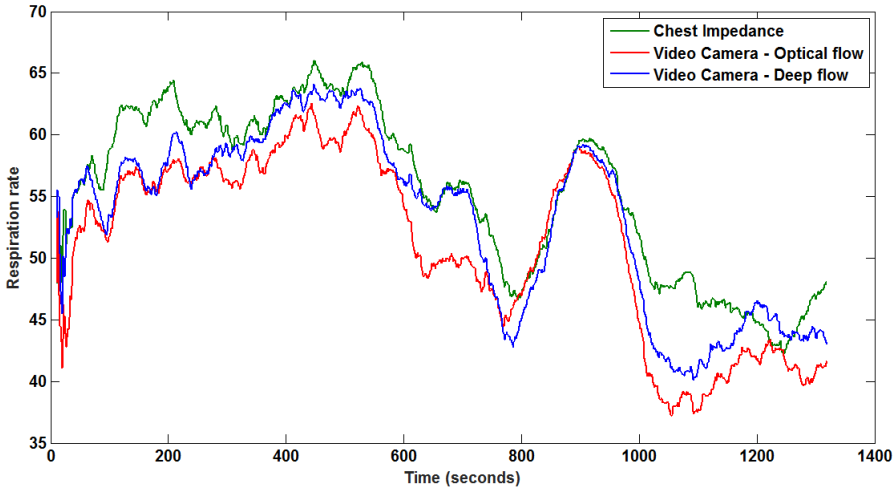
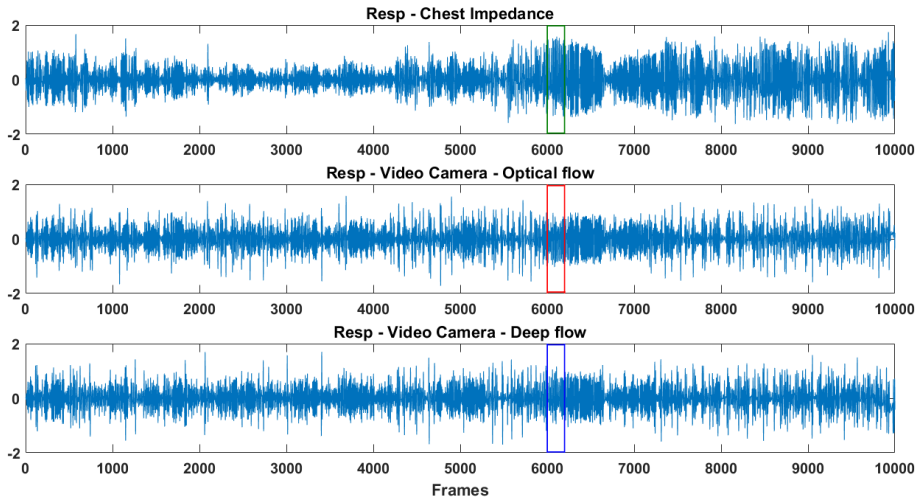


Figure 7.4: – *Monitored respiration from the ChI, and corresponding extracted respiration signals from a segment, based on optical flow and Deep Flow.*

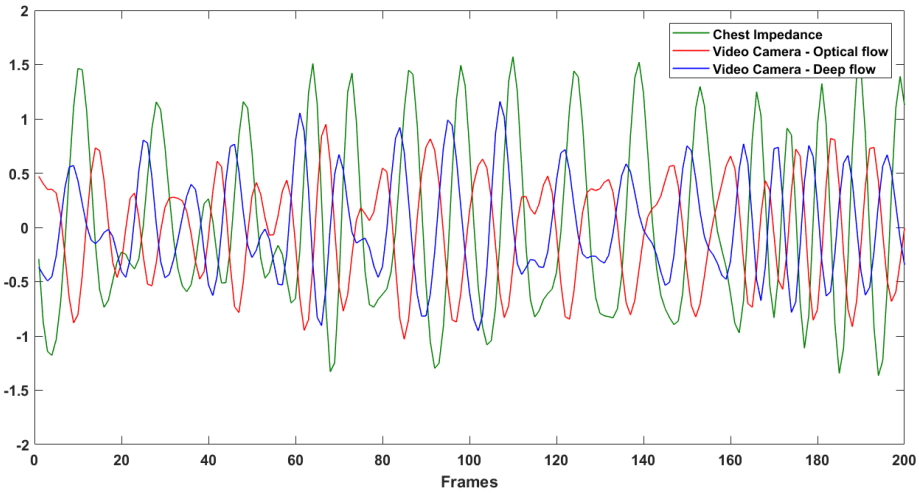
and the overall average RMSE is 4.55, using the proposed deep learning-based method. Despite the limited number of measurements, these preliminary results provide a promising result for further investigation of the neonatal video-based respiration monitoring method. The low value of the overall RMSE for our deep learning-based method (4.55) shows the feasibility to apply our automated respiration for clinical use.

Patient	**Duration**	**Mean ± std**		
ID	**(h:m:s)**	**Reference (ChI)**	**Optical Flow**	**Deep Flow**
1	00:50:13	51.12 ± 6.00	47.13 ± 4.96	48.67 ± 4.69
2	00:22:00	55.41 ± 6.73	51.18 ± 7.35	53.23 ± 7.07
3	00:26:54	48.78 ± 3.58	44.12 ± 2.91	46.75 ± 2.80
4	00:09:22	52.36 ± 5.25	48.75 ± 3.17	50.05 ± 3.72
5	00:06:15	61.61 ± 3.80	51.56 ± 4.29	55.58 ± 3.28

Table 7.1: – *Measurement of the respiration rate for all the videos. Duration of each video, and measured mean and standard deviations of reference and our optical flow-based and Deep Flow-based methods.*



(a)



(b)

Figure 7.5: – *Data patterns for chest impedance and video-based respiration observation: (**a**) re-scaled 1D signals; and (**b**) zoom in of each interval indicated in (**a**).*

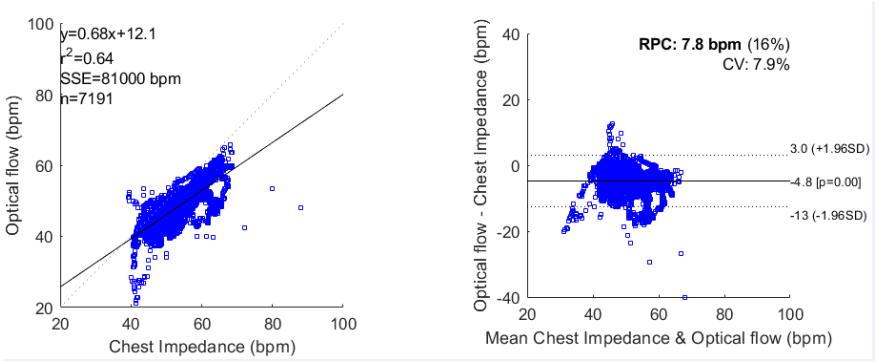
Patient ID	RMSE		CC Coefficients	
	Optical Flow	Deep Flow	Optical Flow	Deep Flow
1	5.09	4.32	0.85	0.82
2	4.94	3.10	0.94	0.95
3	3.86	3.51	0.73	0.75
4	5.75	5.11	0.47	0.52
5	10.85	6.71	0.49	0.66
Average	6.10	4.55	0.70	0.74

Table 7.2: – *Root mean-squared errors (RMSE) and cross-correlation (CC) coefficients of the reference breathing signal from the ChI compared to our optical flow-based and Deep Flow-based results.*

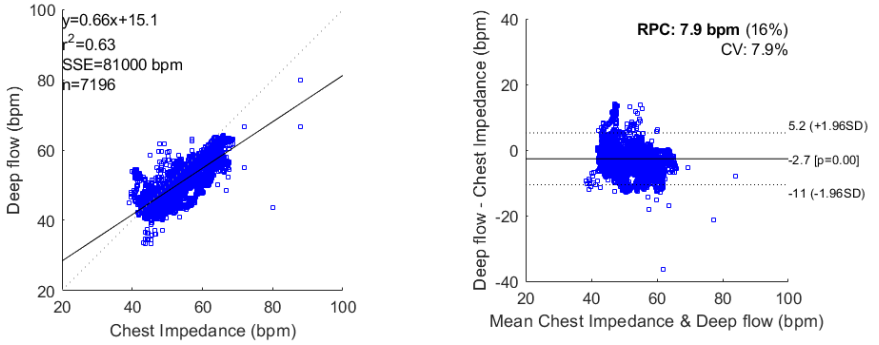
Discussion on systematic errors: From both the correlation and Bland–Altman analyses in Figure 7.6(a) and (b), the Deep Flow-based method produces a lower error than the optical flow-based method, especially when breathing rates are less than 50 bpm. The absolute values of the overall mean errors reduce from 4.8 to 2.7 bpm in the optical flow and Deep Flow cases, respectively. Except for Patient 5, the obtained error in the respiration rate is limited to 5 % or less. The Deep Flow-based method gives a consistently lower error (RMSE) than the optical flow-based method.

We have observed that a systematic residual error occurs, which results from our automated processing pipeline. The proposed automated method consistently underestimates the respiration rate. This phenomenon occurs because our motion filter fails to remove motion caused by interruptions in the video. For example, nurse care-handling of the infants can give occlusions to the visibility of the infant respiratory motion, which results in lack of motion cycles being measured by the camera, leading to a lower respiration frequency. Considering a possible reduction of the systematic residual errors, future work would involve filtering methods, which are typically unsupervised predefined filters. To replace such unsupervised filters for extracting the noise from respiration signals, it seems attractive to investigate a supervised approach.

Future direction on optimizing accuracy: We have compared the effect of two different flow estimation methods on the final respiration calculation. The results show that the Deep Flow approach is both sensitive to homogeneous regions and the boundary area, whereas the conventional approach is more sensitive to the boundary area. This advantage of the Deep Flow-based method



(a) *Optical flow*



(b) *Deep Flow*

Figure 7.6: – *Correlation and Bland–Altman plots comparing respiration rate measurements derived from: (**a**) the chest impedance (ChI) and optical flow-based method; and (**b**) the ChI and Deep Flow-based method. The correlation plots contain the linear regression equation, correlation coefficient (r^2), sum of squared error (SSE), and number of points. The Bland–Altman plots contain the reproducibility coefficient ($RPC = 1.96 \times \sigma$), and also show the coefficient of variation (CV = the standard deviation (σ) as a percentage of the mean), limits of agreement ($LOA = \pm 1.96 \times \sigma$), and the bias offset of the measures.*

improves the whole processing pipeline by increasing both the accuracy and robustness.

We consider that the accuracy of the proposed algorithm for extracting a respiration signal may be affected by the captured image resolution and image sensor noise. If the resolution is rather low, the movement information from different regions can be blended within one pixel, which is not ideal for accurate motion extraction. If the image sensor noise is too high such that the sensor noise components in pixel values dominate the pixel changes induced by respiratory motion, the respiration measurement will be polluted and inaccurate. The quantitative analysis regarding this perspective is a complicated matter and is seen as future work.

Future directions on disease-related analysis: At present, the research work is carried out as a feasibility study. The focus of this study is the installation, adaptation, and validation of camera-based monitoring technology in the NICU setting. The performance on preterm infants related to specific diseases has not been investigated. In the future, we will further validate the proposed algorithm on infants having different health situations. One ongoing study that is performed at the same hospital is the relation between discomfort and gastroesophageal reflux disease (GERD) [77].

Required adaptations for clinical deployment: Our recordings were taken from real clinical practice without interfering with the clinical workflow. Our algorithm works when the respiratory motion can be observed by the camera, even as subtle movement, i.e., infants can be either naked or covered by a blanket (as long as the infant body has contact with the blanket such that the movement information from the thorax-abdomen can still be derived).

Our algorithm relies on the intensity of video frames for motion extraction (no color or chrominance information is used). Therefore, for the nighttime condition, it is possible to measure the respiration signal by using the same proposed software algorithms from our study and just altering to an infrared camera with an infrared lighting source. During low-light conditions, the performance of our system may be affected by the noise generated within the camera sensor.

Extension possibilities with video observation: The major advantage of using a video-based method to monitor respiration is its contact-free operation. Both ChI and polysomnography need electrodes attached to the patient's skin, which increase the risk of pain and skin irritation. Therefore, our method can contribute to the comfort level and convenience of the infant. A further bene-

fit of using a camera is that it enables more measurements than contact-based bio-sensors, including physiological signals (e.g. breathing rate, heart rate, and blood oxygen saturation) [162, 170] and contexture signals (e.g. body motion, activities, and facial expressions) [141, 145, 148, 174]. These possibilities will enrich the functionality of a health monitoring system. Our system can be constructed with a generic webcam and an embedded computing platform, which forms a cost-effective solution. In principle, one camera can monitor multiple subjects/infants simultaneously, as long as they are captured by the camera view, while each contact-based bio-sensor can only monitor one single subject or infant.

7.5 Conclusions

This study has presented an automated processing workflow to estimate respiration signals from videos of premature infants in NICUs. The motion estimation methods of both optical flow and Deep Flow are employed for extracting respiration movement. The Deep Flow approach recruits CNNs for finding the replacement between target images and reference images. The proposed automated extracted respiration signals are compared to the signals being extracted from the regular chest impedance (ChI) measurement. The preliminary results are promising for further investigation of the video-based respiration monitoring method and for applying our automated respiration extraction for infants in NICUs. In most of the cases, the measured RMSE is below 5 %. Experiments have shown that the deep learning-based method outperforms the optical flow-based method both in accuracy (low RMSE) and robustness. Future work can investigate the possibility of directly applying a deep learning framework to estimate respiration rate. For example, an LSTM-based system can effectively incorporate temporal information for a regression task and is expected to further enhance the obtained results.

Chapter 8

Conclusions

8.1 Conclusions of the individual chapters

The objective of this thesis is to develop an automated video-based discomfort detection system for infants. The system aims to alert clinical staff immediately when infants start suffering from discomfort and express themselves accordingly.

Chapter 2 has introduced a facial expression-based method for automated detection of infant discomfort. The infant facial expression is analyzed at the image level for each static video frame. First, a specific pre-processing of face detection and normalization is proposed. Second, a two-phase classification workflow is employed, where Phase 1 is subject-independent, and Phase 2 is subject-dependent. Phase 1 derives geometric and appearance features, while Phase 2 incorporates facial landmark-based template matching. Finally, an SVM classifier is applied to video frames to recognize facial expressions of comfort or discomfort. The method is evaluated using videos from 22 infants. Experimental results show an AUC of 0.87 for the subject-independent phase (Phase 1), and 0.97 for the subject-dependent phase (Phase 2).

Chapter 3 has exploited deep CNN algorithms to address the problem of discomfort detection for infants by the learned analysis of their facial expressions. A dataset of 55 videos about facial expressions, recorded from 24 infants, is used in our study. Given the limited available data for training,

we have employed a pre-trained CNN model, which is followed by fine-tuning the networks using a public dataset with labeled facial expressions (the Shoulder-Pain dataset). The CNNs are further refined with our data of infants. After using twofold cross-validation, we achieve an AUC value of 0.96, which is substantially higher than the results without any pre-training steps (AUC=0.77). The proposed method also achieves better results than Chapter 2, which is based on handcrafted features. By fusing individual frame results over the video sequence, the AUC is further improved from 0.96 to 0.98.

Chapter 4 has investigated a video-based method for automated detection of infant discomfort, based on analyzing facial and body motion. Infant motion trajectories are calculated from frame-to-frame using optical flow. The motion acceleration rate is further derived from the motion magnitudes, which is followed by extracting 18 time-domain and frequency-domain features to identify different motion patterns. An SVM classifier is finally employed to discriminate infant status of comfort or discomfort. The method is evaluated using 183 video segments of 11 infants from 17 heel prick events. Experimental results show an AUC of 0.94 for discomfort detection and an average accuracy of 0.86 when combining all proposed features.

Chapter 5 has introduced a system that first employs the optical flow to estimate infant body motion trajectories across video frames. Following the movement estimation, Log Mel-spectrogram, Mel Frequency Cepstral Coefficients (MFCCs), and Spectral Subband Centroid Frequency (SSCF) features are extracted from the motion signal. This allows a presentation of 1D motion signals by 2D time-frequency representations of the distribution of signal energy. Finally, deep CNNs are applied to the 2D images for the binary comfort/discomfort classification. The performance of the model is assessed using leave-one-infant-out cross-validation. The proposed algorithm is evaluated on a dataset containing 183 video segments recorded from 11 infants during 17 heel prick events, which is a pain stimulus associated with a routine care procedure. Experimental results have shown an area under the receiver operating characteristic curve of 0.985 and an accuracy of 94.2 %, thereby offering a promising possibility to deploy the proposed system in clinical practice. The spectrum representation has improved the AUC of about 4 percent points comparing to the SVM-based classification in Chapter 4.

Chapter 6 has presented a novel and efficient discomfort detection system based on 3D CNN, which achieves an end-to-end solution without the

conventional face detection and tracking steps. The scheme and the approach of this study is leveraging the video characteristics of spatial contextual and temporal (motion) information. The architectures of both 2D and 3D networks are investigated for the task of infant discomfort detection. The 3D CNNs enable to capture both the body motion and the facial expression from infants. The performance of the system is assessed using videos recorded from 24 hospitalized infants by visualizing ROC curves and measuring the AUC values. Additional performance metrics (labeling accuracy) are also calculated. Experimental results show that the proposed system achieves an AUC of 0.99, while the overall labeling accuracy is also 0.99, which confirms the robustness by using a 3D CNN for infant discomfort monitoring, simultaneously capturing both motion and facial expressions. The results in this chapter offer the highest performance comparing to previous chapters, which is explained by the joint learning of spatial and temporal features of the infant behavior.

Chapter 7 has investigated an automated video-based workflow to estimate respiration rates for premature infants in NICUs. Two flow estimation methods are employed to estimate motion between video frames, namely the conventional optical flow-based and deep learning-based flow calculation methods, which are then compared in performance. The respiratory signal is finally extracted via decomposition of the motion matrix. The proposed methods are evaluated by comparing the proposed automated extracted respiration signals to that of the extracted chest-impedance signal on videos of five premature infants. The overall average cross-correlation coefficients are 0.70 for the optical flow-based method and 0.74 for the Deep Flow-based method. The average root mean-squared errors are 6.10 bpm and 4.55 bpm for the optical flow-based and the Deep Flow-based methods, respectively. The experimental results are interesting for further investigation and clinical application of the video-based respiration monitoring method for infants in NICUs. The contribution of this chapter is to benefit further from video-based monitoring by investigating the feasibility of contactless respiration measurement, and further add value to the measurement to the health-state monitoring of infants.

8.2 Discussion on the research questions

In this section, the performance of the proposed methods and systems is discussed with respect to the posed research questions in Chapter 1.

RQ1: Facial features for frame-level discomfort detection

- RQ1.a: Which facial features are relevant and discriminating for characterizing the facial expression for the infant comfort/discomfort detection task?

  In Chapter 2, this question is addressed. Regarding handcrafted features, we have investigated two categories of features, geometric and appearance features, to describe the facial expressions. For geometric features, the areas of left eye, right eye, outer lip contour, and inner lip contour are calculated. For the appearance features, Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBPs) are used to describe the facial texture. From all geometric and appearance features, the most discriminative features are the geometric features that combine the sizes of four facial regions, namely left eye, right eye, outer lip contour, and inner lip contour. The combination of these geometric features achieves the best AUC of 0.85. However, in contrast, LBP is less effective (AUC = 0.71) than the combination of these geometric features. The resulted features are a logical finding because they are important in the description of comfort and discomfort analysis, as used in clinical pain scales.

- RQ1.b: How can we normalize the facial features across different infants, since infants have different levels of facial expressions?

  Across Chapter 2, we have adopted two solutions to normalize the facial information to control the variations among different infants. The first solution is to normalize the face ROI. In Chapter 2, 68 facial landmarks are first detected using the Dlib face landmark detector. Based on the detected 68 landmarks as a reference, the rotation and size variance of faces across infants are corrected. This correction involves the usage of a part of the feature points previously discussed to rotate, scale, and crop the original face image for normalization. The second solution is based on the application of subject-dependent features. For each individual infant, the annotated comfort moments are obtained, of which the faces are predefined as parameterized templates to compare the similarity with an unseen input frame. Compared to the results without template matching features, the AUC of template matching features increases from 0.89 to 0.97. The overall accuracy of results without template matching for the 10-infant dataset is 0.79, and 0.95 for combining the template matching, which amounts to a significant accuracy increase of 20 %. Because of the individual template matching for each infant, the last solution is more

powerful and suited to customize the facial expression levels of different infants.

RQ2: Deep learning-based frame-level discomfort detection

- RQ2.a: Can deep learning be used for this infant comfort/discomfort classification task? If so, in what way?

  This question has been investigated as a possibility in Chapter 3, where first the face ROI is extracted from each video frame, and then deep learning is applied on the face ROIs. The deep learning approach is compared with the handcrafted feature-based method, which shows that the results of deep learning-based methods substantially increase the performance (AUC increases from 0.87 to 0.96). The results strongly indicate that deep learning methods can replace the conventional handcrafted features for infant monitoring applications in state-of-the-art settings.

- RQ2.b: Can deep learning be applied on a small dataset or are special actions required to facilitate this?

  In Chapter 3, it has been discussed that a small dataset is not suitable for fine-tuning a large set of parameters of the entire network at the initial stages of the deep learning framework. The AUC without any pre-training only achieves the value of 0.77. Therefore, a DenseNet model is first pre-trained on ImageNet data (AUC increased to 0.93), followed by a first fine-tuning step of training the model with a relatively large and similar dataset to tune the networks. In our case and for that purpose, the Shoulder-Pain dataset is used, which includes the labeled facial expressions recorded from adults. The second fine-tuning step is leveraging our own data of infants. The results show that with that extra fine-tuning step on the adult dataset, the AUC is further improved from 0.93 to 0.96. However, the performance of joint training with the adult and infant dataset combined has not been investigated, which is considered as future work. The nodes of the output layer can be adjusted to support both populations correspondingly.

- RQ2.c: Is deep learning sufficiently robust to environmental settings and changes?

  In Chapter 3, various changes to the environment and settings have been explored. The performed is evaluated by 1) randomly zooming in and out the face images from 0.8 up to 1.2, 2) randomly rotating the face from

-45 to +45 degrees, and 3) randomly changing the contrast from 0.8 to 1.2. The obtained experimental results have shown similar and consistent performance, which clearly demonstrates that the proposed system is robust to different and varying video conditions. This is explained by applying the data augmentation in the training to mimic such situations.

RQ3: Incorporating temporal information - motion analysis

- RQ3.a: Can we leverage the motion information for infant discomfort detection?

  In Chapters 4, 5, and 6, the motion information is converted from video segments to different representations such as 1D signal, 2D feature maps, or directly using the original video segments as inputs for further comfort classification. The work in these chapters proves the effective use of motion information in various ways for infant discomfort detection. The performance of the individual approaches is addressed below.

- RQ3.b: Can we employ optical flow and handcrafted features for the extracted motion signal to capture the temporal information?

  The results of Chapter 4 have shown that it is feasible to employ optical flow for capturing the infant motion information. The handcrafted features of 18 time/frequency-domain features are further calculated, which include the mean, median, root mean square, autocorrelation, and spectrum features. The highest performance (AUC = 0.94) is achieved when combining all these features together with an SVM classifier.

- RQ3.c: In what way can deep learning facilitate the classification of a 1D motion signal?

  To apply deep learning on the extracted 1D motion signal while avoiding handcrafted features, multiple options are available. The first option is to apply 1D convolutional neural networks. The second is to apply Long Short-Term Memory (LSTM) as part of the learning framework, but this is not yet explored in this thesis because of the initial high computing cost. However, the most sophisticated deep learning development is for 2D images. Therefore, as an alternative, in Chapter 5, the 1D signal is converted into 2D representations, after which ResNet is applied for classification. This leads to a high AUC value of 0.985, which clearly outperforms the previous 1D classification system based on handcrafted features.

- RQ3.d: Can we use CNNs to jointly process both spatial and temporal information?

  A number of existing CNNs are only capable of handling 2D inputs due to the inherent network structure. The 2D CNNs are limited to the spatial information, while the temporal information across the video frames is not exploited. In Chapter 6, 3D CNNs are proposed to extract features from both the spatial and temporal domains by performing 3D convolution and 3D pooling. This approach is able to capture both the object appearance/contextual information and motion information in a single optimization framework, such that the generated features assemble the spatial and temporal semantics. A thorough benchmark is performed between a 2D CNN and a 3D CNN, where the best result of 2D CNN yields the AUC of 0.91, while the 3D CNN obtains the AUC of 0.99. Therefore, it can be concluded that 3D CNNs can exploit both spatial and temporal information jointly to provide an improved prediction. It should be indicated though that there is a clear complexity trade-off to be made between the two networks. The execution time performance of 3D CNN is comparable with the 2D CNN implementation, but the hardware of the experiments has provided sufficient computation power in both cases, so that the execution time indication has limited value. Furthermore, the 3D CNN requires about 500k parameters, which is not indicating a very high complexity. To make the real trade-off, this aspect should be further studied in details, but we estimate that the 3D CNN seems to be feasible with respect to computing complexity.

- RQ3.e: How can an additional motion channel give guidance to CNNs to focus on areas related to discomfort?

  In Chapter 6, the benefit of using optical flow measurement to draw the attention of 3D CNNs has been investigated. The followed approach is completely data-driven when video segments are used to train a CNN model. Usually, for data-driven approaches, a large amount of training data is required for optimal results. In Chapter 6, the training of CNNs is additionally guided by the motion information encoded by optical flow in combination with the normal video channels. This combination yields an extra AUC gain of up to 3 percent in terms of classification performance.

RQ4: Quantitative physiological signal measurement

- RQ4.a: How can we quantitatively extract physiological signals through videos?

  In Chapter 7, this question is addressed by investigating two approaches, namely the conventional optical flow-based and deep learning-based flow estimation methods. The motion estimation methods are employed to estimate pixel motion vectors between adjacent video frames. The spatial statistics of the flow matrix are further analyzed by applying robust principal component analysis (PCA). The first Eigenvalue of the covariance matrix of flows represents the major motion component. The experimental results show that the deep learning-based method clearly outperforms the optical flow-based method in accuracy and robustness.

- RQ4.b: Are the results calculated from videos reliable by comparing them with the current standard methods (chest impedance)?

  Chest impedance (ChI) is the current standard for cardio-respiratory monitoring in preterm infants. The proposed video-based algorithms in Chapter 7 are evaluated by comparing the proposed automated extracted respiration signals to that extracted from chest impedance. The cross-correlation coefficients and the RMSE are computed as measures to compare the respiration rates. Five videos of premature infants were recorded for evaluation, and ChI signals were also recorded simultaneously as reference standards. The experimental results show that the overall average cross-correlation coefficients are 0.70 for the optical flow-based method and 0.74 for the Deep Flow-based method. The average root mean-squared errors are 6.10 bpm and 4.55 bpm for the optical flow-based and the Deep Flow-based methods, respectively. Considering that newborn babies usually breathe between 40 and 60 bpm, the results of the proposed methods show the feasibility of applying the video-based contactless respiration monitoring.

8.3 Outlook on AI for infant monitoring

Machine learning and computer vision have strengthened many aspects of human visual perception to identify clinically meaningful patterns, e.g., in the medical imaging field. CNNs have been widely used for a variety of tasks from medical image processing, computer-aided diagnosis, and prediction of

clinical data, etc. Novel and efficient artificial intelligence and machine learning algorithms are rapidly emerging. For further work of the infant monitoring application, novel techniques, such as Long Short-Term Memory (LSTM) and the newly designed Transformer model can be investigated for better leveraging the temporal information embedded in videos.

LSTM [52] computes raw time-series data. For future investigation, the input can be selected as 1D signal segments of motion information derived from optical flow. It is also possible to utilize multi-dimensional input, which requires an initial step of extracting frame feature vectors using pre-trained CNNs. More recently, the Transformer model has been introduced [163], which is based on self-attention. Self-attention allows for more direct information flow across the whole signal sequence, and thus more direct gradient flow. Furthermore, this concept facilitates faster training, since most operations can be calculated in parallel for the standard cross-entropy loss. The Transformer model has become a very successful model in various applications, which promotes the interest of a benchmark of our infant monitoring tasks.

On the other hand, to improve the effectiveness of artificial intelligence models for an optimal discomfort assessment, future work could involve incorporating more information or signals such as audio signals and corresponding physiological measurements of vital signs, and other contextual factors (e.g. gender, gestational age, presence of parents) of infants into the same networks. Moreover, another reason for considering combining physiological measures is that it has been noticed that extreme premature/sick infants are limited to behaviorally express discomfort/pain. Therefore, it is important to consider both behavior and physiological indicators for monitoring.

One limitation of the current work is that the discomfort/pain moments have been mostly recorded during acute painful procedures. It has been found that infant behavioral and physiological responses to chronic pain can be different from acute pain. Therefore, monitoring and analyzing different types of pain is essential for an efficient and reliable assessment. Furthermore, to describe the infant status more precisely, the discomfort level can be specified by providing grades from the system.

From the clinical perspective, a fully contactless infant monitoring system is highly desirable to avoid the wires and connectors in physiological monitoring and improve the comfort quality of infants in bed accordingly. Thus, video-based physiological measurements can be further investigated, such as video-based heart rate estimation. The current research only focuses on identifying infant discomfort moments. The proposed algorithms can also be derived for predicting/detecting specific diseases (e.g. sepsis). Besides, it has

been widely accepted that excessive false and nuisance alarms are a real clinical and economical problem. Future work of the system can be integrated with the proposed artificial intelligence models and combined with the current NICU monitoring system to successfully assist in alarm management, especially to reduce false alarms and provide continuous monitoring in a safe, attractive way.

Bibliography

[1] Abbas K Abbas, Konrad Heimann, Katrin Jergus, Thorsten Orlikowsky, and Steffen Leonhardt. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *Biomedical engineering online*, 10(1):93, 2011. (Cited on page 114.)

[2] R Acharya, Ashwin Kumar, PS Bhat, CM Lim, N Kannathal, SM Krishnan, et al. Classification of cardiac abnormalities using heart rate signals. *Medical and Biological Engineering and Computing*, 42(3):288–293, 2004. (Cited on pages 19 and 69.)

[3] FS Afsar. Skin care for preterm and term neonates. *Clinical and Experimental Dermatology: Clinical dermatology*, 34(8):855–858, 2009. (Cited on page 112.)

[4] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. (Cited on page 24.)

[5] Bruce Ambuel, Kim W Hamlett, Celeste M Marx, and Jeffrey L Blumer. Assessing distress in pediatric intensive care environments: the comfort scale. *Journal of pediatric psychology*, 17(1):95–109, 1992. (Cited on pages 3 and 69.)

[6] L Antognoli, P Marchionni, S Nobile, VP Carnielli, and L Scalise. Assessment of cardio-respiratory rates by non-invasive measurement methods

in hospitalized preterm neonates. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5. IEEE, 2018. (Cited on page 115.)

[7] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009. (Cited on pages 20 and 44.)

[8] Gary Baker, Mark Norman, Mohan Karunanithi, and Colin Sullivan. Contactless monitoring for sleep disordered-breathing, respiratory and cardiac co-morbidity in an elderly independent living cohort, 2015. (Cited on page 112.)

[9] John L Barron, David J Fleet, Steven S Beauchemin, and TA Burkitt. Performance of optical flow techniques. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 236–242. IEEE, 1992. (Cited on pages 113, 117, and 118.)

[10] Stacy Beck, Daniel Wojdyla, Lale Say, Ana Pilar Betran, Mario Merialdi, Jennifer Harris Requejo, Craig Rubens, Ramkumar Menon, and Paul FA Van Look. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bulletin of the World Health Organization*, 88:31–38, 2010. (Cited on page 68.)

[11] RE Behrman and A Stith Butler. Institute of medicine committee on understanding premature birth and assuring healthy outcomes board on health sciences outcomes: preterm birth: causes, consequences, and prevention. *Preterm birth: causes, consequences, and prevention, National Academies Press, Washington, DC*, 2007. (Cited on page 2.)

[12] Igal Bilik and Peter Khomchuk. Minimum divergence approaches for robust classification of ground moving targets. *IEEE Transactions on Aerospace and Electronic Systems*, 48(1):581–603, 2012. (Cited on page 82.)

[13] J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986. (Cited on page 120.)

[14] Hannah Blencowe, Simon Cousens, Mikkel Z Oestergaard, Doris Chou, Ann-Beth Moller, Rajesh Narwal, Alma Adler, Claudia Vera Garcia,

Sarah Rohde, Lale Say, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*, 379(9832):2162–2172, 2012. (Cited on page 68.)

[15] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997. (Cited on page 27.)

[16] Sheryl Brahnam, Chao-Fa Chuang, Frank Y Shih, and Melinda R Slack. Machine recognition and representation of neonatal facial displays of acute pain. *Artificial intelligence in medicine*, 36(3):211–222, 2006. (Cited on pages 22 and 27.)

[17] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, March 2011. (Cited on page 118.)

[18] Luca Cattani, Davide Alinovi, Gianluigi Ferrari, Riccardo Raheli, Elena Pavlidis, Carlotta Spagnoli, and Francesco Pisani. Monitoring infants by automatic video processing: A unified approach to motion analysis. *Computers in Biology and Medicine*, 80:158–165, 2017. (Cited on page 70.)

[19] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3D facial expressions. *arXiv preprint arXiv:1802.00542*, 2018. (Cited on page 22.)

[20] Wei Chen, Idowu Ayoola, Sidarto Bambang Oetomo, and Loe Feijs. Non-invasive blood oxygen saturation monitoring for neonates using reflectance pulse oximeter. In *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pages 1530–1535. IEEE, 2010. (Cited on page 7.)

[21] Wei Chen, Sietse Dols, Sidarto Bambang Oetomo, and Loe Feijs. Monitoring body temperature of newborn infants at neonatal intensive care units using wearable sensors. In *Proceedings of the Fifth International Conference on Body Area Networks*, pages 188–194, 2010. (Cited on page 7.)

[22] Victor Chernick, Fred Heldrich, and Mary Ellen Avery. Periodic breathing of premature infants. *The Journal of pediatrics*, 64(3):330–340, 1964. (Cited on page 111.)

[23] Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multi-source transfer learning with convolutional neural networks for lung pattern analysis. *IEEE Journal of Biomedical & Health Informatics*, PP(99):1–1, 2016. (Cited on page 45.)

[24] Xiaomei Cong, Jing Wu, Dorothy Vittner, Wanli Xu, Naveed Hussain, Shari Galvin, Megan Fitzsimons, Jacqueline M McGrath, and Wendy A Henderson. The impact of cumulative pain/stress on neurobehavioral development of preterm infants in the NICU. *Early human development*, 108:9–16, 2017. (Cited on page 2.)

[25] Anis Davoudi, Kumar Rohit Malhotra, Benjamin Shickel, Scott Siegel, Seth Williams, Matthew Ruppert, Emel Bihorac, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Azra Bihorac, et al. The intelligent icu pilot study: using artificial intelligence technology for autonomous patient monitoring. *arXiv preprint arXiv:1804.10201*, 2018. (Cited on page 77.)

[26] Philip De Chazal, Emer O'Hare, Niall Fox, and Conor Heneghan. Assessment of sleep/wake patterns using a non-contact biomotion sensor. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 514–517. IEEE, 2008. (Cited on page 114.)

[27] F. Deng, J. Dong, X. Wang, Y. Fang, Y. Liu, Z. Yu, J. Liu, and F. Chen. Design and implementation of a noncontact sleep monitoring system using infrared cameras and motion sensor. *IEEE Transactions on Instrumentation and Measurement*, 67(7):1555–1563, July 2018. (Cited on page 113.)

[28] Oscar Déniz, Gloria Bueno, Jesús Salido, and Fernando De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011. (Cited on page 24.)

[29] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015. (Cited on page 93.)

[30] Guiguang Ding, Yuchen Guo, Kai Chen, Chaoqun Chu, Jungong Han, and Qionghai Dai. Decode: deep confidence network for robust image classification. *IEEE Transactions on Image Processing*, 2019. (Cited on page 44.)

[31] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. (Cited on page 59.)

[32] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, 2016. (Cited on page 93.)

[33] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, pages 363–370. Springer, 2003. (Cited on pages 71, 72, 81, and 100.)

[34] Maria Fitzgerald. The development of nociceptive circuits. *Nature Reviews Neuroscience*, 6(7):507–520, 2005. (Cited on page 2.)

[35] Maria Fitzgerald, Catherine Millard, and Neil McIntosh. Cutaneous hypersensitivity following peripheral tissue damage in newborn infants and its reversal with topical anaesthesia. *Pain*, 39(1):31–36, 1989. (Cited on page 2.)

[36] E Fotiadou, S Zinger, W Tjon A Ten, S Bambang Oetomo, et al. Video-based facial discomfort analysis for infants. In *IS&T/SPIE Electronic Imaging*, pages 90290F–90290F. International Society for Optics and Photonics, 2014. (Cited on pages 21 and 70.)

[37] E. Fotiadou, S. Zinger, W. E. Tjon a Ten, S. Bambang Oetomo, and P. H. N. de With. Video-based facial discomfort analysis for infants. In *Proceedings Volume 9029, Visual Information Processing and Communication*, volume 3, page 9029, 2014. (Cited on pages 42 and 43.)

[38] Behnood Gholami, Wassim M Haddad, and Allen R Tannenbaum. Relevance vector machine learning for neonate pain intensity assessment using digital imaging. *IEEE Transactions on biomedical engineering*, 57(6):1457–1466, 2010. (Cited on pages 22 and 27.)

[39] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011. (Cited on page 82.)

[40] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. (Cited on page 93.)

[41] Ruth E Grunau, Michael F Whitfield, Julianne Petrie-Thomas, Anne R Synnes, Ivan L Cepeda, Adi Keidar, Marilyn Rogers, Margot MacKay, Philippa Hubber-Richard, and Debra Johannesen. Neonatal pain, parenting stress and interaction, in relation to cognitive and motor development at 8 and 18 months in preterm infants. *Pain*, 143(1):138–146, 2009. (Cited on page 2.)

[42] Ruth Eckstein Grunau and Mai T Tu. Long-term consequences of pain in human neonates. *Pain in neonates and infants*, 3:45–55, 2007. (Cited on page 2.)

[43] R.V. Grunau and K.D. Craig. Pain expression in neonates: facial action and cry. *Pain*, 28(3):395–410, 1987. (Cited on page 21.)

[44] P. Gupta, B. Bhowmick, and A. Pal. Accurate heart-rate estimation from face videos using quality-based fusion. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, pages 4132–4136, September 2017. (Cited on page 114.)

[45] Z. Hammal and J. F. Cohn. Towards multimodal pain assessment for research and clinical use. In *Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, volume 3, pages 13–17, 2014. (Cited on pages 19 and 69.)

[46] Zakia Hammal and Jeffrey F. Cohn. Automatic detection of pain intensity. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 47–52, New York, NY, USA, 2012. ACM. (Cited on page 44.)

[47] Jungong Han, L. Hazelhoff, and P.H.N. With, de. *Neonatal monitoring based on facial expression analysis*, pages 303–323. IGI Global, 2012. (Cited on page 44.)

[48] Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. (Cited on page 74.)

[49] Lykele Hazelhoff, Jungong Han, Sidarto Bambang-Oetomo, and Peter H. N. de With. Behavioral state detection of newborns based on facial expression analysis. In Jacques Blanc-Talon, Wilfried Philips, Dan

Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 698–709, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. (Cited on page 44.)

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. (Cited on pages 49 and 83.)

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645, 2016. (Cited on page 47.)

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. (Cited on page 135.)

[53] Shin Hoochang, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. (Cited on page 45.)

[54] Rex Hsieh, Yuya Mochizuki, Takaya Asano, Marika Higashida, and Akihiko Shirai. Real baby-real family: Vr entertainment baby interaction system. In *ACM SIGGRAPH 2017 Emerging Technologies*, page 20. ACM, 2017. (Cited on page 22.)

[55] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016. (Cited on pages 47, 49, and 50.)

[56] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. (Cited on page 47.)

[57] Hyungkeun Jee, Kyunghee Lee, and Sungbum Pan. Eye and face detection using svm. In *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004.*, pages 577–580. IEEE, 2004. (Cited on page 18.)

[58] In Cheol Jeong and Joseph Finkelstein. Introducing contactless blood pressure assessment using a high speed video camera. *Journal of medical systems*, 40(4):77, 2016. (Cited on page 7.)

[59] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. (Cited on page 97.)

[60] Dongsheng Jiang, Weiqiang Dou, Luc Vosters, Xiayu Xu, Yue Sun, and Tao Tan. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. *Japanese journal of radiology*, 36(9):566–574, 2018. (Cited on page 162.)

[61] C Celeste Johnston, Bonnie J Stevens, Fang Yang, and Linda Horton. Differential response to pain by very premature neonates. *Pain*, 61(3):471–479, 1995. (Cited on page 2.)

[62] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015. (Cited on page 94.)

[63] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE, 2000. (Cited on page 93.)

[64] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society, 2014. (Cited on pages 22 and 47.)

[65] Siavash Khallaghi, C Antonio Sánchez, Abtin Rasoulian, Yue Sun, Farhad Imani, Amir Khojaste, Orcun Goksel, Cesare Romagnoli, Hamidreza Abdi, Silvia Chang, et al. Biomechanically constrained surface registration: Application to MR-TRUS fusion for prostate interventions. *IEEE transactions on medical imaging*, 34(11):2404–2414, 2015. (Cited on page 162.)

[66] Reza Kharghanian, Ali Peiravi, and Farshad Moradi. Pain detection from facial images using unsupervised feature learning approach. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 419–422. IEEE, 2016. (Cited on pages 20 and 44.)

[67] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. (Cited on pages 22 and 46.)

[68] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. (Cited on page 51.)

[69] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 100.)

[70] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. (Cited on page 83.)

[71] R Klein. Bland-altman and correlation plot. *Mathworks File Exchange*, 2014. (Cited on page 120.)

[72] Ninah Koolen, Olivier Decroupet, Anneleen Dereymaeker, Katrien Jansen, Jan Vervisch, Vladimir Matic, Bart Vanrumste, Gunnar Naulaers, Sabine Van Huffel, and Maarten De Vos. Automated respiration detection from neonatal video data. In *ICPRAM (2)*, pages 164–169, 2015. (Cited on page 114.)

[73] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, 16(1):172–187, 2007. (Cited on pages 20, 43, and 70.)

[74] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. (Cited on page 47.)

[75] Jocelyn Lawrence, Denise Alcock, Patrick McGrath, J Kay, S Brock MacMurray, and C Dulberg. The development of a tool to assess neonatal pain. *Neonatal network: NN*, 12(6):59, 1993. (Cited on page 3.)

[76] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. (Cited on page 41.)

[77] Cheng Li, Arash Pourtaherian, Lonneke Van Onzenoort, Walter E Tjon a Ten, and Peter HN de With. Infant monitoring system for real-time and remote discomfort detection. *IEEE Transactions on Consumer Electronics*, 66(4):336–345, 2020. (Cited on page 125.)

[78] Kuan Li, Yi Jin, Muhammad Waqar Akram, Ruize Han, and Jiongwei Chen. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. *The Visual Computer*, 36(2):391–404, 2020. (Cited on page 94.)

[79] Zhang Li, Fan Huang, Jiong Zhang, Behdad Dashtbozorg, Samaneh Abbasi-Sureshjani, Yue Sun, Xi Long, Qifeng Yu, Bart ter Haar Romeny, and Tao Tan. Multi-modal and multi-vendor retina image registration. *Biomedical optics express*, 9(2):410–422, 2018. (Cited on page 162.)

[80] Zhang Li, Zheng Zhong, Yang Li, Tianyu Zhang, Liangxin Gao, Dakai Jin, Yue Sun, Xianghua Ye, Li Yu, Zheyu Hu, et al. From community acquired pneumonia to covid-19: A deep learning based method for quantitative analysis of covid-19 on thick-section CT scans. *European Radiology*, 30(12):6828–6837, 2020. (Cited on page 161.)

[81] Viveca Lindh, Urban Wiklund, and Stellan Håkansson. Heel lancing in term new-born infants: an evaluation of pain by frequency domain analysis of heart rate variability. *Pain*, 80(1-2):143–148, 1999. (Cited on page 19.)

[82] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009. (Cited on pages 20 and 44.)

[83] Kuang Liu, Mingmin Zhang, and Zhigeng Pan. Facial expression recognition with cnn ensemble. In *2016 international conference on cyberworlds (CW)*, pages 163–166. IEEE, 2016. (Cited on page 93.)

[84] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, volume 270, pages 1–11, 2000. (Cited on page 82.)

[85] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017. (Cited on page 45.)

[86] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. (Cited on page 24.)

[87] Shangzhen Luan, Chen Chen, Baochang Zhang, Jungong Han, and Jianzhuang Liu. Gabor convolutional networks. *IEEE Transactions on Image Processing*, 27(9):4357–4366, 2018. (Cited on page 94.)

[88] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. (Cited on page 93.)

[89] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain 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*, pages 57–64. IEEE, 2011. (Cited on pages 20, 44, 46, 50, and 62.)

[90] Carolyn H Lund, Lourdes B Nonato, Joanne M Kuller, Linda S Franck, Chris Cullander, and David K Durand. Disruption of barrier function in neonatal skin associated with adhesive removal. *The Journal of pediatrics*, 131(3):367–372, 1997. (Cited on page 112.)

[91] Bahram Marami, Shahin Sirouspour, Suha Ghoul, Shadi Emami Abarghouei, Yue Sun, and Aaron Fenster. Non-rigid MRI-TRUS registration in targeted prostate biopsy. In *Medical Imaging 2015: Image Processing*, volume 9413, page 941332. International Society for Optics and Photonics, 2015. (Cited on page 164.)

[92] Gregory Matthews, Barry Sudduth, and Michael Burrow. A non-contact vital signs monitor. *Critical Reviews in Biomedical Engineering*, 28(1&2), 2000. (Cited on page 112.)

[93] Ronald Melzack and Joel Katz. The gate control theory: Reaching for the brain. *Pain: psychological perspectives*, pages 13–34, 2004. (Cited on page 69.)

[94] Yuki Mima and Kaoru Arakawa. Cause estimation of younger babies' cries from the frequency analyses of the voice-classification of hunger, sleepiness, and discomfort. In *Intelligent Signal Processing and Communications, 2006. ISPACS'06. International Symposium on*, pages 29–32. IEEE, 2006. (Cited on pages 19 and 70.)

[95] Lan Min, Yue Sun, and Peter H. N. de With. Video-based infant monitoring using a CNN-LSTM scheme. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 1159717. International Society for Optics and Photonics, 2021. (Cited on page 163.)

[96] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. (Cited on page 93.)

[97] N. Neshov and A. Manolova. Pain detection from facial characteristics using supervised descent method. In *Proc. IEEE 8th Int. Conf. Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 251–256, September 2015. (Cited on page 44.)

[98] J Norden, R Hannallah, P Getson, R O'Donnell, G Kelliher, and N Walker. Reliability of an objective pain scale in children. *Journal of pain and symptom management*, 6(3):196, 1991. (Cited on pages 3 and 69.)

[99] US Department of Health, Human Services, et al. Acute pain management in infants, children, and adolescents: Operative and medical procedures. *Agency for Health Care Policy and Research: Rockville, Maryland*, 1992. (Cited on page 2.)

[100] American Academy of Pediatrics, Fetus, Newborn Committee, et al. Prevention and management of pain in the neonate: an update. *Pediatrics*, 118(5):2231–2241, 2006. (Cited on page 2.)

[101] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. (Cited on pages 24 and 25.)

[102] Edgar Osuna, Robert Freund, and Federico Girosit. Training support vector machines: an application to face detection. In *Proceedings of IEEE*

computer society conference on computer vision and pattern recognition, pages 130–136. IEEE, 1997. (Cited on page 18.)

[103] Gayle Giboney Page. Are there long-term consequences of pain in newborn or very young infants? *The Journal of perinatal education*, 13(3):10, 2004. (Cited on pages 2 and 69.)

[104] Arpan Pal, Aniruddha Sinha, Anirban Dutta Choudhury, Tanushyam Chattopadyay, and Aishwarya Visvanathan. A robust heart rate detection using smart-phone video. In *Proceedings of the 3rd ACM MobiHoc workshop on Pervasive wireless healthcare*, pages 43–48, 2013. (Cited on page 7.)

[105] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. (Cited on page 49.)

[106] J Anthony Parker, Robert V Kenyon, and Donald E Troxel. Comparison of interpolating methods for image resampling. *IEEE Transactions on medical imaging*, 2(1):31–39, 1983. (Cited on page 96.)

[107] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. (Cited on pages 113 and 119.)

[108] Joann R Petrini, Todd Dias, Marie C McCormick, Maria L Massolo, Nancy S Green, and Gabriel J Escobar. Increased risk of adverse neurological development for late preterm infants. *The Journal of pediatrics*, 154(2):169–176, 2009. (Cited on pages 2 and 69.)

[109] Christian F Poets, Valerie A Stebbens, Martin P Samuels, and David P Southall. The relationship between bradycardia, apnea, and hypoxemia in preterm infants. *Pediatric research*, 34(2):144, 1993. (Cited on page 111.)

[110] A. P. Prathosh, P. Praveena, L. K. Mestha, and S. Bharadwaj. Estimation of respiratory pattern from video using selective ensemble aggregation. *IEEE Transactions on Signal Processing*, 65(11):2902–2916, June 2017. (Cited on page 114.)

[111] Heinz FR Prechtl. The behavioural states of the newborn infant (a review). *Brain research*, 76(2):185–212, 1974. (Cited on page 112.)

[112] HFR Prechtl, K Theorell, and AW Blair. Behavioural state cycles in abnormal infants. *Developmental Medicine & Child Neurology*, 15(5):606–615, 1973. (Cited on page 112.)

[113] Kenneth M Prkachin and Patricia E Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008. (Cited on page 51.)

[114] Wu Qiu, Martin Rajchl, Fumin Guo, Yue Sun, Eranga Ukwatta, Aaron Fenster, and Jing Yuan. 3D prostate TRUS segmentation using globally optimized volume-preserving prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 796–803. Springer, 2014. (Cited on page 164.)

[115] Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Efficient 3D multi-region prostate MRI segmentation using dual optimization. In *International Conference on Information Processing in Medical Imaging*, pages 304–315. Springer, 2013. (Cited on page 164.)

[116] Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Fast globally optimal segmentation of 3d prostate mri with axial symmetry prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 198–205. Springer, 2013. (Cited on page 164.)

[117] Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Dual optimization based prostate zonal segmentation in 3D MR images. *Medical image analysis*, 18(4):660–673, 2014. (Cited on page 162.)

[118] Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Prostate segmentation: an efficient convex optimization approach with axial symmetry using 3-D TRUS and MR images. *IEEE transactions on medical imaging*, 33(4):947–960, 2014. (Cited on page 162.)

[119] AF Quiceno-Manrique, JB Alonso-Hernandez, CM Travieso-Gonzalez, MA Ferrer-Ballester, and G Castellanos-Dominguez. Detection of obstructive sleep apnea in ecg recordings using time-frequency distributions and dynamic features. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5559–5562. IEEE, 2009. (Cited on page 82.)

[120] Arushi Raghuvanshi and Vivek Choksi. Facial expression recognition with convolutional neural networks. *CS231n Course Projects*, 2016. (Cited on page 45.)

[121] Tonse NK Raju, Rosemary D Higgins, Ann R Stark, and Kenneth J Leveno. Optimizing care and outcome for late-preterm (near-term) infants: a summary of the workshop sponsored by the national institute of child health and human development. *Pediatrics*, 118(3):1207–1214, 2006. (Cited on page 2.)

[122] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M Gonfaus, Kamal Nasrollahi, Thomas B Moeslund, and F Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, 2017. (Cited on page 43.)

[123] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *International Symposium on Visual Computing*, pages 234–243. Springer, 2013. (Cited on page 51.)

[124] Justin Salamon and Juan Pablo Bello. Feature learning with deep scattering for urban sound analysis. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 724–728. IEEE, 2015. (Cited on page 82.)

[125] Steven Michael Sale. Neonatal apnoea. *Best Practice & Research Clinical Anaesthesiology*, 24(3):323–336, 2010. (Cited on page 111.)

[126] Martin Schiavenato, Jacquie F Byers, Paul Scovanner, James M McMahon, Yinglin Xia, Naiji Lu, and Hua He. Neonatal pain facial expression: Evaluating the primal face of pain. *Pain*, 138(2):460–471, 2008. (Cited on pages 16 and 41.)

[127] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. (Cited on page 50.)

[128] Caifeng Shan. An efficient approach to smile detection. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 759 – 764. IEEE, 2011. (Cited on page 43.)

[129] Caifeng Shan and Ralph Braspenning. Recognizing facial expressions automatically from video. In *Handbook of ambient intelligence and smart environments*, pages 479–509. Springer, Boston, MA, 20010. (Cited on page 43.)

[130] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages II–370. IEEE, 2005. (Cited on page 24.)

[131] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. (Cited on pages 20, 43, and 70.)

[132] Clyde Shavers, Robert Li, and Gary Lebby. An svm-based approach to face detection. In *2006 Proceeding of the Thirty-Eighth Southeastern Symposium on System Theory*, pages 362–366. IEEE, 2006. (Cited on page 18.)

[133] Gilberto Sierra, Victor F Lanzo, and Valery Telfort. Non-invasive monitoring of respiratory rate, heart rate and apnea, February 4 2014. US Patent 8,641,631. (Cited on page 7.)

[134] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang. Automated assessment of children's postoperative pain using computer vision. *Pediatrics*, 136(1):124–131, 2015. (Cited on pages 21 and 42.)

[135] Karan Sikka, Alex A Ahmed, Damaris Diaz, Matthew S Goodwin, Kenneth D Craig, Marian S Bartlett, and Jeannie S Huang. Automated assessment of children's postoperative pain using computer vision. *Pediatrics*, 136(1):e124, 2015. (Cited on page 77.)

[136] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. (Cited on page 97.)

[137] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. (Cited on page 47.)

[138] Bonnie Stevens, Celeste Johnston, Patricia Petryshen, and Anna Taddio. Premature infant pain profile: development and initial validation. *The Clinical journal of pain*, 12(1):13–22, 1996. (Cited on page 3.)

[139] Yue Sun, Jingjing Hu, Wenjin Wang, Min He, and Peter H. N. de With. Camera-based discomfort detection using multi-channel attention 3D-CNN for hospitalized infants. *Quantitative Imaging in Medicine and Surgery (accepted)*, 2021. (Cited on pages 14 and 161.)

[140] Yue Sun, Deedee Kommers, Tao Tan, Wenjin Wang, Xi Long, Caifeng Shan, Carola van Pul, Ronald M. Aarts, Peter Andriessen, and Peter H. N. de With. Automated discomfort detection for premature infants in NICU using time-frequency feature-images and CNNs. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 113144B. International Society for Optics and Photonics, 2020. (Cited on pages 14 and 163.)

[141] Yue Sun, Deedee Kommers, Wenjin Wang, Rohan Joshi, Caifeng Shan, Tao Tan, Ronald M. Aarts, Carola van Pul, Peter Andriessen, and Peter H. N. de With. Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5995–5999. IEEE, 2019. (Cited on pages 13, 86, 94, 126, and 163.)

[142] Yue Sun, Deedee Kommers, Wenjin Wang, Rohan Joshi, Caifeng Shan, Tao Tan, Ronald M. Aarts, Carola van Pul, Peter Andriessen, and Peter H. N. de With. Video-based discomfort monitoring for premature infants in NICU. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, 2019. (Cited on page 163.)

[143] Yue Sun, Wu Qiu, Cesare Romagnoli, and Aaron Fenster. 3D non-rigid surface-based MR-TRUS registration for image-guided prostate biopsy. In *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 9036, page 90362J. International Society for Optics and Photonics, 2014. (Cited on page 164.)

[144] Yue Sun, Wu Qiu, Jing Yuan, Cesare Romagnoli, and Aaron Fenster. Three-dimensional nonrigid landmark-based magnetic resonance to transrectal ultrasound registration for image-guided prostate biopsy. *Journal of Medical Imaging*, 2(2):025002, 2015. (Cited on page 162.)

[145] Yue Sun, Caifeng Shan, Tao Tan, Xi Long, Arash Pourtaherian, Svitlana Zinger, and Peter H. N. de With. Video-based discomfort detection for infants. *Machine Vision and Applications*, 30(5):933–944, 2019. (Cited on pages 13, 126, and 161.)

[146] Yue Sun, Caifeng Shan, Tao Tan, Xi Long, Arash Pourtaherian, Svitlana Zinger, and Peter H. N. With. Video-based discomfort detection for infants. *Machine Vision and Applications*, page 1, August 2018. (Cited on pages 42, 43, 47, 59, 61, and 64.)

[147] Yue Sun, Caifeng Shan, Tao Tan, Tong Tong, Wenjin Wang, Arash Pourtaherian, and Peter H. N. de With. Detecting discomfort in infants through facial expressions. *Physiological Measurement*, 40(11):115006, 2019. (Cited on pages 13, 93, and 161.)

[148] Yue Sun, Caifeng Shan, Tao Tan, Tong Tong, Wenjin Wang, Arash Pourtaherian, and Peter H. N. de With. Detecting discomfort in infants through facial expressions. *Physiological Measurement*, 2019. (Cited on page 126.)

[149] Yue Sun, Wenjin Wang, Xi Long, Mohammed Meftah, Tao Tan, Caifeng Shan, Ronald M. Aarts, and Peter H. N. de With. Respiration monitoring for premature neonates in NICU. *Applied Sciences*, 9(23):5246, 2019. (Cited on pages 14 and 161.)

[150] Yue Sun, Jing Yuan, Wu Qiu, Martin Rajchl, Cesare Romagnoli, and Aaron Fenster. Three-dimensional nonrigid MR-TRUS registration using dual optimization. *IEEE transactions on medical imaging*, 34(5):1085–1095, 2014. (Cited on page 162.)

[151] Yue Sun, Jing Yuan, Martin Rajchl, Wu Qiu, Cesare Romagnoli, and Aaron Fenster. Efficient convex optimization approach to 3D non-rigid MR-TRUS registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 195–202. Springer, 2013. (Cited on page 164.)

[152] Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based discomfort detection for premature infants. In *Medical Image Understanding and Analysis (MIUA)*, Edinburgh, UK, 2017. (Cited on page 163.)

[153] Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based monitoring and assessing discomfort in premature infants. In *6th Dutch Bio-Medical Engineering Conference*, Egmond aan Zee, The Netherlands, 2017. (Cited on page 165.)

[154] Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based discomfort detection for premature infants. In *11th Biomedica Summit*, Eindhoven, the Netherlands, 2017. (Cited on page 165.)

[155] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. (Cited on page 16.)

[156] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. (Cited on page 49.)

[157] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. (Cited on page 45.)

[158] Tao Tan, Zhang Li, Haixia Liu, Ping Liu, Wenfang Tang, Hui Li, Yue Sun, Yusheng Yan, Keyu Li, Tao Xu, et al. Optimize transfer learning for lung diseases in bronchoscopy using a new concept: sequential fine-tuning. *arXiv preprint arXiv:1802.03617*, 2018. (Cited on page 49.)

[159] AS Tolba, AH El-Baz, and AA El-Harby. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103, 2006. (Cited on page 19.)

[160] Md Zia Uddin, Weria Khaksar, and Jim Torresen. Facial expression recognition using salient features and convolutional neural network. *IEEE Access*, 5:26146–26161, 2017. (Cited on page 93.)

[161] Beatriz O Valeri, Liisa Holsti, and Maria BM Linhares. Neonatal pain and developmental outcomes in children born preterm: a systematic review. *The Clinical journal of pain*, 31(4):355–362, 2015. (Cited on page 2.)

[162] Ron Van Luijtelaar, Wenjin Wang, Sander Stuijk, and Gerard de Haan. Automatic roi detection for camera-based pulse-rate measurement. In *Asian Conference on Computer Vision*, pages 360–374. Springer, 2014. (Cited on page 126.)

[163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. (Cited on page 135.)

[164] KV Veena and Dominic Mathew. Speaker identification and verification of noisy speech using multitaper mfcc and gaussian mixture models. In *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–4. IEEE, 2015. (Cited on page 82.)

[165] Mauricio Villarroel, Alessandro Guazzi, João Jorge, Sara Davis, Peter Watkinson, Gabrielle Green, Asha Shenvi, Kenny McCormick, and Lionel Tarassenko. Continuous non-contact vital sign monitoring in neonatal intensive care unit. *Healthcare Technology Letters*, 1(3):87–91, 2014. (Cited on page 77.)

[166] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. (Cited on page 17.)

[167] Ngoc-Son Vu and Alice Caplier. Face recognition with patterns of oriented edge magnitudes. In *European conference on computer vision*, pages 313–326. Springer, 2010. (Cited on page 24.)

[168] Feng Wang, Xiang Xiang, Chang Liu, Trac D Tran, Austin Reiter, Gregory D Hager, Harry Quon, Jian Cheng, and Alan L Yuille. Regularizing face verification nets for pain intensity regression. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 1087–1091. IEEE, 2017. (Cited on pages 45 and 51.)

[169] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800–807, 2012. (Cited on page 19.)

[170] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Robust heart rate from fitness videos. *Physiological measurement*, 38(6):1023, 2017. (Cited on page 126.)

[171] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proc. IEEE Int. Conf. Computer Vision*, pages 1385–1392, December 2013. (Cited on pages 113, 117, and 118.)

[172] Jan Werth, Louis Atallah, Peter Andriessen, Xi Long, Elly Zwartkruis-Pelgrim, and Ronald M. Aarts. Unobtrusive sleep state measurements in preterm infants–a review. *Sleep medicine reviews*, 32:109–122, 2017. (Cited on page 114.)

[173] R Whit Hall and KJS Anand. Short-and long-term impact of neonatal pain and stress. *NeoReviews*, 6:69–75, 2005. (Cited on page 2.)

[174] Ying Wu and Thomas S Huang. Vision-based gesture recognition: A review. In *International Gesture Workshop*, pages 103–115. Springer, 1999. (Cited on page 126.)

[175] Jie Xing, Chao Chen, Qinyang Lu, Xun Cai, Aijun Yu, Yi Xu, Xiaoling Xia, Yue Sun, Jing Xiao, and Lingyun Huang. Using bi-rads stratifications as auxiliary information for breast masses classification in ultrasound images. *IEEE Journal of Biomedical and Health Informatics (accepted)*, 2020. (Cited on page 161.)

[176] Xiaojing Xu, Kenneth D Craig, Damaris Diaz, Matthew S Goodwin, Murat Akcakaya, Büşra Tuğçe Susam, Jeannie S Huang, and Virginia R de Sa. Automated pain detection in facial videos of children using human-assisted transfer learning. In *International Workshop on Artificial Intelligence in Health*, pages 162–180. Springer, 2018. (Cited on page 42.)

[177] Xin Yang, Mingyue Ding, Liantang Lou, Ming Yuchi, Qiu Wu, and Yue Sun. Common carotid artery lumen segmentation in b-mode ultrasound transverse view images. *International Journal of Image, Graphics and Signal Processing*, 3(5):15, 2011. (Cited on page 163.)

[178] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Eprint Arxiv*, 27:3320–3328, 2014. (Cited on page 46.)

[179] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proc. BMVC*, pages 1–12, 2019. (Cited on page 97.)

[180] Jing Yuan, Wu Qiu, Eranga Ukwatta, Martin Rajchl, Yue Sun, and Aaron Fenster. An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior. *Prostate MR Image Segmentation Challenge, MICCAI*, 7512:82–89, 2012. (Cited on page 163.)

[181] Jing Yuan, Eranga Ukwatta, Wu Qiu, Martin Rajchl, Yue Sun, Xue-Cheng Tai, and Aaron Fenster. Jointly segmenting prostate zones in 3D mris by globally optimized coupled level-sets. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 12–25. Springer, 2013. (Cited on page 164.)

[182] Ghada Zamzami, Gabriel Ruiz, Dmitry Goldgof, Rangachar Kasturi, Yu Sun, and Terri Ashmeade. Pain assessment in infants: Towards spotting pain expression based on infants' facial strain. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 5, pages 1–5. IEEE, 2015. (Cited on page 70.)

[183] Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof, Rangachar Kasturi, Terri Ashmeade, and Yu Sun. An approach for automated multimodal analysis of infants' pain. In *23rd International Conference on Pattern Recognition (ICPR 2016)*, volume 3, pages 4148–4153, 2016. (Cited on page 19.)

[184] Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof, Rangachar Kasturi, Terri Ashmeade, and Yu Sun. An approach for automated multimodal analysis of infants' pain. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4148–4153. IEEE, 2016. (Cited on page 77.)

[185] Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof, Rangachar Kasturi, Yu Sun, and Terri Ashmeade. Automated pain assessment in neonates. In *20th Scandinavian Conference on Image Analysis (SCIA 2017)*, volume 3, pages 350–361, 2017. (Cited on page 19.)

[186] Baochang Zhang, Wankou Yang, Ze Wang, Lian Zhuo, Jungong Han, and Xiantong Zhen. The structure transfer machine theory and applications. *IEEE Transactions on Image Processing*, 29:2889–2902, 2019. (Cited on page 94.)

[187] Dan Zhang, Fan Huang, Maziyar Khansari, Tos TJM Berendschot, Xiayu Xu, Behdad Dashtbozorg, Yue Sun, Jiong Zhang, and Tao Tan. Automatic corneal nerve fiber segmentation and geometric biomarker quantification. *The European Physical Journal Plus*, 135(2):1–16, 2020. (Cited on page 162.)

[188] J. Zhao, J. Han, and L. Shao. Unconstrained face recognition using a set-to-set distance measure on deep learned features. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2679–2689, Oct 2018. (Cited on page 44.)

[189] Jianfeng Zhao, Xia Mao, and Jian Zhang. Learning deep facial expression features from image and optical flow sequences using 3d cnn. *The Visual Computer*, 34(10):1461–1475, 2018. (Cited on page 94.)

[190] Jiaojiao Zhao, Jungong Han, and Ling Shao. Unconstrained face recognition using a set-to-set distance measure on deep learned features. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2679–2689, 2017. (Cited on page 94.)

[191] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3466–3474, June 2016. (Cited on pages 55 and 56.)

[192] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3474, 2016. (Cited on page 51.)

[193] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. (Cited on page 59.)

[194] J. Zhou, X. Hong, F. Su, and G. Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1535–1543, June 2016. (Cited on pages 55 and 56.)

[195] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 84–92, 2016. (Cited on page 51.)

Publication list

Journals for This Thesis

- Yue Sun, Jingjing Hu, Wenjin Wang, Min He, and Peter H. N. de With. Camera-based discomfort detection using multi-channel attention 3D-CNN for hospitalized infants. *Quantitative Imaging in Medicine and Surgery (accepted)*, 2021

- Yue Sun, Wenjin Wang, Xi Long, Mohammed Meftah, Tao Tan, Caifeng Shan, Ronald M. Aarts, and Peter H. N. de With. Respiration monitoring for premature neonates in NICU. *Applied Sciences*, 9(23):5246, 2019

- Yue Sun, Caifeng Shan, Tao Tan, Tong Tong, Wenjin Wang, Arash Pourtaherian, and Peter H. N. de With. Detecting discomfort in infants through facial expressions. *Physiological Measurement*, 40(11):115006, 2019

- Yue Sun, Caifeng Shan, Tao Tan, Xi Long, Arash Pourtaherian, Svitlana Zinger, and Peter H. N. de With. Video-based discomfort detection for infants. *Machine Vision and Applications*, 30(5):933–944, 2019

Other Journals

- Jie Xing, Chao Chen, Qinyang Lu, Xun Cai, Aijun Yu, Yi Xu, Xiaoling Xia, Yue Sun, Jing Xiao, and Lingyun Huang. Using bi-rads stratifications as auxiliary information for breast masses classification in ultrasound images. *IEEE Journal of Biomedical and Health Informatics (accepted)*, 2020

- Zhang Li, Zheng Zhong, Yang Li, Tianyu Zhang, Liangxin Gao, Dakai Jin, Yue Sun, Xianghua Ye, Li Yu, Zheyu Hu, et al. From community acquired pneumonia to covid-19: A deep learning based method for quanti-

tative analysis of covid-19 on thick-section CT scans. *European Radiology*, 30(12):6828–6837, 2020

- Dan Zhang, Fan Huang, Maziyar Khansari, Tos TJM Berendschot, Xiayu Xu, Behdad Dashtbozorg, Yue Sun, Jiong Zhang, and Tao Tan. Automatic corneal nerve fiber segmentation and geometric biomarker quantification. *The European Physical Journal Plus*, 135(2):1–16, 2020

- Dongsheng Jiang, Weiqiang Dou, Luc Vosters, Xiayu Xu, Yue Sun, and Tao Tan. Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. *Japanese journal of radiology*, 36(9):566–574, 2018

- Zhang Li, Fan Huang, Jiong Zhang, Behdad Dashtbozorg, Samaneh Abbasi-Sureshjani, Yue Sun, Xi Long, Qifeng Yu, Bart ter Haar Romeny, and Tao Tan. Multi-modal and multi-vendor retina image registration. *Biomedical optics express*, 9(2):410–422, 2018

- Yue Sun, Wu Qiu, Jing Yuan, Cesare Romagnoli, and Aaron Fenster. Three-dimensional nonrigid landmark-based magnetic resonance to transrectal ultrasound registration for image-guided prostate biopsy. *Journal of Medical Imaging*, 2(2):025002, 2015

- Siavash Khallaghi, C Antonio Sánchez, Abtin Rasoulian, Yue Sun, Farhad Imani, Amir Khojaste, Orcun Goksel, Cesare Romagnoli, Hamidreza Abdi, Silvia Chang, et al. Biomechanically constrained surface registration: Application to MR-TRUS fusion for prostate interventions. *IEEE transactions on medical imaging*, 34(11):2404–2414, 2015

- Yue Sun, Jing Yuan, Wu Qiu, Martin Rajchl, Cesare Romagnoli, and Aaron Fenster. Three-dimensional nonrigid MR-TRUS registration using dual optimization. *IEEE transactions on medical imaging*, 34(5):1085–1095, 2014

- Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Prostate segmentation: an efficient convex optimization approach with axial symmetry using 3-D TRUS and MR images. *IEEE transactions on medical imaging*, 33(4):947–960, 2014

- Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Dual optimization based prostate zonal segmentation in 3D MR images. *Medical image analysis*, 18(4):660–673, 2014

- Jing Yuan, Wu Qiu, Eranga Ukwatta, Martin Rajchl, Yue Sun, and Aaron Fenster. An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior. *Prostate MR Image Segmentation Challenge, MICCAI*, 7512:82–89, 2012

- Xin Yang, Mingyue Ding, Liantang Lou, Ming Yuchi, Qiu Wu, and Yue Sun. Common carotid artery lumen segmentation in b-mode ultrasound transverse view images. *International Journal of Image, Graphics and Signal Processing*, 3(5):15, 2011

International Conference Contributions for This Thesis

- Lan Min, Yue Sun, and Peter H. N. de With. Video-based infant monitoring using a CNN-LSTM scheme. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 1159717. International Society for Optics and Photonics, 2021

- Yue Sun, Deedee Kommers, Tao Tan, Wenjin Wang, Xi Long, Caifeng Shan, Carola van Pul, Ronald M. Aarts, Peter Andriessen, and Peter H. N. de With. Automated discomfort detection for premature infants in NICU using time-frequency feature-images and CNNs. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 113144B. International Society for Optics and Photonics, 2020

- Yue Sun, Deedee Kommers, Wenjin Wang, Rohan Joshi, Caifeng Shan, Tao Tan, Ronald M. Aarts, Carola van Pul, Peter Andriessen, and Peter H. N. de With. Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5995–5999. IEEE, 2019

- Yue Sun, Deedee Kommers, Wenjin Wang, Rohan Joshi, Caifeng Shan, Tao Tan, Ronald M. Aarts, Carola van Pul, Peter Andriessen, and Peter H. N. de With. Video-based discomfort monitoring for premature infants in NICU. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, 2019

- Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based discomfort detection for premature infants. In *Medical Image Understanding and Analysis (MIUA)*, Edinburgh, UK, 2017

Other International Conference Contributions

- Bahram Marami, Shahin Sirouspour, Suha Ghoul, Shadi Emami Abarghouei, Yue Sun, and Aaron Fenster. Non-rigid MRI-TRUS registration in targeted prostate biopsy. In *Medical Imaging 2015: Image Processing*, volume 9413, page 941332. International Society for Optics and Photonics, 2015

- Yue Sun, Wu Qiu, Cesare Romagnoli, and Aaron Fenster. 3D non-rigid surface-based MR-TRUS registration for image-guided prostate biopsy. In *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 9036, page 90362J. International Society for Optics and Photonics, 2014

- Wu Qiu, Martin Rajchl, Fumin Guo, Yue Sun, Eranga Ukwatta, Aaron Fenster, and Jing Yuan. 3D prostate TRUS segmentation using globally optimized volume-preserving prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 796–803. Springer, 2014

- Yue Sun, Jing Yuan, Martin Rajchl, Wu Qiu, Cesare Romagnoli, and Aaron Fenster. Efficient convex optimization approach to 3D non-rigid MR-TRUS registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 195–202. Springer, 2013

- Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Efficient 3D multi-region prostate MRI segmentation using dual optimization. In *International Conference on Information Processing in Medical Imaging*, pages 304–315. Springer, 2013

- Wu Qiu, Jing Yuan, Eranga Ukwatta, Yue Sun, Martin Rajchl, and Aaron Fenster. Fast globally optimal segmentation of 3d prostate mri with axial symmetry prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 198–205. Springer, 2013

- Jing Yuan, Eranga Ukwatta, Wu Qiu, Martin Rajchl, Yue Sun, Xue-Cheng Tai, and Aaron Fenster. Jointly segmenting prostate zones in 3D mris by globally optimized coupled level-sets. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 12–25. Springer, 2013

Regional Conference Contributions

- Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based discomfort detection for premature infants. In *11th Biomedica Summit*, Eindhoven, the Netherlands, 2017
- Yue Sun, Svitlana Zinger, Sidarto Bambang Oetomo, and Peter H. N. de With. Video-based monitoring and assessing discomfort in premature infants. In *6th Dutch Bio-Medical Engineering Conference*, Egmond aan Zee, The Netherlands, 2017

Acronyms

AAM Active Appearance Model

AI Artificial Intelligence

AUC Area Under the Curve

AUs Action Units

CC Cross-Correlation

ChI Chest Impedance

CI Confidence Interval

CNN Convolutional Neural Network

CV Computer Vision

DCT Discrete Cosine Transform

ECG Electrocardiogram

FACS Facial Action Coding System

HOG Histogram of Oriented Gradients

HP Heel Prick

HR Heart Rate

HRV Heart Rate Variability

LBPs Local Binary Patterns

LMSpec Log Mel-Spectrogram

MAE Mean Absolute Error

MFCCs Mel-Frequency Cepstral Coefficients

MSE Mean Squared Error

NICU Neonatal Intensive Care Unit

PCA Principal Component Analysis

PCC Pearson Correlation Coefficient

PFP Primal Face of Pain

RBF Radial Basis Function

RMSE Root Mean-Squared Error

ROC Receiver Operating Characteristic

ROI Region Of Interest

RR Respiratory Rate

SSCF Spectral Subband Centroid Frequency

SVM Support Vector Machine

Acknowledgments

First and foremost, Prof.dr.ir. Peter H.N. de With, beste Peter, I would like to thank you for opening the door of VCA to me, as well as for the support, guidance, and confidence in me. The scene when I first met you for my interview is still so vivid to me. Your explanation of the baby monitoring project impressed me, and I started to believe that we will be the pioneer in this field since then. Although several other medical imaging groups in the Netherlands were approaching me at the same time, I made a firm decision to pursue my PhD degree at VCA studying the topic of video-based baby monitoring. You provided me excellent supervision. Thanks to your knowledge on the topic, your ability to seeing things from an adequate perceptive. My personal and professional skills have improved substantially. What I learned from you was not only academic knowledge but a positive attitude and a passion for science and technology. No matter how busy your schedule was, you always made time to answer my questions, helped me explore ideas, reviewed my papers even during holidays, and afforded me endless patience and encouragement. To me, it is a treasure for a lifetime to be one of your students and to be a part of De With group.

Next, I would express my gratitude to my co-promotor Prof.dr. Ronald M. Aarts, for the patient guidance, encouragement and advice you have provided throughout my PhD time. You have enhanced my academic vision, and thank you for always responding to my questions and queries promptly.

I would like to sincerely thank Prof.dr. Caifeng Shan for the guidance and support and for the time Dr. Shan spent helping me prepare papers. Your contribution has increased the quality of my research and thesis significantly. I would like to acknowledge Dr. Sveta Zinger. Dear Sveta, thank you for getting me through the door of this project and for the strong support of this project.

I would like to thank Deedee for bringing me to the real NICU and allowing me to observe the babies by their sides. Thank you for your strong medical background and insight. You helped me understand the real clinical problems that the infants are experiencing, and that is exactly where they need us. Thank you Rohan for generating the great idea for identifying discomfort labels and connecting me to the great MMC people. Thank you, Carola and Peter (Andriessen) for the exciting collaboration. We are working hard together for a new baby-monitoring generation, in the past, present, and to be continued. Thank all the medical staff from the NICU department of Máxima Medical Center, Veldhoven for the cooperation and assistance in collecting the infant videos. I enjoyed the time that I got up in the early mornings to join your standup meetings for the morning shift and then of course the numerous heel prick procedures.

I would like to thank Dr. Arash Pourtaherian, Dr. Wenjing Wang, and Dr. Xi Long for their strong support and help on this project. Xi, Wenjin, and Caifeng have introduced me to Philips and invited me as a researcher at Philips. I would like to thank Ihor and Elly for hosting me in Philips and allowing me to perform scientific research in an industry setting.

I would like to thank all the VCA members. We together have enjoyed the time spent at VCA, conference trips, and social events. Thank lab members including but not limited to Xin, Cheng, Francesca, Hongxu, Panos, Anweshan, Hani, Chenyang, Patrick, Tim, Liang, Farhad, Amir, Marco, and Marco, etc. Thank Xin for sharing your nice home-cooked food. Thank you Cheng for being a teammate. Thank Lan Min for working together. As a mentor for your Master's graduation project, you make my PhD's life a wonderful experience. Thank you Anja for your countless help with patience.

Last but not least, I would like to thank my family and friends. Thank all the new friends I met in the Netherlands, Yang, Qi, Jiankang, Matt, Hongchao, Yuan, Ruisheng, Luc, Xiong, Lin, Bin, Xin, Qinghua, etc. Also thank my friends in Canada, the US, and China. Time passes, things change, but our minds are always in sync. It is my honor to have you in my life.

I have to thank my husband, the love of my life, Tao, for keeping things going and for always showing how proud he is of me.

Thank my parents for your love, for inspiring me to follow my dreams and being by my side throughout my journey.

Curriculum Vitae

Yue Sun received her BSc degree in Biomedical Engineering from Huazhong University of Science and Technology, Wuhan, China in 2010. She worked as a research assistant at Robarts Research Institute, London, Ontario, Canada from 2011 to 2014. In 2014, she completed her MSc. in Biomedical Engineering at the University of Western Ontario, London, Ontario, Canada. She worked as a software developer from 2014 to 2015 at Center for Imaging Technology Commercialization (CIMTEC), London, Ontario, Canada. She moved to the Netherlands in 2016 and started working as a PhD researcher in the Signal Processing Systems, Video Coding and Architectures research group at Eindhoven University of Technology, Eindhoven, the Netherlands. From June 2020, she started contributing to the IMAGEN project as a postdoc researcher by investigating novel deep learning based methods for animal detection/tracking and behavior analysis. Her research interests include video/image processing, machine learning, medical image analysis, and facial expression and behavior analysis.