GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS, TELECOMMUNICATIONS AND INFORMATICS

The author of the PhD dissertation: Alicja Kwaśniewska
Scientific discipline: Biomedical Engineering

DOCTORAL DISSERTATION

Title of PhD dissertation: Thermal Images Analysis Methods using Deep Learning Techniques for the Needs of Remote Medical Diagnostics

Title of PhD dissertation (in Polish): Metody analizy obrazów termograficznych z wykorzystaniem technik uczenia głębokiego dla potrzeb zdalnej diagnostyki medycznej

Supervisor *signature*
Jacek Rumiński, Ph. D., D. Sc., E. Eng., Assoc. Prof.

Gdańsk, year 2020

Abstract

Remote medical diagnostic solutions have recently gained more importance due to global demographic shifts and play a key role in evaluation of health status during epidemic. Contactless estimation of vital signs with image processing techniques is especially important since it allows for obtaining health status without the use of additional sensors. Thermography enables us to reveal additional details, imperceptible in images acquired with standard visible light cameras, yet, low resolution is its significant limitation. In the presented doctoral dissertation, original artificial intelligence solutions were proposed based on performed analysis of innovative thermal image processing methods using Deep Learning techniques for the needs of remote medical diagnostics. Possibility of modifying architecture of deep neural network designed for classification of visible light images in such a way that distribution of extracted features will be recreated enabling detection of facial areas from low resolution thermal data was verified in conducted experiments. Effectiveness of the proposed deep neural network architecture was demonstrated in practical applications, increasing resolution of thermal images and leading to better image quality metrics in comparison to state-of-the-art convolutional models, as well as increasing accuracy of facial areas detection, contactless estimation of respiratory rate and person recognition.

Streszczenie

Rozwiązania zdalnej diagnostyki medycznej zyskują na znaczeniu w świetle globalnych przemian demograficznych, a także pełnią istotną rolę w ocenie stanu zdrowia podczas epidemii. Bezkontaktowy pomiar parametrów życiowych z wykorzystaniem przetwarzania obrazów jest w szczególności istotny z uwagi na możliwość uzyskania informacji o stanie zdrowia bez użycia dodatkowych sensorów. Termografia pozwala na pozyskanie danych niedostępnych przy użyciu kamer światła widzialnego, jednakże jej istotnym ograniczeniem jest niska rozdzielczość rejestrowanych danych. W ramach rozprawy zaproponowano autorskie rozwiązania sztucznej inteligencji na podstawie dokonanej analizy innowacyjnych metod przetwarzania obrazow termograficznych z wykorzystaniem technik uczenia głebokiego dla potrzeb zdalnej diagnostyki medycznej. W przeprowadzonych badaniach zweryfikowano możliwość zmodyfikowania architektury głębokiej sieci neuronowej przeznaczonej do klasyfikacji obrazow uzyskanych w promieniowaniu widzialnym w celu odtworzenia rozkładu wydobytych cech i tym samym umożliwienia detekcji obszarow twarzy z niskiej rozdzielczosci obrazow termograficznych. Skuteczność zaproponowanej architektury głebokiej sieci neuronowej została potwierdzona w praktycznych aplikacjach, umożliwiając zwiększenie rozdzielczości obrazów termograficznych oraz ulepszenie metryk jakości obrazu w porównaniu ze stanem wiedzy w zakresie modeli konwolucyjnych, a także poprawiając dokładność detekcji obszarów twarzy, bezkontaktowego pomiaru częstości oddychania oraz rozpoznawania osób.

Thank You

I would like to express my sincere thanks, deepest appreciation and gratitude to my Supervisor, Jacek Rumiński, Ph. D., D. Sc., E. Eng., Assoc. Prof., for his unwavering guidance, encouragement, support and constructive criticism throughout my studies.

I am extremely grateful to Professor Jerzy Wtorek, Ph. D., D. Sc., E. Eng., Assoc. Prof., Professor Antoni Nowakowski, Ph. D., D. Sc., E. Eng., all other professors, employees and colleagues at Biomedical Engineering Department, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology for igniting the spark which conceived my passion to Biomedical Engineering. I am deeply indebted to everyone, who provided insight and expertise that greatly assisted the study with special thanks to Professor Paul Rad, Ph. D., Assoc. Prof., for an opportunity to pursue part of my research at the University of Texas San Antonio, Open Cloud Institute and acting as my External Advisor during that time.

Furthermore, this dissertation would not be possible without constant support of my Parents, Grandparents, Gosia, Krzysiek, and the rest of family, who were always there for me, encouraging to always be eager to learn and go one step further. Special thanks to Maciek for his love and profound belief in my abilities. I also very much appreciate all heartening words provided by my friends when I needed them most.

Finally, I would love to acknowledge the assistance of Ida, who joined us during final stages of my studies, giving me unlimited happiness and strength.

Thank you! This dissertation would not be complete without you.

Acknowledgement

This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology; NCBiR, FWF, SNSF, ANR and FNRS in the framework of the ERA-NET CHIST-ERA II, project eGLASSES — The interactive eyeglasses for mobile, perceptual computing and Intel Corporation, USA.

Abbreviations

AI Artificial Intelligence. 17, 23, 77, 113, 149

CNN Convolutional Neural Network. 17, 36, 54, 81, 114, 116, 150

DL Deep Learning. 17, 23, 43, 53, 83, 113, 149, 157

DNN Deep Neural Network. 17, 20, 35, 43, 54, 79, 114, 116, 151, 157

DRCN Deeply Recursive Convolutional Network. 84, 151

DRESNet Deeply Residual Embedding and Supervised-recursion. 90, 113, 150, 158

DRRN Deep Recursive Residual Network. 84, 122, 151

EVM Eulerian Video Magnification. 125

GAN Generative Adversarial Network. 20, 84, 119, 154

HR High Resolution. 80, 117

IoU Intersection over Union. 41, 65, 116, 118, 120, 180

LR Low Resolution. 80, 117

PSNR Peak Signal-to-Noise Ratio. 20, 79, 82, 86, 103, 116, 120, 123, 124, 153, 180, 183

ResNet Residual Network. 36, 84

RoI Region of Interest. 29, 39, 79, 115, 124, 140

RR Respiratory Rate. 113

SISR Single Image Super Resolution. 80, 150

SR Super Resolution. 21, 77, 79, 80, 116, 149

SRCNN Super Resolution Convolutional Neural Network. 83

SSD Single Shot Detector. 39, 53, 116, 158

SSIM Structural Similarity Index Metric. 79, 82, 103, 116, 124, 153, 183

Nomenclature

D The number of recursions in non-linear mapping subnetwork. 89

E The number of residuals in feature extraction subnetwork. 89

FE The feature extraction subnetwork. 83

NLM The non-linear mapping subnetwork. 83

R The reconstruction subnetwork. 83

U The number of residuals in non-linear mapping subnetwork. 89

Contents

Chapter 1

Introduction

Recent advances in technologies and increased self-awareness of societies has revolutionized current healthcare definition. Modern medical systems are expected to support and cover various subsectors of medicine, e.g. computer-aided diagnosis, remote monitoring, therapy support, healthy lifestyle tracking, security, and more. Hence, we can observe an increasing focus on Artificial Intelligence (AI) studies and many of possible medical use cases have benefited by progress made in Deep Neural Network (DNN) development, e.g. by supporting clinical decisions with predictions obtained from AI algorithms [1].

Since images have one of the biggest contribution to overall Big Data resources [2], the majority of solutions utilize computer vision and image processing algorithms. A key Deep Neural Network (DNN) architecture that led to a breakthrough in image recognition studies is based on convolution operations (Convolutional Neural Network (CNN)) [3] incorporating local connection patterns shared between different locations in two dimensional data. Examples of medical solutions that take advantage of CNN architecture are boundless, from breast cancer identification using mammograms [4] or thermograms [5], lung [6, 7] and colon [8] cancer detection, bone scans analysis [9], eye diseases detection [10], and more. AI is also widely used for improving quality and resolution of medical images, so that the diagnosis may be more accurate, e.g. in microscopy [11], Magnetic Resonance Imaging [12] or Computed Tomography [13]. AI-based medical market is expected to further expand, covering more and more conventional approaches used so far.

At the same time, some concerns about possibilities of replacing physicians and fully automating diagnostic procedures arise since Deep Learning (DL) models work as black-boxes making predictions that are difficult to justify. According to the research conducted by Ahuja A. [14], AI will rather support and augment current professional diagnostic instead of being its substitution. However, there are also other subsectors of medicine which allow for obtaining information about health status without supervision of professional diagnosticians and specialized acquisition devices.

Those subsectors are mainly concerned around telemedicine solutions which aim at providing health services at a distance. Such systems have mainly a support function, meaning that they are designed to increase self-awareness of people and provide basic solutions for monitoring of health status outside medical facilities. At the same time, they do not aim at replacing professional diagnostic procedures. Therefore, there is a chance that AI will be widely used in remote medical diagnostic sector.

1.1 Artificial Intelligence in Medicine

Back in 2003, the U.S. government has already been spending $26B on healthcare Research and Development and this growth is continuously progressing [15]. As presented in the market analysis performed by Accenture [16], key AI-based healthcare applications can potentially save $150 billion in annual healthcare expenses in the United States by 2026. Moreover, AI health market size is expected to grow by 40 percent, raising from $600M in 2014 to $6.6B in 2021. This growth is already visible in expansion of medical startups. The number of AI healthcare deals has grown more than 3 times from 2012 to mid-2016. Similar trend can be also observed in other countries, especially those that are dealing with the problem of super-aged societies. A recent report conducted by Global Market Insights shows that the global telemedicine market will expand from $38.3B valuation to $130.5B by 2025 [17]. Indian and China telemedicine market is predicted to grow at 22.4% and 23%, respectively. Accenture research also specifies AI use cases that will lead to the best near-term gains based on their application, likelihood of adoption and value to the health economy [16].

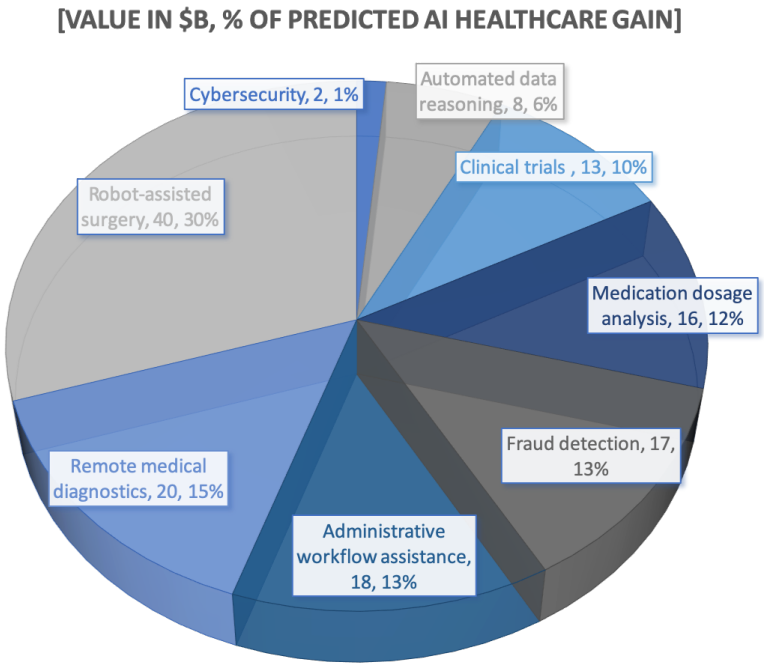


Figure 1.1. Participation of different AI use cases in the total healthcare near-term impact

The chart presented in Fig. 1.1 shows near-term value of these applications. $8B of future AI value in the healthcare sector is represented by solutions focused on automated data reasoning, creating a huge demand for innovative computer vision algorithms. What's more, remote medical diagnostics is projected to generate a second highest gain in the healthcare market in next few years. Thus, we are mainly interested how AI can transform this sector and whether it can ease accessing accurate information about primary health indicators without supervision to lessen a burden on medical professionals. Specifically, we believe there is a need of expanding research on remote diagnostics solutions to the thermal domain, as it can provide additional medical information (temperature patterns can be used for pain analysis [18], sleep detection [19], evaluation of facial muscle paralysis [20] or respiratory rate estimation [21]), while being insensitive to different illumination condition and ensuring better privacy [22] than visible light data.

1.2 Artificial Intelligence for Thermal Image Processing

Although development of image processing algorithms in visible light domain is more advanced than in thermography, some studies targeting remote medical diagnostics using thermal data have already been conducted. Proposed machine learning-based approaches address e.g. breast cancer detection [23], face expression recognition [24], vital signs monitoring [25], and other applications.

Most of remote medical diagnostic studies require detection of human body parts and faces in order to analyse changes present in those regions. Some accurate image processing algorithms for thermal face segmentation have already been proposed in studies conducted by Marzec M. et al. [26, 27, 28]. The work presented in [26] focused on detecting facial areas using anthropometric measurements and relations between facial regions, while mitigating common problems occurring in face segmentation tasks, e.g. various body poses, different scales and influence of background objects. In later studies by Marzec M. et al. [27], the proposed facial pattern was further improved to support various positions of eye regions and thereby increasing robustness to head rotations. Bertozzi M. et al. [29] proposed to perform body detection by comparing prepared human model with candidates generated as warm symmetrical objects, filtered from false positives using shape information. Template-based techniques was also explored in [30] where persons were detected using edge contour maps and Sobel filtering. Human shape information was further exploited in study conducted by Wang W. et al. [30]. It has also been shown that valuable information for locating objects in thermal data can be obtained from interest point detectors [31], Haar features [32] or histograms of oriented gradient [33]. Body and face areas detection problem can also be solved using machine learning algorithms. Human head proportions and neural networks were utilized for facial areas segmentation in research conducted by Koprowski R. et al. [34]. Later, in the study introduced by Marzec M. et al. [35], neural network was applied to classify eyes regions pre-designated using information about brightness distribution (warmer spots in eye corners).

Despite producing satisfactory results, mentioned algorithms have one main disadvantage. They require specified sets of features that characterize different classes of objects. Then, decisions are made by correlating each feature with expected outcomes. Yet, an exact collection of features is frequently hard to define and can change over time or vary depending on personal or environment factors. As a result, machine learning solutions based on hand-crafted features work best in strictly defined conditions and can't generalize well to changing environments. Since remote medical diagnostics are performed without supervision, it is especially important to ensure that they are robust to outliers in feature representations associated with various problems that may occur in thermal imaging. Examples of some potential factors that may have influence of system accuracy include lack of color and texture information that may impact visibility of features significant for predictions, presence of thermal radiation reflections, low resolution leading to small spatial sizes of interesting components, thus their similarity, and others.

DL mitigates some accuracy issues caused by different factors impacting descriptors of human detection by providing ability to automatically learn representations given a set of training samples instead of manually defining sets of features which describe various objects. As shown in [36], thermal features of a face can be extracted using Deep Boltzmann Machine, outperforming previous machine learning solutions based on hand-crafted features. Other DL architectures can also be utilized for thermal image processing, as presented in preliminary studies on pedestrian detection using convolutional model [37] or occupancy estimation with Recurrent Neural Network [38].

1.3 Limitations of Existing AI Solutions

Although DL has addressed various problems of previous, more conventional machine learning techniques, some limitations of existing neural networks are still valid for thermal image processing applications. First problem is a lack of publicly available thermal datasets. One of reasons for recent advances in Deep Neural Network (DNN) is Big Data trend [39], as providing models with enough data to learn proper dependencies is a key for achieving human-like accuracy in image processing tasks. This finding has been also confirmed for thermal images in a study presented by Kopaczka M. et al. [25], who showed that Deep Neural Networks work well in thermal spectrum, but only when appropriate training database is available. This problem can be solved by using an approach known as transfer learning, which allows for utilizing weights of models optimized using one set of images on a different one. Some attempts for making use of this approach in thermal images have already been done. Abbott R. et al. [40] proposed to utilize features extracted from high resolution thermal images in classification of low resolution thermal images. In this way, although lower quality data were more blurred and less detailed, it was still possible to achieve high prediction accuracy, as the network has already been aware of object characteristics that it learnt from images of higher resolution. Yet, to the best of our knowledge, detailed analysis of making use of visible light data features for thermal image processing hasn't been conducted yet. We believe that this could be beneficial, as much more samples are available for a visible light spectrum than for a thermal one [41].

Secondly, it is worth noting that thermal images have different representation of features than visible light data. Hence, a direct application of existing neural networks designed for extraction of high frequency features may not produce satisfactory results. To mitigate a problem of a gap between different imaging domains, Zhang T. et al. [42] introduced Generative Adversarial Network (GAN) aimed at translating images between domains. However, visible light data acquisition is a limitation for monitoring and diagnostic applications due to privacy concerns and thus translation between domains is not applicable in such cases.

Furthermore, as mentioned in [43], common thermal data acquisition devices suffer from some technical challenges, e.g. low resolution, poor Peak Signal-to-Noise Ratio (PSNR), reflections and halos around objects with significantly higher/lower temperature. All of those factors affect also shape information [30] leading to difficulties in classifying and detecting objects. Taking it into account, we see the need to design and evaluate algorithm for thermal image enhancement that would allow for eliminating those constraints and thereby increasing accuracy of remote medical diagnostics.

1.4 Goal and Thesis of the Presented Doctoral Dissertation

In the view of foregoing, the presented work aims at designing novel methods of thermal image processing using Deep Neural Networks in order to enhance their quality and thereby increase accuracy of facial areas detection for the needs of remote medical diagnostic solutions. It's important to note that the goal is not to outperform other conventional image processing techniques used for similar studies, but to verify applicability of DL to thermal image domain and propose innovative DNN architectures for thermal image processing. The presented doctoral dissertation constitutes an integrated synthesis derived from our studies published in a wide range of publications (specified in Appendix A) and expands them by additional experiments on thermal image processing with

DL. Following thesis were formulated as a part of research problems and evaluation conducted in this study:

I) Architecture of Deep Neural Network designed for classification of visible light images can be modified in such a way that distribution of extracted features will be recreated enabling detection of facial areas from low resolution thermal data.

II) Proposed architecture of Deep Convolutional Neural Networks allows for increasing resolution of thermal images leading to improvement of facial areas detection accuracy.

First thesis aims at verifying whether knowledge learnt for visible light data, characterized by a presence of high frequency features, can be utilized for processing of thermal images that are more blurred. In addition, we also want to evaluate if classification model, which produces a vector of high abstract features that are mapped to output categories, can be modified during the inference to restore features distribution. In this way, classification networks could be used for detection of objects (in our case facial areas) without a need to retrain them. All experiments proposed for verification of this thesis are conducted on low resolution thermal images in order to address possible scenarios of remote medical diagnostics, where usually only low-cost sensors are available.

The second thesis follows-up on the first one by mitigating the problem of dealing with low resolution data in telemedicine solutions. Specifically, we define criteria of neural network architecture that would allow for enhancing thermal images. In-depth analysis of different number and placement of blocks used in CNN will be performed in order to find configuration of a model which leads to the best image quality metrics for data restored from low resolution inputs.

Besides studies on proposed thesis, we perform additional experiments to evaluate whether introduced techniques and solutions could be applied in potential practical applications of remote medical diagnostics. Conducted benchmark evaluation will include different steps of end-to-end vital signs monitoring solutions, i.e. person recognition, facial areas detection, contactless estimation of vital signs and obtaining of other medically useful information from extracted vital signs patterns in order to determine whether proposed solutions can lead to better accuracy of the whole remote medical diagnostic pipeline. Defined thesis are further divided into detailed tasks, which after completion will support the specified goal of the presented dissertation:

1. Critical analysis of state-of-the-art methods applied to object feature extraction, image classification and areas detection with special focus on AI algorithms
2. Acquisition of thermal databases used for training and evaluation of examined NNs
3. Facial features detection
   (a) Evaluation of existing CNNs on collected datasets and verification of possibility to transfer knowledge learnt on visible light data to thermal images
   (b) Proposal and implementation of a novel flow in DL classification models enabling detection of facial areas from low resolution thermal data
   (c) Experimental analysis performed to compare obtained evaluation metrics for state-of-the-art DL solutions and the proposed model.
4. Thermal data resolution enhancement
   (a) Critical analysis of existing Super Resolution (SR) algorithms and their limitations for thermal image processing

(b) Proposal and implementation of a novel NN architecture designed specifically with thermal data characteristic in mind in order to enhance image resolution

(c) Experimental analysis performed to compare image quality and AI evaluation metrics for state-of-the-art DL solutions and the proposed model.

5. Proposal, design and evaluation of practical remote medical diagnostic applications based on thermal image processing that could benefit from introduced DL techniques and neural network architectures

6. Identification of future work directions and possible improvements of introduced solutions.

1.5 Organization of the Work

The rest of the work is organized into chapters addressing each of specified detailed tasks. Chapter 2 introduces algorithms used for object detection, with a special focus on analysis of thermal images of a face. Presented methods include techniques used for defining features acting as descriptors of objects, followed by conventional image processing and machine learning approaches, and finally more recent AI methods. In Chapter 3 we provide specifications of thermal datasets acquired for enhancing thermal data and detecting areas usable for non-contact vital signs estimation. All collected sets were obtained with possible scenarios of remote medical diagnostics in mind in order to verify robustness of proposed methods for such applications.

The task of facial feature detection from thermal images is studied by us in Chapter 4. At first, we evaluate performance of state-of-the-art neural networks on thermal data collected by us and conduct experiments with transferring knowledge from visible light domain to thermography. Then, we propose an innovative modification of deep classification models to restore features distribution and detect facial areas from thermal images, reducing latency and improving model accuracy. Finally, we provide details of experiments performed with novel DL architecture based on capsules, what results in generating a solution insensitive to different body poses.

Chapter 5 contains details of in-depth analysis of existing Super Resolution solutions and their applications, with the main focus on a thermal image domain. Moreover, a novel DNN architecture, designed by us with thermal features characteristic in mind is introduced and compared with other existing solutions on a wide set of thermal datasets, both publicly available and acquired by us. To the best of our knowledge, this is the first attempt to design CNN specifically dedicated to thermal image enhancement by addressing more distant dependencies between interesting components caused by the heat flow in objects.

The next Chapter (6) focuses on evaluation of proposed DL detection and thermal image enhancement models in possible non-contact vital signs monitoring applications. Various potential steps of remote medical diagnostics i.e. person recognition, facial areas detection, and extraction of vital signs are analyzed. In addition, we also perform further experiments with estimated vital signs by analysing influence of emotional response on calculated respiratory and heart rates. Ideas for future work and improvement of proposed algorithms are described in Chapter 7.

Finally, the work is concluded in Chapter 8, providing summary of achieved results with special focus on confirming and verifying completeness of theses formulated in the presented dissertation. Additionally, we also outline novel outcomes and innovative contribution to state-of-the-art techniques of thermal image processing for the needs of remote medical diagnostics.

Chapter 2

Detection of Facial Areas

2.1 Introduction and Overview

In medical applications, as well as other solutions that utilize image processing algorithms, it is crucial to distinguish and recognize various regions of the object present in images or video sequences. Some examples include breast cancer identification from mammograms [4] or thermograms [5], lung [6, 7] and colon [8] cancer diagnostic, or bone scans analysis [9]. Remote medical diagnostics can also benefit from such algorithms improving accuracy of e.g. processing of facial areas [44] for vital signs extraction. Many of the mentioned practical problems could be solved with detection and segmentation techniques that allow for identifying boundaries between image regions and objects as a whole.

In this chapter, we introduce approaches commonly used for object classification, detection and segmentation with a special focus on algorithms utilized for analysis of thermal images of a face. At first we describe conventional image processing techniques that can be used for defining boundaries between image regions, including algorithms based on differences among pixels present in neighbouring areas (e.g. points, lines, and edges), followed by methods that use similarities of image regions.

Then, we present more complex statistical and learning algorithms that take advantages of data distribution or correlations with other samples from training sets. Described methods that utilize various object features, e.g. defined manually, extracted with algorithms explained in the first section, and finally also learn as a part of applied Artificial Intelligence (AI) pipeline to perform data clustering, classification or detection. Specifically, we divide statistical and learning methods into two categories. The first one uses predefined features and patterns. The second one automatically extracts and adjusts feature representations by feeding algorithms with exemplary labelled images in the training phase.

Finally, we introduce metrics commonly used for evaluation of Deep Learning (DL) models including parameters utilized for classification, detection and inference performance. Results of tests performed in the rest of the work will be assessed based on those metrics in order to verify proposed dissertation theses.

2.2 Conventional Image Processing Techniques

2.2.1 Discontinuities among Pixel Values

One of the traditional ways of image analysis in order to separate objects present in a video frame (or a single image) is filtration and thresholding. Various studies have already been conducted in this area. We will focus mainly on techniques applied to face and facial areas detection, as our work aims at solutions of remote diagnostics using signals extracted from facial regions.

Edge and Line Detection

A common approach to facial areas detection is to take advantage of either a skin color, a specific shape of extracted objects or both. Edge detection is often utilized to obtain shape information. Sobel filtering is frequently used for this purpose, as presented by Huang et al. [45], who combined Sobel techniquee with Two-Stage Multithreshold Otsu method for determining body areas or by Singh A. et al. [46] in a study on face and eyes areas extraction.

Since many of medical images are characterized by complex shapes, that can't be easily detected using horizontal or vertical Sobel masks, a common approach is to approximate edges by a combination of shorter horizontal and vertical components [47]. This process leads to accurate results, but also may significantly increase the processing time. Another commonly used edge detection algorithm is Prewitt operator, which has been proved to perform accurate segmentation of facial areas from thermal images [48]. The best facial features can also be determined using a group of edge detection filters, as presented in [49]. The combination of edge detectors applied in the proposed study included Sobel, Prewitt and Roberts filters, what allowed for the accurate background removal. Edge detection techniques can also utilize magnitude and gradient vector calculated using horizontal and vertical derivatives. These partial derivatives are given by the average of values of neighbouring pixels. In the Canny edge detection algorithm, the local maximum of the gradient magnitude in the direction of the gradient defines edge pixels, what can be utilized for facial expression analysis [50] or chin contour detection [51].

Processing of images acquired in visible spectrum of radiation can also be based on analysis of one of their most important characteristic - color information. Previous studies proved that it can be utilized for face detection task. Skin segmentation is often performed on pixel-by-pixel basis, using specific chromaticity space to determine skin color ranges [52]. Yet, this method is sensitive to changes in lighting conditions. Moreover, this approach can't be directly utilized for thermal imaging, as a very important difference of thermal images is that data obtained from electromagnetic radiation with wavelengths longer than those of visible light, are not perceivable as color to the human eye. A different range of electromagnetic spectrum wavelengths is captured to form an image. This range is outside values that produce a chemical response in retina. In order to produce data which will be perceived by humans, captured signals are converted to various values of brightness, e.g. by using look up tables. Yet, this information can't be considered as representation of features itself, but rather as a generated visualization. Additionally, it's worth noting that features in thermal images are more blurred and contrast between adjacent regions much lower than in visible light data characterized by high frequency components. This is caused by heat flow between objects and thereby blending of temperatures of neighbouring objects visualized in thermal imaging, leading to smoother boundaries between image components. This problem is further described in the next chapter (Chapter 4).

Points Detection

Detection of facial landmarks favors many applications, such as face recognition, tracking or modelling due to ability of capturing and describing face characteristic. Thus, detection of facial points is usually a first step in those applications. Yet, an important problem is that such landmarks can drastically differ in terms of values distribution in the input data, thus it is necessary to add a proper context to them, so that their interpretability will become feasible. This problem was solved by Herpers R. et al. [53] with the means of basic filters operations that were: (1) searching for predefined orientation and scale; (2) determining line orientation; (3) tracking the detected line by moving the filter in the given direction. However,the performed experiments showed that if an image is characterized by a low contrast, a person is wearing contact lens or if shadows disturb facial areas, a detection of keypoints fail.

Various studies utilizing face landmarks detection have been conducted in visible light image domain. Lai J. et al. [54] proposed a method for facial points detection using horizontal projection of a binary image, which assumes that number of peaks in the projection corresponds to horizontal location of each facial point. Another common approach is to take advantage of geometrical relations between landmarks to determine their location [55]. For example, the study presented in [55], showed that detection of a mouth area can be simplified by analysing only a face region marked under top, left and right sides of eyebrows.

Similar studies were also performed for thermal imaging. Marzec M. et al. [26] introduced a face areas detection algorithm based on a head size, anthropometric measurements of facial areas and relations between them (e.g. detection of a face symmetry axis, characteristic facial points and regions). In further studies by Marzec M. et al. [27] the proposed face pattern was expanded to support various possible positions of eye regions, improving robustness of the algorithm to head rotations. Another group of solutions utilizing facial landmark detection are based on Interest Point Detectors (IPDs). Dey T. and Deb T. [56] analysed the Features from Accelerated Segment Test (FAST) Corner Detector on two face databases (IRIS [57] and UGC-DDMC [56]). Yet, although the IRIS dataset contains thermal images as well, experiments were performed only on visible light data.

IPDs could be also applied to thermal images. In our study focused on facial feature tracking from thermal images of a size 320x240 pixels [58], we compared performance of 3 IPDs: SIFT[59], SURF[60] and ORB[61]. The selection of these algorithms for our research was motivated by the differences in characteristics of visible light and thermal data. Since utilized thermal images had rather low resolution, facial features were represented by close contrast values. In addition, due to heat flow in objects, regions present in thermal imagery are usually blurred with no clear borders among different areas. Thus, utilization of brightness distribution or face geometry (as in many studies on visible images) may not work. On the other hand, SIFT, SURF and ORB Detectors have already been proved to provide excellent performance for selection of scale, geometric shape and rotation invariant features [62], what turns out to be beneficial for dealing with thermal data challenges as well. Conducted experiments showed that the displacement of the detected nose area in comparison to the manually marked region was the smallest for SIFT (7.0±1.9%) and the biggest for ORB (9.9±2.2%). For SURF it was equal 8.9±2.7%. As expected, for more intensive motion, the accuracy was worse (biggest displacement) - turning head left (ORB 21.9±6.0%, SIFT 15.6±5.6%, SURF 18.0±7.7%); turning head right (ORB 16.6±5.2%, SIFT 11.2±4.8%, SURF 14.6±6.7%). All of the IPDs achieved real-time performance 23.9ms (ORB), 19.7ms (SIFT), 27.6ms (SURF).

Thresholding

Thresholding allows for categorizing image components into groups using information obtained from pixel intensities, e.g. values above (or below) a given threshold are preserved (or set as a new value, e.g. white), while all others are assigned a different value, e.g. black. In this way, interesting image components (foreground) may be segmented from all other objects (background). According to the classification proposed by Sezgin et al. [63], thresholding can be divided into different groups based on information they are exploiting, e.g. histogram characteristic (e.g. peaks), entropy, object attributes (e.g. shape similarity), spatial correlation using e.g. probability distributions (Otsu thresholding [64]), local pixel values or specific components of the image [65].

The more advanced versions of thresholding are often used to accommodate various illumination conditions, i.e. adaptive thresholding. Whereas the basic thresholding utilizes the fixed value of the threshold for all pixels, the adaptive thresholding changes it depending on the neighbouring pixel intensities. As a result, it's less prone to strong illuminations gradients or other lighting artifacts. The robustness of adaptive thresholds was confirmed in study on face detection from complex images [66] and facial features extraction in indoor and outdoor environments [54]. An iterative thresholding has been also adapted to extract skin segments from skin probability maps [67].

The thresholding operation in thermal imaging can be based on utilization of 1) differences in thermal emittance values between body and other objects (e.g. thermal emittance for skin equals 0.98, for polished iron 0.1) and/or 2) differences in objects temperatures. The latter is associated with the fact that human body has usually a higher temperature value (assuming ambient temperature within the range of a room temperature) than other objects which do not emit heat. Taking it into account, the face region can be easily detected by selecting intensities above some specific pre-defined threshold value. Fig. 2.1 shows a thermal image of a face, which contains much brighter components than the background. This difference has been also visualized on the histogram in Fig. 2.2, that clearly presents two separate peaks, the one with higher values represents a brighter region of a face, while the left hand side peak corresponds to background values. By using specific threshold values, two areas can be easily separated. Due to ease of separating important components from noise with thresholding, this operation is usually applied to thermal data in a pre-processing step of other object detection algorithms. In our previous work, we demonstrated that pattern matching and active contour face detection techniques can be simplified by extracting a facial area from background data with the selection of the peak at the right-hand side on the thermal image histogram (higher intensities values) [68] (see Fig. 2.2).

Figure 2.1. Visualization of differences between color intensities in a facial area and a background

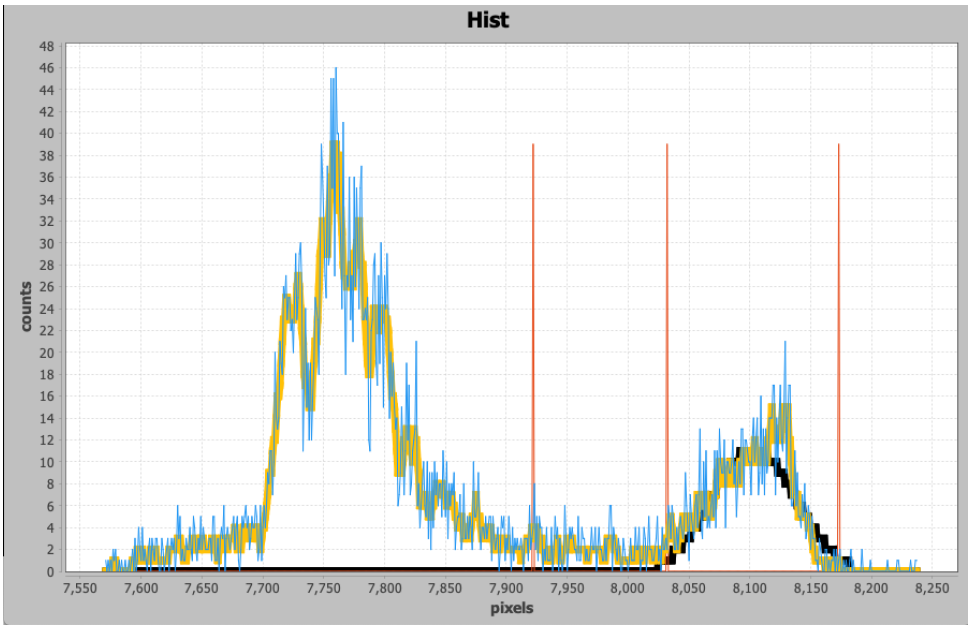


Figure 2.2. Histogram of a face image with two clearly distinguishable peaks, where the right hand side peak represents the facial region

2.2.2 Similarities of Image Regions

Methods described in the previous subsection are used for region segmentation based on discontinuities among pixel intensities within those regions. Here, we will present algorithms that utilize similarities of neighbouring areas/pixels within an image or similarities between pre-defined patterns/sub-images and specific image areas.

Region Growing

The Seeded Region Growing (SRG) method utilizes similarity criterion which specifies whether neighbouring pixels are similar to seed pixels pre-defined for each image region, e.g. nostril area. If the neighbouring pixel is similar to the seed pixel, it is assigned to a corresponding image region.

Fan J. et al. [69] proposed to apply SRG to an automatic face detection task using centroids of image regions obtained from color information as initial seeds. In addition, SRG was also utilized for human segmentation based on seeds defined as detected facial regions. Presented method aimed at performing semantic segmentation, meaning all segmented bodies/body parts were treated as one entity, contrary to instance segmentation, which allows for distinguishing various instances of each class. Our studies, though, assume the presence of a single person in a camera frame and therefore we focus mainly on semantic segmentation. Multi-person environments will be considered by us in a future work.

Region Growing (RG) approach has been also studied for face detection and recognition in a thermal domain. Zheng Y. [70] presented a face recognition solution using region growing and morphology operators to segment human bodies. Facial areas were detected using derivatives of horizontal and vertical projections. Additionally, if a volunteer was wearing eyeglasses, they were removed with RG before matching faces. In the work proposed by Cheong Y. et al., the Zhong's method was improved by using a hybrid approach with Otsu segmentation. As a result, the main disadvantage of RG - the long processing time - was overcome [71].

Quad-Trees

The results of Region Growing heavily depend on the selection of initial seeds. Another regions similarity-based approach, known as Quad Trees, deals with this disadvantage by merging or splitting disjointed regions that the image is divided into at the beginning. The Quad Trees partitioning was used for fractal image coding in the face recognition study presented in [72]. However, experiments performed on 180 grayscale images from 90 subjects showed that the accuracy of face recognition is lower than for other iterative reconstruction methods and that the method should be further developed.

Watershed

The intuition behind the Watershed algorithm is analogous to both RG and QT methods. The advantage of the Watershed algorithms is that it allows for automatic selection of markers corresponding to objects that we want to separate. Usually minima of images (called basins) are selected as such markers. The Watershed method exploits topological characteristic of an image to perform segmentation. It is done by flooding basins until all pixels (from foreground objects to be separated) are flooded and created basins meet at watershed ridge lines. The Watershed algorithm has been successfully used for human/face detection both in visible (for initial separation of image regions later classified with neural network [73]) and thermal image (for objects contour completion [74]) domains.

Patterns

Pattern matching approach, especially image correlation technique, is based on pre-defined templates that contain the same information as the target object to be detected. Many of the pattern matching solutions utilize facial features previously extracted with, e.g. edge detection algorithms [75, 76], color [77], or shape information [78] to build facial patterns. The study conducted by Gao Y. and Leung M. used Line Edge Maps (LEM) constructed using Sobel detector to perform line-based face coding [75]. Suzuki Y. and Shibata T. [76] developed Projected Principal-Edge Distribution (PPED) used for building a face representation. The pattern matching algorithm presented in [77] takes advantage of pre-built head models using color analysis and fuzzy theory.

Most of the studies on pattern matching techniques focus on visible light images and only a limited number of methods have been proposed for thermal imagery. Seal A. et al. took advantage of the Local Binary Pattern (LBP) to perform texture matching of thermal face images [79]. In our previous work [68], we also evaluated accuracy of a pattern matching algorithm aimed at detecting faces from thermal images and compared it with the Active Contour (AC) technique. Active contour models, also known as snakes, are a subsection of pattern matching methods that have been successfully applied to various computer vision tasks, including medical applications [80]. The main advantage of active contours is their ability to dynamically adjust to an object shape. Thus, they are often called deformable models.

Yet, the performed analysis showed that the AC algorithm is both slower and less accurate than the pattern-matching technique on the collected thermal dataset. Furthermore, since face areas are built from complex shapes, Snake model is frequently difficult to apply to the face detection task, as it can only detect a single close contour. Some solutions to mitigate this limitation have been proposed Wu h. et al. [81], but their accuracy was dependant on lighting conditions and body poses.

2.3 Statistical and Learning Methods

In this section we present another group of methods, which utilize correlations between different inputs or probabilistic description of a sample (e.g. frequencies of feature occurrence). Such techniques can learn mappings between inputs and expected outcomes (classification) or exploit representations of samples to divide them into different categories (clustering). Since decision rules used for making predictions can be adjusted based on cost of these decisions in order to find the most optimal model, statistical and learning methods tend to lead to better accuracy and easier adaptability than conventional image processing techniques.

2.3.1 Hand-crafted Features

Clustering

A separate group of image processing learning techniques are based on clustering algorithms, which allow for separating images into groups and in this way detecting specific regions, e.g. facial areas. K-means is one of the most popular clustering methods utilized for different image processing tasks. The intuition behind K-means lies in an iterative selection of centroids for k output categories. At first, centroids are chosen randomly and all samples are assigned to categories based on a distance to a closest centroid. Then, centroids are updated by selecting a sample which is the center of each created group, e.g. calculated as an average value of all group members. After that, samples are reassigned to new groups based on distances from newly generated group centers. The process is repetitively applied until none of examples changes its group.

Interesting clustering-based approach to face detection task has been proposed by Segundo P. et al. [82]. In the presented study, authors divided image regions into two clusters: the facial area (first) and the rest of data (second), i.e. background and other objects that were not considered as Region of Interest (RoI). Although this approach turned out to efficiently extract faces from background components, it required additional filtering operations for separation of the rest of a body, e.g. neck or hair regions.

As shown in [24], K-means is also useful for processing of thermal images of a face. In case of thermal data, facial areas detection can be done by utilization of information about image pixel intensities. According to results presented by Trujillo L. et al. [24], the most significant areas are those where we can observe strongest brightness changes. After extraction of those regions, K-means clustering can be applied for dividing them into individual facial parts, e.g. eyes or nose regions. Another application of k-means in thermal imaging focuses on face verification. Crespo D. et al. [83] proposed to make use of Scale-Invariant Feature Transform (SIFT) descriptors, which are invariant to image scale and rotation. After feature extraction, they were clustered using K-means to build a representation of each person. The limitation of the presented study is sensitiveness to facial poses, as facial temperature distributions is used for making predictions.

Classification

Viola Jones

Haar cascade-based Viola-Jones method [32] became very popular a few decades back and turned out to be a huge breakthrough in the face detection and recognition field. P. Viola and M. Jones managed to outperform previous solutions by introducing AdaBoost - a learning algorithm for selecting best classifiers based on calculated Haar features extractors. Selected classifiers

were consolidated into a cascade to form a final face detector. This approach significantly reduces required computation and thus also processing time, as the final decision is positive only if decisions of all previous classifiers are positive. Since invention of the Viola-Jones algorithm it has been widely applied in various face detection studies, e.g. as a hybrid approach with the Canny Edge detector [84]. Further improvements of the Viola-Jones method included for example the replacement of classical Haar features with polygonal features [85]. The problem of annotating objects with polygonal regions was solved by decomposing a rectilinear polygonal integral into a finite sum of rectangular integrals. It has also been proved that replacement of a rectangular region with a polygonal shape led to better accuracy of objects detection by eliminating the presence of unwanted regions, e.g. background elements.

Viola-Jones algorithm has been verified in thermal image domain as well. Basbrain A. et al. [86] experimented with different features (Haar, Histogram of Oriented Gradients and Local Binary Patterns) in order to determine which ones are best for detecting faces from thermal images with Viola-Jones algorithm. In addition, authors showed that proper image pre-processing is crucial for accurate predictions. In our study focused on thermal face detection from sequences recorded with a portable thermal camera [68] we also compared Viola-Jones approach with methods based on a face geometric, i.e. pattern matching and active contour algorithms. According to achieved results Viola-Jones algorithm produced detection accuracy of 90.5±4.34% at a shortest processing time of 23ms. This outcome indicates the suitability of Viola-Jones method for real-time face detection applications. However, it has been also shown that the presented approach is sensitive to movements performed by volunteers.

Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning binary classification method. During the learning procedure, training samples are represented as points in space, separated by a hyperplane based on categories that they belong to. The goal of model optimization is to make this gap as wide as possible. Classification of new samples is done by mapping them into the same space and assigning to a category corresponding to the side of the gap that they occupy.

SVM has already been applied for facial areas detection both in visible light and thermal spectrum. A very common approach is to make use of Principle Component Analysis as a feature extractor and then divide produced representations into classes corresponding to different facial regions using SVM [87]. Other studies proposed to utilize SVM with features defined using, e.g. Gabor key points or Histogram of Oriented Gradients (HOG), showing improvement of 8.75% in comparison with Principal Component Analysis (PCA) [88]. Gumus E. et al. [89] proved robustness of Wavelet based SVM algorithm over other methods, i.e. distance classifier in a face recognition task. In study conducted by Jee H. et al. [90], eye regions candidates were defined using color, edge, and binary information and then classified using SVM. Comparison of SVM with neural network has been also performed by showing a significant improvement of face classification accuracy for SVM with Radial Basis Function (RBF) kernel, which allows for making predictions from data that is not linearly separable [87]. RBF-based SVM has also turned out to be useful for discriminating facial poses , what may be useful for face alignment in contactless vital estimation studies, since in such cases a frontal pose is usually desired.

Thermal image detection task also benefited from application of SVM. Evaluation of various face recognition algorithms in thermal imaging has been discussed by Floody D. et al. [91] resulting in higher performance of SVM comparing to Local Binary Patterns (LBP) and Trade-off correlation filters (TOF). SVM method with Haar-like features has been used in a research conducted by

Wang S. et al. [92] and demonstrated high accuracy in an eye detection task. Similarly, Martinez B. et al. [93] proposed to localize eyes and nose regions from thermal images using their representations defined with Haar features. To classify pixel clusters, SVM was utilized on image sub-parts extracted using the sliding window approach. Analysis of possibility to apply SVM to infrared images has been also analysed by Wang X. et al. [94] in a research focused on auto-localization of a face, resulting in false negative error of 3.73% and processing time of 200-500ms. Contrary to results achieved in visible light data [87], study on identification of thermal faces showed robustness of SVM with linear kernel over radial kernels [95]. Yet, according to achieved results, high accuracy is achieved only for fusion of visible and thermal features. The proportion of features proposed by authors is 70% visual and 30% infrared, what indicates that higher resolution and presence of high frequency data is essential for person recognition.

Although SVM leads to satisfactory results for many face detection applications, it may not perform well for overlapping classes. Also, choosing a kernel function appropriate for a given task is not easy. Very often linear kernels perform well, but then problems that are linearly inseparable can't be solved.

K-Nearest Neighbours

Another popular learning algorithm often used to solve classification and regression problems is based on relations between training examples in the feature space. In a classification setting, an output category is chosen as the most frequent class of k neighbouring samples. In case of a regression task, the result is calculated as the average of those nearest neighbours. Due to utilization of information from adjacent examples, this approach is known as k-nearest neighbours (kNN).

The effectiveness of kNN over SVM was proved in the study conducted by Parveen P. and Thuraisingham B. [96] aimed at real-time face recognition from visible light data. Other face classification, detection and recognition applications based on kNN have also been proposed. Some examples used in visible light image processing include kNN classifiers based on eigen vectors extracted from color components using PCA [97]. Zheng Y. et al. proved that combination of visible and thermal modalities lead to reduction of false acceptance rates in a face recognition system based on the kNN method [98].

A different study based on thermal image processing and kNN focused on recognition of facial expressions using information about temperature differences [99]. However, the achieved recognition rate of 61.62% isn't satisfactory for medical applications especially requiring high reliability. Other approaches utilized in thermal imaging made use of facial region histograms [100] or orthogonal moments [101] utilized as features for kNN classifier.

Although kNN has been successfully applied for thermal face detection, its main drawback is the selection of the number of nearest neighbours that should be taken into account for classification decision. With the increasing k results are more stable but computational overhead increases. This limitation is also valid for K-means algorithm (in case of K-means k meaning number of groups). Thus, we should know beforehand how many regions are present in our data. This prerequisite knowledge, e.g. of number of faces visible in a frame is very hard to define and that's why k-means-based solutions may not generalize well to new samples. As suggested by Park C. and Kim H. it is beneficial to post-process outputs using e.g. edge detectors or color mapping in other to reduce number of mis-classified pixels [102]. Yet, this can influence the processing time, what in case of generating responses about health status may be crucial. There are also some empirical approaches for determining the optimum number of categories/clusters, such as model fitting (e.g. with mixture model) or visual techniques (e.g. elbow rule).

Bayesian model

Bayesian model is a statistical technique which has been widely applied in literature for different image perception tasks and became a base for various medical detection and segmentation algorithms. Key concepts of Bayesian algorithm is a knowledge about a priori probability specifying a probability of samples belonging to each class and a conditional probability used to define a probability that a value of a sample belonging to a given class will be in a given range. Bayesian-based solutions have became popular for analysis of facial areas, e.g. in skin color models [103]. In order to determine skin/non-skin conditional probability density functions the authors of [103] proposed to use normalized histograms of skin and non-skin pixels. The performed analysis of the skin Bayesian model proved that its reliability is insensitive to different color spaces, e.g. RGB, YCbCr, or HSV, as long as all channels are used. Nguyen D. et al. introduced face detection and lip extraction technique with Bayesian theory [104]. In addition, authors proposed to use edge representations for objects classification what lead to reduction of the model size 2417 times comparing to previous Bayesian-based modeling approaches.

In thermal image domain, it's also possible to make use of Bayesian classification method by using information about temperature distribution. As specified in [105], the main advantage that this solution possess over facial geometric-based approaches is higher permanency of used features and proneness to changing illumination factors. In the presented approach face recognition task was solved by estimating pose of a person and then matching extracted thermal points with corresponding samples stored as templates in a database. In the study conducted by Gordon C. et al. [106], Bayesian method was proved to outperform Viola-Jones algorithm on a thermal datasets. Yet, the required step for achieving such results was to retrain a model on thermal data, as visible light features could not have been transferred to thermal imaging. This indicates that there is a huge demand for providing datasets of thermal images, what we address in Chapter 3. It has also been shown [106] that the accuracy of face recognition using Bayesian classification applied to visible light data and thermal images drops significantly if PCA is used. For more advanced feature extraction methods, no such drop is observed. Taking it into account, we believe that proper representation of data is crucial for final performance and there is a need to research solutions that allow for automatic extraction of important facial features. Such techniques will be discussed by us in a following section (Section 2.3.2). Moreover, Bayesian theory assumes that the probability of distribution of all classes is known in advance. Although such probabilities can be estimated, in many cases they are not known. As a result, it may not be suitable for real life problems.

Shallow Neural Networks

Although statistical methods, such as Bayesian model, are effective for different image processing problems, they are not capable of providing optimal solutions for more complex tasks. One of the reasons for this is the need to make assumption about a type of data distribution, as such techniques make predictions by drawing samples from provided inputs. Neural Networks (NNs) allow for finding generalizable data patterns directly from examples, so no assumptions have to be made As a result NNs are more resistant to outliers, especially in case of unknown data nature or missing values. This is a huge advantage, as extraction of data patterns allows for solving more advanced image processing tasks.

Neural networks were inspired by biological neural system structure, thus they are composed of connected group of neurons. Those connections are represented by weights, which are adjusted to model the problem represented by provided data. While working with neural network, we can distinguish two tasks. First one, training, aims at adjusting model weights, so that produced pre-

dictions are as accurate as possible. In a supervised setting analysed by us, this is achieved by feeding the network with training samples, producing outputs and comparing them with provided labels. This procedure is repeated until a difference (e.g. defined by log loss function) between outcomes and their corresponding expected values is minimal. After each feed forward pass, the backpropagation algorithm is utilized to update the weights with respect to the calculated prediction error. Once optimal weights are learnt, the network can be used for a second task, known as inference. At this step, no more updates to the model are done and previously learnt network parameters are used for calculating outcomes of new samples.

In this subsection, we focus on shallow neural networks, the ones that contain only a single layer. The simplest form of neural network is known as a perceptron. As shown in Fig. 2.3, the perceptron consists of multiple inputs connected to one output. The strength of each connection is represented by weights, which are tuned during the training procedure. The final decision is made using thresholding operation applied to the output calculated as a sum of all weighted inputs. Logistic regression can also be modeled as another simple type of NN. In this case output of the network is followed by logistic function which allows for converting produced scores into probabilities of classes, as presented in Fig. 2.4.

Some attempts for utilization of linear regression for face recognition tasks have already been made. Naseem I. et al. [107] proposed a modular linear regression classification algorithm which mitigates the problem of facial features occlusion using novel Distance-based Evidence Fusion. In a study on thermal face detector [108], different facial representations were constructed and compared in order to define the best set of features for image analysis, i.e. Haar wavelet coefficients and Local Binary Patterns. Classification was performed with a shallow NN, resulting in higher detection accuracy for Haar wavelet feature extractor. As specified by authors, due to the use of NN no prior knowledge of face geometry is required, but because of using hand-crafted features it is applicable only to frontal views and constant backgrounds.

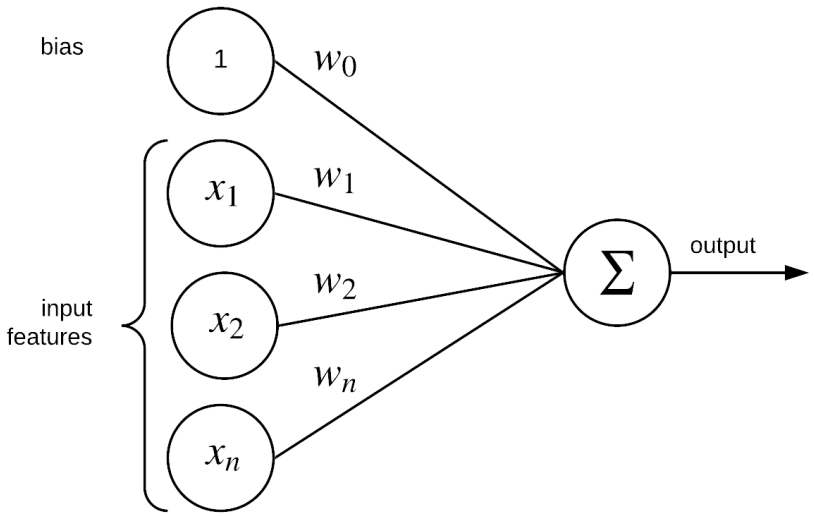


Figure 2.3. Simplest neural network architecture: perceptron

Main drawbacks of shallow networks are twofold. First, they can only be used for solving problems with linear decision boundaries. Secondly, as shown by exemplary NN-based solutions,

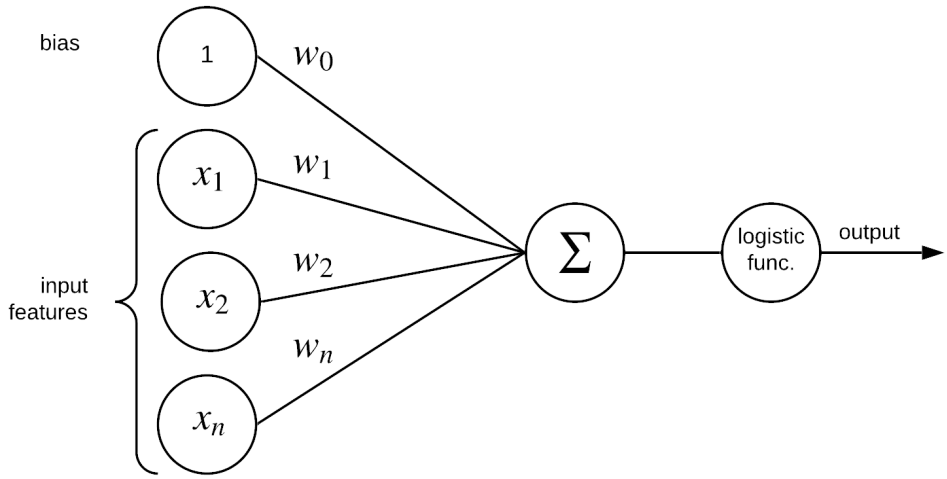


Figure 2.4. Logistic regression classifier

shallow models are not able to examine data directly. There is a need for defining sets of features, e.g. edges extracted with Canny filter or Local Binary patterns that could be used for making predictions by correlating them with expected outcomes. Both problems can be addressed by techniques introduced with DL, e.g. representation learning, weight sharing and solutions for modelling of non-linearities, explained in the following section.

2.3.2 Representation Learning and Modelling of Non-linearities

Motivation

Although shallow NN, e.g. perceptron or logistic regression can learn data patterns directly from examples, usually the complexity of the network is not sufficient to model data representations. Instead, features are manually designed, what requires a lot of effort and time. We may say that those simple machine learning techniques are representation-dependant, as we have to define parameters important for a given task. For example, in logistic regression-based disease classification, specialists have to define features specific for a given condition, as algorithm is not able to examine provided images directly. Machine learning system can only learn how to make proper predictions using correlations between features determined by a human and various observed outcomes. As one may note, this is a huge disadvantage since many problems are frequently hard to described with enough complexity using strictly defined knowledge. Moreover, choice of handcrafted features meaningful for specific image processing task is not straightforward, as it requires correlation of each feature with an outcome and very often results may be influenced by varying lighting conditions or different personal factors, e.g. age or gender [41]. Although in many cases (especially for strictly defined measurement conditions) handcrafted features are effective, the task of making proper predictions becomes more difficult for complex inputs acquired in dynamic environments.

Selected features may also contain a lot of meaningless inputs and, as a result, solutions are not optimal. There are different techniques for reduction of data dimensionality to preserve only most essential features for a given prediction, e.g. PCA applied to find most optimal faces vectors [49]. Yet, finding the right set of features is still very challenging. One idea for mitigating this problem

is to learn not only mappings between features and categories, but also data representation itself. This approach of automatic extraction of features without human supervision or action and then analysis of how they map to specific outputs is known as representation learning. One of the main most developed in recent years techniques utilizing this approach is DL (DL). By learning representations, Deep Neural Network (DNN) is able to acquire patterns from raw data which are used for making predictions that may seem subjective.

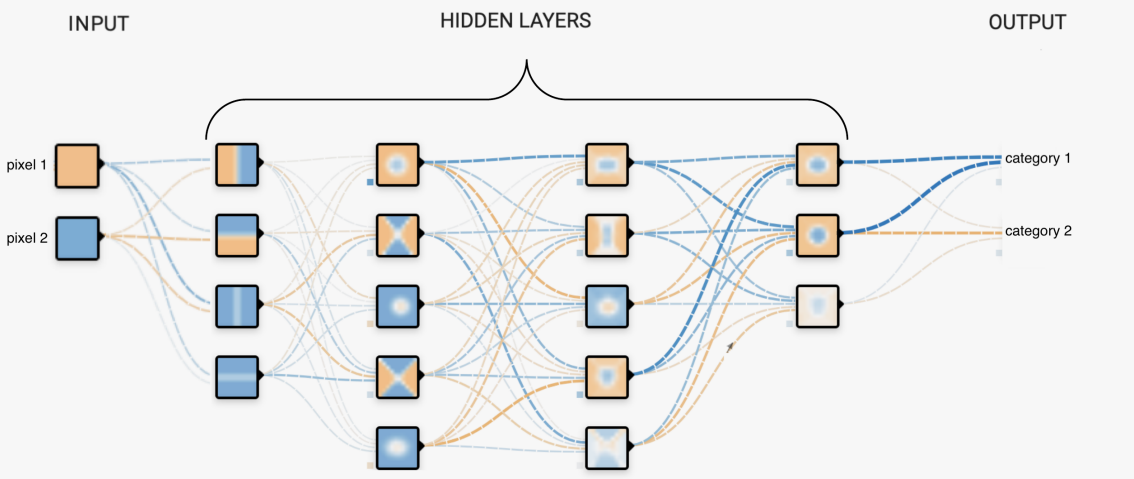


Figure 2.5. Simplified visualization: sequence of simple nested mappings constructing a Deep Neural Network, at each level more complex features are extracted

A Deep Neural Network (DNN) is a specific type of NN, which uses more layers, denoted as *hidden* due to their placement between input and output layers. DL allows for reduction of data variation and automatic generation of data representations by dividing the complex direct input-output mappings into a sequence of nested mappings between respective representations extracted from subsequent layers of the model (Fig. 2.5). As can be seen, a more sophisticated object can be presented by combining features of complexity increasing at each layer, starting from e.g. points and lines that are then formed into contours, object parts and finally a full image. The classic example of DNN is a multilayer perceptron [39], which represent mapping of input to output values using complex mathematical functions defined with a sequence of simpler ones. Since more complex problems require execution of more instructions, it can be easily deduced that deeper networks allow for achieving higher accuracy in various computer vision problems. Moreover, contrary to shallow NN, deep models can solve problems with non-linear decision boundaries due to the use of multiple layers and non linear activation functions. This feature makes them capable of modelling complex real-world problems and achieving human-like performance.

Other key concepts used in DL are distributed learning and weight sharing. They assume that features are shared across multiple possible samples (distributed learning) and across different objects present within a single input (weight sharing). Thus, kernels with the same weights can be applied to different locations of the input to produce accurate predictions (as specific objects are built from similar lower-level representations, e.g. lines, edges, etc). As a result, very complex problems can be realized by using reduced number of neurons. For example, suppose we would like to perform emotion and gender (male vs. female) recognition. One way of achieving this would be to use 2N neurons, each neuron per one possible combination (male-sad, female-happy, etc.), where N represents number of emotions to be classified. Using the concept of distributed representation

we will need 2+N neurons only, as neurons responsible for generating emotion representations are able to learn from all inputs, not only from samples belonging to one gender.

By using representation learning and weight sharing approaches, DL eliminates a need for defining any prerequisite knowledge, contrary to previously described learning methods. For example, k-means algorithm applied to face detection task [82] require definition of a number of faces. Frequently, such details are impossible to determine what results in system being not able to generalize to new data. This problem is not valid if Deep Neural Networks are used, as they can automatically adapt to data representation and extract patterns irrespective of their location or number of instances.

Although DL has already been known for many years, only recently gained much more attraction. There are different factors that led to this progress. One of them is the big data trend which allows us to provide models with enough data to learn proper correlations between inputs and outputs. Another reason for DL being successful nowadays are much better hardware capabilities with bigger storage and faster computation units, which make it possible to train deeper networks that better represent processes of human brain neurons. Last but not least, recent advances in research on neural networks has led to the development of much better regularization techniques, so training of models became more feasible. All those aspects resulted in the huge progress in neural network design, enabling various applications to achieve human-like perception performance.

Classification and Recognition

Since the reinvention of DL, the image classification research has significantly expanded, doubling the size of models roughly every 30 months [39]. The architecture of neural network commonly used for image recognition task is known as Convolutional Neural Network (CNN), as the basic building block used in this structure is a convolution operation. Similarly to introduced earlier edge detection filters, the convolution operation produces a single output for an input pixel by calculating a weighted average of neighbouring values surrounding it (within a given window size). The weights are automatically adjusted during the learning procedure, contrary to edge detection masks which use pre-defined unalterable values for their kernels. Apart from convolution window size, there are two main parameters associated with this operation. First of them is stride, which specifies the number of pixels by which the kernel is moved across the input (see Fig. 2.6). The second one, called padding, is used to define whether border pixels should be taken into account (*same* padding) or should be skipped (*valid* padding). In case of padding set to *same*, the input has to be surrounded with zeros in order to preserve the same output size as the input (number of additional rows and columns up to the half of the convolution window size rounded down, so that the border pixels are centered at the kernel, see Fig. 2.7).

The pioneer CNN network (named AlexNet after its inventor Alex Krizhevsky), which started a huge breakthrough in image recognition, was built from only 5 convolutional layers [109] and still showed that it's capable of achieving record breaking results even for very complex data. Since then, we have observed a revolution of network depth producing better accuracy with every additional layer, e.g. VGG with 19 layers [110] reduced top-5 error by 9% or GoogleNet with 22 layers by 9.5% comparing to AlexNet [111]. Later, it turned out that simple stacking of more layers is not enough, as training become more complex because of the difficulties with propagating gradients through the increased depth (vanishing gradient problem [112]). The solution to this problem came with the invention of residual blocks. Authors of Residual Network (ResNet) architecture [113] proved that a feedforward network with skip connections is easier to train, achieving state of the art

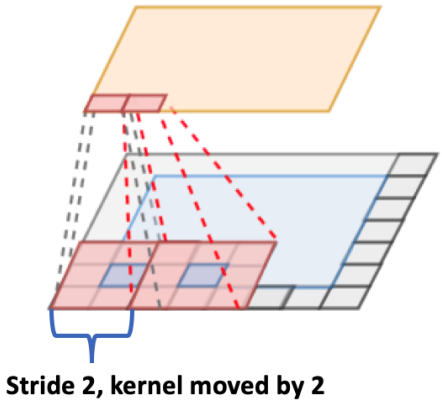
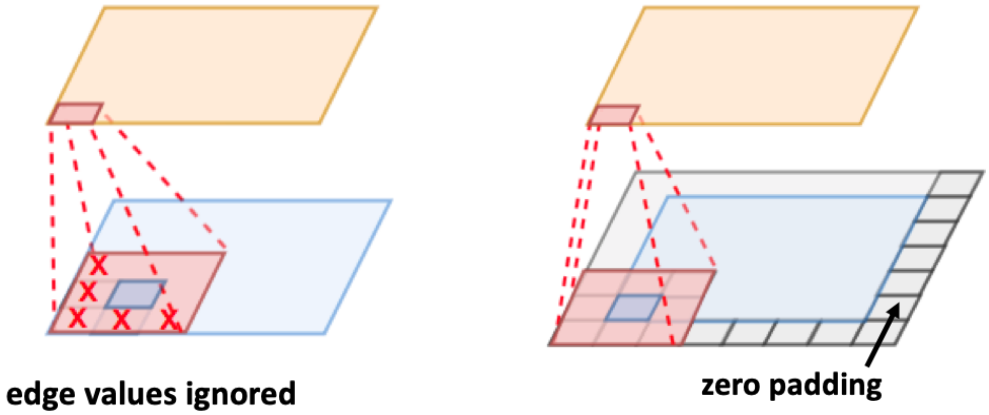


Figure 2.6. Visualization of a stride parameter indicating a number of pixels by which convolutional filter is moved across an input



(a) Padding valid

(b) Padding same

Figure 2.7. Explanation of the padding parameter indicating a way how edge values are handled during applying convolutional operations

performance in object classification task. Skip connections are those that skip one or more stacked layers, implementing residual mapping. Moreover, they do not introduce additional parameters and computational complexity remains constant. Since the introduction of residual network, this idea has been widely applied for other computer vision tasks, including image enhancement [114] and denoising [115].

The huge progress in Deep Neural Networks advances has led to significant improvement of image recognition accuracy. Thus, DL (DL) started to be applied to various image perception tasks, including face classification. CNNs have already been successfully used for distinguishing facial from non facial images and pose estimation [116], facial points (nose tip and mouth corners) classification [117], gender classification [118], or face verification and recognition [119]. Cheung B. deterred automated systems by using CNN to recognize real facial images from generated avatars [120]. The introduced convolutional network consisted of only 6 layers, producing accuracy of 99% in ICMLA 2012 Face Recognition Challenge. In the study conducted by Anderson R. et al. [121] the performance of state-of-the-art convolutional networks was compared in facial attractiveness

classification task. As previously confirmed, the network with residual blocks outperformed other solutions that were using just a stack of convolutional networks, i.e. VGG and Inception. In addition, authors also analysed the influence of applying Haar Cascades as the preprocessing method, improving accuracy of both Inception v3 and ResNet50 models.

Although most of studies on CNN-based face classification focus on visible light data (as presented so far), some attempts to utilize CNNs for thermal face classification have also been made. The superiority of DL-based techniques over methods based on manually selected features was proved in the study conducted by Simon M. O. et al. [122], where CNN was used on the RGB-D-T dataset. The optimized version of the GoogleNet network has been also used on thermal inputs for person identification task [123]. Seo J. and Chung I. [124] made use of thermal data in order to determine face liveness with a relatively simple network built from 4 convolutional layers. CNN-based model has been also applied to the task of matching thermal images against corresponding visible light data [125]. Another cross-domain face classification task was implemented with the means of deep convolutional auto-encoders, showing the accuracy increase of 7% [126]. Non-linear mapping between visible and thermal image modalities can be also realized with the means of deep feed forward CNN, named Deep Perceptual Mapping (DPM) [127]. Biometric authentication also benefited from the use of DL on thermal data. Grudzien A. et al. [128] explored the robustness of siamese convolutional model built on top of the VGG network in the task of comparing two samples against each other for verification purposes. Yet, the comparison to other, more recent, CNN architectures, e.g. residual or inception blocks based, hasn't been presented. Sayed M. and Baker F. [129] proposed to determine person's identity from thermal input using CNN. Due to the use of thermal imaging, the proposed solution could be successfully applied to person recognition at night in video surveillance systems. The next application of processing thermal facial images with CNNs was introduced by Obi-Alago A. et al. in their study on face signature abnormality classification [130]. Yet, the specifics of the network architecture were not given. Similar research of face abnormality detection, specifically sobriety classification, was proposed by Menon S. et al. [131]. The presented study was motivated by the algorithm for low resolution thermal face detection proposed by us and described in details in Chapter 4. Yet, CNN was only used for face classification without detection of its regions.

Detection and Segmentation

A common practice used in object detection is to determine the presence of each category within selected bounding boxes using features extracted with convolutions. Thus, CNN blocks are used as the backbone for feature extraction, on top of which additional operations are applied that allow to determine probability of categories and adjust coordinates of an object location.

Over past few years, different DL-based object detection architectures have been invented, proving their advantage over methods based on manually selected features, e.g. as presented in [132], where authors showed that DNNs are better than Haar cascades for finding face orientation. The pioneer work in CNN-based object detectors was proposed by Girshick R. et al. with the introduction of Regions with CNNs features (RCNN) model [133]. The first consideration of how to approach face detection using CNNs focused on the sliding window method and features extraction at each window location [134]. However, as mentioned by the authors of RCNN, this leads to many difficulties with determining the precise object position, as the receptive field after a few convolutional layers is very large. Thus, the alternative is to use *recognition using regions* approach, which takes advantage of pre-defined regions instead of checking for object presence at each window

location. This approach was used in RCNN which at first generates category-independent proposals of bounding boxes and then applies CNN to extract a fixed-length vector representing features present in each proposed bounding box. Classification is performed with category-specific linear SVMs, what is the main drawback of this architecture due to the decrease in a processing speed. This limitation was mitgated by replacing SVMs with a fully connected layer (*fc*) in NN architecture named Fast R-CNN [135]. In order to be able to make use of the *fc* layer, extracted features had to be cast to fixed-length vectors. This step is achieved in Fast R-CNN with a RoI pooling layer. The *fc* layer is then branched into two outputs: probability for each category produced using softmax function and a second layer that produces four real-valued numbers corresponding to bounding box coordinates.

Although the training and testing speed has been improved in Fast R-CNN comparing to RCNN, both architectures have a significant limitation - a requirement of obtaining pre-defined region proposals using a selective search algorithm. This problem was addressed by Faster R-CNN [136], a model which utilizes a separate neural network aimed at generating region proposals, in this way eliminating the time consuming selective search process. Further improvements to object detection networks have been achieved by using a single CNN to obtain bounding boxes and perform object classification within a single forward pass. This optimization turned out to be crucial for real-time performance, as presented in some recent models, such as You Only Look Once (YOLO) [137] or Single Shot Detector (SSD) [138]. The main difference of those architectures is that they look at the whole image instead of only specific regions that have the highest probability of containing the object. In YOLO architecture, an image is divided into a grid and bounding boxes are generated for each cell within this grid. Then, the model outputs probabilities and offset values for each bounding box. Finally, bounding boxes with probabilities above a specific threshold are preserved and used to produce model outputs. In SSD network, a similar concept of defining bounding boxes for the generated grid is utilized. The main difference from the YOLO model is that SSD makes use of multi-scale feature maps to generate independent object detectors responsible for localizing different scale objects. In this way, the accuracy of detecting objects of various sizes is improved.

Another approach to object detection problem was proposed by Zhang Z. et al. in their research on face detection and alignment [139]. The introduced framework was built using a unified cascade of CNNs and multi-task learning (MTCNN - Multitask Cascaded CNN). In the first step, the Proposal Network (P-Net) was used to produce bounding box candidates that were then fed into the Refine Network (R-Net) for selection of the positive outputs. The last network in the cascade, the Output Network (O-Net) aimed at producing positions of facial landmarks.

Similarly to the classification task, the majority of DL models have been evaluated on visible light data only and only a few studies were conducted for thermal face detection. For example, the robustness of MTCNN in the thermal domain was compared against two other landmark detection models originally developed for visible light images: the Deep Alignment Network (DAN) and a Multi-class Patch-based fully Convolutional Neural Network (PBC) [140]. The presented results showed that the DAN architecture has the best ability to adapt to the thermal domain. However, it has been also shown that even small errors in the face alignment lead to huge drop of recognition accuracy. Since the number of studies on thermal face detection is limited, we focused on exploring this area of research in our work to verify robustness of DL-based techniques in thermal imaging. Some of our findings have been already published [58, 68, 41, 141]. We present details of those studies and further expand them in the following sections of the dissertation.

2.3.3 Evaluation of DL Models

After model training and adjustment of weights in such a way that produced predictions are as close to provided labels as possible, performance of a network should be evaluated on previously unseen data. In this way, we can estimate if a model was properly optimized and whether we can trust produced predictions. In general, we can say that a goal of evaluation is to estimate the generalization accuracy by testing networks on future (unseen/out-of-sample) data.

In order to provide quantitative analysis of a model performance, various evaluation metrics are deployed. The choice of specific ones depends on a problem that a network is supposed to solve, e.g. classification, regression, detection, etc. There are also some evaluation metrics which could be applied to various tasks, e.g. sensitivity (recall), precision, specificity (true negative rate) etc. In this section we will focus on describing metrics useful for machine learning problem evaluated and addressed by us in this study. Since we are considering remote medical diagnostic applications, some additional evaluation is required to determine if they could be run on target platforms which are usually resource constraint devices. Thus, we also define metrics which specify processing performance.

Classification

One of the most common evaluation metrics for classification problems is accuracy. This metric indicates a ratio of correctly classified samples to an overall number of examples. For binary classification problem it can be defined as:

$$Acc = \frac{TP + TN}{N} \tag{2.1}$$

where TP is a number of True Positive predictions, TN is a number of True Negative Predictions and N is a number of all examples in a test set. Accuracy is a relatively simple metric for determining general model performance, yet it may be not sufficient, especially for problems where datasets are imbalanced, meaning one class is much more numerous than the other. Class imbalance problem is frequently an issue in medical databases often characterized only by a limited number of specific condition representatives. Let's assume we have a dataset with 10 positive cases and 90 negative cases. Our classifier outputs negative decision for all examples, so the accuracy equals $Acc = \frac{90}{90+10} = 90\%$. At first we may think that the achieved performance is very good, but in fact our model is not useful at all. This problem is known as accuracy paradox and is a reason for a need to make use of other metrics as well.

Taking it into account, accuracy should be always accompanied by other metrics that are not sensitive to such problems. Evaluation parameters commonly used for classification tasks are precision and recall. Precision, also known as positive predictive value tells us how many samples from all positive predictions are in fact truly positive, what can be denoted as:

$$Prec = \frac{TP}{TP + FP} \tag{2.2}$$

Recall on the other hand specifies sensitivity of a classifier, i.e. tells us how many samples from all positive examples present in the test set the classifies was able to pick up. Thus, recall is also known as true positive rate and is defined as:

$$Rec = \frac{TP}{TP + FN} \tag{2.3}$$

Detection

Since detection models aim at producing coordinates of objects' positions apart from probability of each region belonging to a specific category, evaluation applied to this task require metrics that take into account a size and a placement of produced detections. Intersection over Union (IoU) allows for making such comparison by analysing overlap and union areas between bounding boxes (BB) detected by a model and their corresponding ground-truth (GT) examples marked manually in test images:

$$IoU = \frac{GT \cap BB}{GT \cup BB} \tag{2.4}$$

Fig. 2.8 illustrates this relation. IoU is a relatively simple, yet effective way for evaluation of object detection models and will be used fr our studies described in the dissertation.

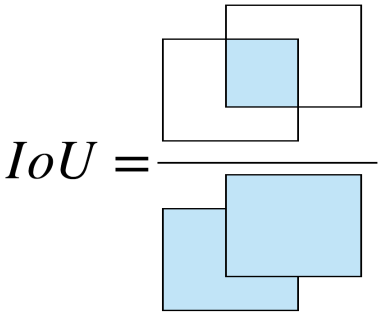


Figure 2.8. Visualization of the IoU metric used for evaluation of Deep Neural Networks

Another metric used for object detection evaluation is Average Precision (AP), which associates model scores with detected bounding boxes and estimates performance at various level of confidence. Before introducing details of AP, it's important to take into account relation between precision/recall and IoU threshold set to determine whether produced output is positive or negative. For example, if our model produces 10 outputs: 4 with IoU values equal 0.5 and 6 above 0.6. Setting threshold to a value > 0.5, would result in 4 boxes being false positives, while setting threshold equal to 0.5 would produce 10 true positive outcomes. Following the Pascal VOC convention [142], the final value of AP is calculated as average value of precision (if multiple values are present, the maximum one is selected) for 11 predefined recall values: $[0 : 1.0 : 0.1]$:

$$AP = \frac{1}{11} \sum_{Recall \in \{0:1.0:0.1\}} Precision(Recall) \tag{2.5}$$

The AP metric may be further extended by taking various IoU thresholds into consideration. Different IoU values are used to determine whether detected object is true or false prediction, e.g. if IoU threshold is defined at 0.5, then all results for which relation between overlap and union areas is lower than that will be treated as false results. Hence, one can observe that the choice of the IoU threshold will have influence on AP metric leading to more false positives or false negatives outputs. The final detection evaluation metric is then formulated as mean value of APs for various IoU thresholds and across all classes and is referred to as mAP - mean Average Precision.

Execution Time

Target platforms considered by us in the presented work include resource-constraint devices which may be embedded in wearable equipment, such as glasses or watches (e.g. to support healthy

lifestyle or perform diagnosis at a distance) or be a part of smart home infrastructure if applications aim at providing home monitoring of human subjects. Taking it into account, in our research we also measure inference and training time of proposed DL models in order to evaluate their applicability to such systems. Training time is considered, because some future studies may include scenarios where model is retrained directly on the edge device, e.g. to add a new person to a household health monitoring system. Inference time is defined as a time of processing a batch of inputs in a forward pass, while a training time is a sum of forward and backward pass.

2.4 Problems

Undoubtedly, the research on thermal face and areas classification and detection has significantly progressed over past few years and results are much more promising due to advances introduced by DL. Yet, there is still a room for improvement in this area, as most of existing solutions are designed specifically for visible light data that have a different representation than thermal images. First of all, visible light sequences are characterized by the presence of high frequency features, such as edges or corners, while thermal data are more blurred and a contrast between adjacent regions is very smooth due to the heat flow in objects. Thus, the use of kernels learnt (on visible light data) to extract high frequency features may not be sufficient for a thermal domain. Secondly, even though availability of cost-efficient higher resolution thermal camera has recently expanded, possibility of acquiring images of quality similar to ones obtained with visible light cameras is still unachievable. Taking it into account, there is a need to improve resolution of acquired thermal sequences in a post-processing phase. Another problem is a lack of publicly available datasets that could be used for model training and very often, even if provided, they do not contain proper annotations, what makes them unusable for supervised training without a time-consuming effort of data preparation.

2.5 Summary

This chapter overviewed types of existing methods used for image classification and object detection, with a special focus put on solutions applied to thermal imaging. The comparison of techniques based on manually selected features and/or learnt mappings between inputs and outputs to various latest DL architectures were also described. Finally, we identified problems related to applicability of described solutions to thermal images. These problems are addressed in the following chapters which cover detailed explanation of proposed novel DL techniques and NN architectures designed for thermal data processing.

Chapter 3

Datasets

3.1 Introduction and Overview

One of the reasons why Deep Learning (DL) became more approachable and easier for engineers without an expert knowledge about artificial neural networks is the increased availability of training sets. Certainly, some knowledge is still required to achieve a descent performance, but the task of generating a successful DL model is easier with access to more data [39, 41]. This trend is clearly visible in the growing number of samples included in publicly available datasets. Since 1990s sizes of available training sets have expanded by few orders of magnitude, from MNIST (handwritten digits) [143] and CIFAR10 (more complex data divided into 10 categories) [144] containing tens of thousands of examples, to ImageNet [145] and ImageNet10k [146] sets which contain ten of millions samples. Unfortunately, because of some limitations of thermal imaging, such as higher cost of available cameras, availability of thermal datasets is much lower, e.g. IRIS Face Dataset with 4k images [57].

Usually, the limited number of available images makes it difficult to train a very Deep Neural Network (DNN) from scratch. Some techniques to address this problem have been already developed. Specifically, a widely used approach to transfer already learnt knowledge from one domain to another is known as transfer learning. The intuition behind it is that some low level features (corners, edges, basic geometric shapes) are common for various tasks regardless of a used domain. Thus, we can restore weights already learnt on huge amount of data from lower level layers and only retrain only a few top layers in order to adjust our model to a novel task. This technique was further explored by us and described in details in Chapter 4.

In this chapter we focus on providing specifications of thermal datasets acquired by us to verify possibility of applying DL algorithms to thermal data processing, specifically resolution enhancement, detection of facial areas usable for non-contact vital signs estimation and evaluation of other possible remote medical diagnostic applications. The motivation for our own databases acquisition is threefold: 1) even with transfer learning some samples are still required to re-purpose a model to a different task; 2) breathing activity must be recorded during obtaining sequences in order to implement remote diagnostics solutions; 3) we haven't found any thermal dataset which contain samples in original raw format. Most of publicly available sets have been already converted to standard image formats (e.g. 8-bit PNG), causing some loss of the precision, as raw thermal data usually have higher bit resolution, e.g. 14 bits.

3.2 Thermal Sequences Collection

Since we examine various remote diagnostics solutions based on processing of thermal imagery in our work, we collected datasets taking into account different potential applications: evaluation of possibility to apply DL algorithms t thermal data, facial features detection from thermal images, analysis of respiratory activity, and emotion recognition from changes in vital signs patterns. Separate data collections scenarios were considered and applied in order to address each of these solutions. This work mainly focuses on evaluation of DL detection and resolution enhancement models and an effect of applying them in contactless vital signs monitoring solutions. In addition, we also evaluate possibility to recognize emotions from extracted vital signs, as this information may have a huge potential in medical applications for evaluating e.g. pain levels [18], facial paralysis [20], or neuropsychiatric disorders [147]. Taking it into account, this section covers details of data acquisition and post-processing procedures performed by us in order to prepare inputs for evaluated applications.

3.2.1 Evaluation of DL Algorithms on Thermal Data

Imaging Hardware

The goal of our study is to evaluate existing DL solutions and propose novel neural network architectures aimed at processing images acquired in thermal image domain, i.e. measurements in a range of 8–12 μm (Long-Wave Infrared). In thermography, intensities of electromagnetic radiation are represented as digital values that are then used to construct a final image by assigning grayscale or indexed color intensities to digital values using e.g. color lookup tables.

The first database collected by us was used as a reference set, used for initial validation of possibility to apply DL algorithms to thermal images. At a first step we focused on facial areas detection from low resolution images without applying additional image enhancement techniques. Since FLIR® Lepton camera was used for data acquisition (see Fig. 3.1), this set is referred thereafter as Lepton-IE (initial evaluation).

The FLIR® Lepton module, was used in our research because of its huge potential in smart home and remote healthcare applications that are the main focus of this work. A relatively low cost (one tenth the cost of standard IR devices) and a small form factor of the device (a size of a circuit board $< 1\text{cm}^2$) allows for embedding it in existing devices, such as smartphones, kitchen appliances or smart home infrastructure, enabling various medical diagnostics solutions, e.g. estimation of respiratory rate using wearable devices [148] or multi modality-based elderly monitoring platform [149]. A thermal sensor incorporated in the Lepton module records data at 9 FPS in 80×60 pixels spatial resolution in a 14-bit dynamic range.

Data Collection Procedure

Over past years DL has enabled various computer vision applications leading to human-like accuracy in image classification, detection and segmentation tasks [39]. Since we are mainly interested in thermal image processing, as it allows for obtaining medical information not detectable from visible light spectrum (e.g. breathing patterns), our first goal was to determine if Deep Neural Networks could achieve high accuracy also for thermal imagery.

Our initial evaluation was related to possibility of detecting facial areas from low resolution sequences. The low resolution is one of main aspects considered by us, because platforms targeted

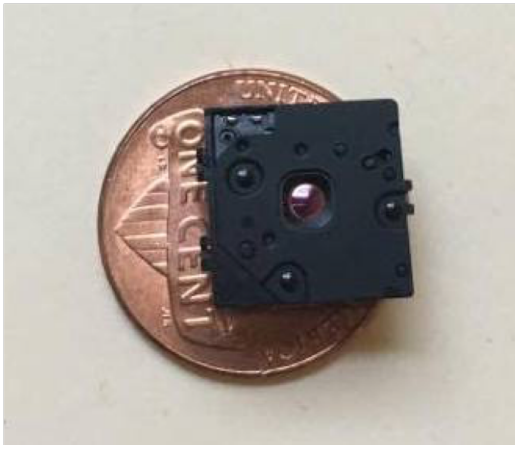


Figure 3.1. FLIR® thermal Lepton camera module of resolution 80x60 used for data collection

in our work (e.g. smart glasses developed for the eGlasses project [150]) have size limitations and thereby allow for embedding only small thermal sensors into them. The examined wearable platform setup is presented in Fig. 3.2.

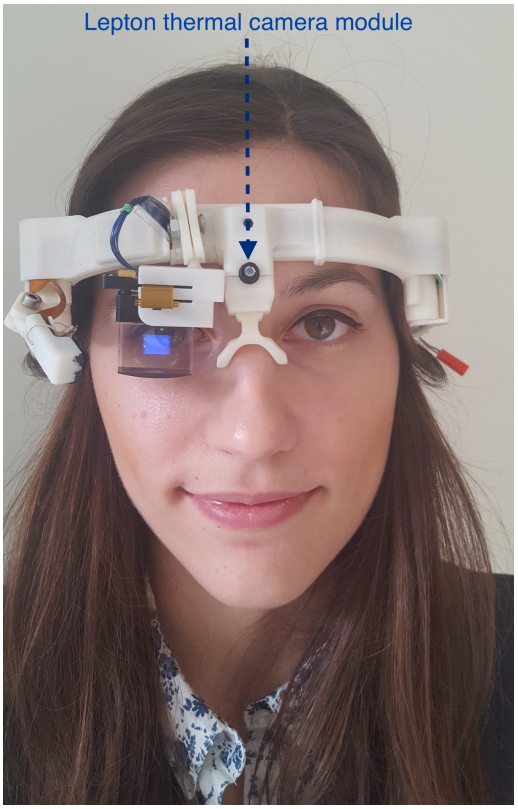


Figure 3.2. Wearable platform developed for eGlasses project; photo of the dissertation author

Thermal images were collected at a distance of 1 meter from volunteers assuming their immobility and frontal view of a face. During data acquisition we wanted to simulate possible scenarios of an elderly person living alone, thus some background objects were present in recorded sequences, as they would also appear in real-life conditions (e.g. displays, furniture, etc). Since indoor appli-

cations are taken into account in the presented work, measurements were taken at an ambient temperature of 23-27°C. The database was constructed by recording 60-sec sequences from 26 healthy volunteers (age: 26.8 ± 8.1), who were notified about a purpose of study and agreed for taking part in it. To ensure data variability, only every ˜30^{th} frame from all acquired sequences was used for extraction of face and eyes areas, resulting in 624 images in each of those categories. For a nostril area we wanted to obtain images with a clearly distinguishable temperature difference in order to make sure that breathing patterns were recorded and reflected in pixel intensities changes. Thus, more dense selection was utilized by preserving samples for every ˜2-sec periods (every ˜16 frame), resulting in 983 images in a nose category. All data was saved in 8-bit PNG format by scaling full input data range to the full output range.

(a) Negative samples (other objects: display, windows, remote controller, hand)

(b) Positive samples (face images)

Figure 3.3. Examples of images acquired with the Lepton camera for our initial studies on evaluating DL applicability to thermal data

In addition, a second dataset was acquired using the same Lepton camera in order to evaluate possible limitations of contactless estimation of vital signs. 11 healthy volunteers of age 31.1±10.6 took part in experiments. In particular, the goal was to determine if movements performed by subjects and/or diagnostician will have influence on accuracy of face detection. This database referred to as Lepton-IE-M, meaning that was mainly used for initial evaluation (IE) of Motion (M) influence on thermal image processing with Deep Neural Networks. Various possible scenarios where motion content could affect quality of collected data were examined. At first, we wanted to limit possible motion of volunteers by asking them to perform a specific task (silent text reading), as it has been previously proved by us [151] that in this way subjects remain more still. Moreover, such tasks are similar to those potentially applied in real-life applications of remote medical diagnostics, where health information is collected in a non-disruptive way during daily activities, e.g. watching television, reading a book, etc. The second use case considered by us aimed at verifying influence of small (almost involuntary) head movements. In this scenario volunteers were not given any additional task, so theoretically motion content should be higher. In both first and second case the camera was placed on a tripod to eliminate potential artifacts introduced by displacement of acquisition device. Finally, we also wanted to examine a problem of movements performed by diagnostician/specialists, e.g. when remote medical diagnostics is performed as a part of routine physical examination or during admitting/monitoring patients in hospitals. In such cases, a camera

could be embedded in a wearable platform, such as smart glasses. To simulate such use case, the Lepton camera was placed on developed by us eGlasses platform - smart electronic glasses designed to enable innovations in human-machine interaction studies by utilization of various sensors (camera, optical sensors, etc.) [152, 153]). During data acquisition, glasses were worn by another person, who was looking at examined subjects. Again, volunteers were not performing any additional task, so potentially both subject's and diagnostician's motion content should be present in collected data. 25091 positive (face) cases and 25002 negative (other class: e.g. objects present in a laboratory room) cases were extracted from all recorded sequences and saved as 8-bit PNG images. Examples of saved frames are shown in Fig. 3.3.

3.2.2 Analysis of Facial Regions Detection and Extraction of Respiratory Activity from Detected Areas

Imaging Hardware

A database collected with the Lepton sensor (Lepton-IE) for initial evaluation of DL techniques in a thermal domain was extended in order to apply it to other studies aimed at thermal image super-resolution and contactless extraction of respiratory rates. Our main focus was to evaluate possibility of detecting facial areas from extremely low resolution thermal images, where features important for predictions may be blurred, making them almost indistinguishable. After region detection, we wanted to verify if marked areas carry information useful for contactless estimation of vital signs by analysing color changes within those regions.

Additionally, to preserve variety of data and ensure that algorithms don't depend on one specific representation, thermal image sequences were also acquired using another thermal sensor: FLIR® SC3000, presented in Fig. 3.4. The FLIR® SC3000 thermal camera records temperature in a range from 20°C to +80°C in a noise reduction mode at 30 Frames Per Second (FPS) in 320 × 240 pixels spatial resolution. A default raw data format of this camera has resolution of 14 bits.

Figure 3.4. FLIR® SC3000 thermal camera used for data acquisition

Data Collection Procedure

Before experiments, participants involved in the data collection process were informed about the aim of the study and details about a data acquisition process. Signed informed consents were obtained from volunteers who took part in experiments. All procedures were performed according to information obtained from the regional, institutional Bioethical Commission. All experiments were carried out in a laboratory room at the ambient temperature of 23 to 27°C.

For facial areas detection and respiratory rate analysis study, we collected data using both the FLIR® Lepton and the FLIR® SC3000 cameras in order to compare whether a size of a focal plane array has influence on accuracy of breathing rate estimation. Datasets collected for analysed scenarios are hereafter referred as SC3000-ADRA and Lepton-ADRA databases (FLIR® SC3000 Areas Detection and Respiration Analysis and FLIR® Lepton Areas Detection and Respiration Analysis, respectively).

The SC3000-ADRA database contains thermal sequences recorded for 40 healthy volunteers of an age 34.11±12. The camera was placed on a tripod at a height of approximately 1.1m and at a distance of 1.2m from participants' heads. 2-min sequences (sampling frequency fs=30Hz) were acquired for each volunteer while looking towards the camera. To obtain reference information about a respiratory rate, participants were asked to point finger up while inhaling and down while exhaling. The movement of a finger was visible in the recorded sequences and used in post processing step to calculate the ground-truth number of breaths per minute for each person.

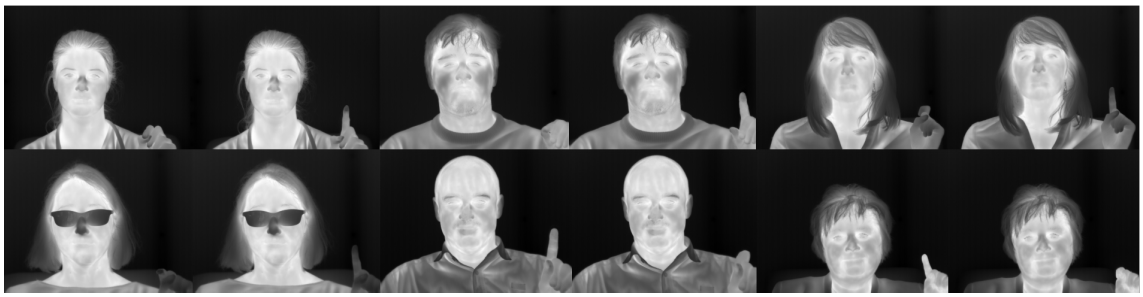
The second database, Lepton-ADRA, is a collection of 1-minute video sequences (sampling frequency fs=9Hz) recorded for 31 healthy volunteers of an age 26±8.1. For data acquisition, the thermal camera was placed on a tripod at a height of approximately 1.1m and at a distance of 0.5m from participants' heads. We decided to collect data at a shorter distance than for the SC3000-ADRA database due to the much lower spatial resolution of the Lepton sensor. The ground-truth respiratory rate data were obtained with the respiratory monitor belt (Vernier RMB). The belt was strapped around a chest of a volunteer to collect a pressure during expansion and contraction of a body during breathing activity.

Examples of samples from each dataset are shown in Fig. 3.5.

Image Post-processing and Datasets Sizes

At first we extracted frames from collected thermal sequences. The frame rate for the SC3000-ADRA database was set to 30 FPS resulting in 3600 frames per volunteer (120-second recording per person). To introduce data augmentation (which aims at increasing number of samples in acquired dataset and thereby increasing data variety, what is potentially beneficial for model training) images in the Lepton-ADRA dataset were up-sampled from 9FPS to 12FPS. Then, similarly to the SC3000-ADRA set, recorded sequences were divided into single frames producing 720 samples per person.

For implementations of the proposed DL algorithms we utilize standard libraries and frameworks which by default support BMP, GIF, JPEG, or PNG image formats. Since, the raw data format of data captured for both SC3000-ADRA and Lepton-ADRA databaseswas set to 14 bits, acquired data had to be properly converted. The usual practice is to linearly down-scale raw data to 8-bit image formats. We followed this approach, producing 144000 8-bit PNG images in the SC3000-ADRA set (40 volunteers, 3600 frames per person) and 22320 8-bit PNG images in the



(a) SC3000-ADRA set, reference respiratory activity recorded with finger bend movement: finger up indicates inhaling, down - exhaling

(b) Lepton-ADRA set, reference respiratory activity recorded with respiratory monitoring belt

Figure 3.5. Examples of images acquired for Facial Regions Detection and Extraction of Respiratory Activity studies, respiratory activity can be seen by noticing differences in color intensities of nostrils area between image pairs of each volunteer (especially visible in Lepton set, for which camera was facing subjects upward)

Lepton-ADRA set (31 volunteers, 720 frames per person). However, this procedure may lead to a loss of some important details due to the bit compression.

Thus, we additionally generated 16-bit PNG images from raw data by upscaling the bit-depth from 14 to 16 to avoid contrast decrease. As a result, two additional sets were created with the same number of examples, but higher bit resolution (144000 in SC3000-ADRA and 22320 in Lepton-ADRA). In order to distinguish sets from each other, we use the following name convention: *camera name-abbreviation of target application-bit representation of produced images*, i.e. SC3000-ADRA-16 is a dataset collected with the FLIR® SC3000 sensor for studies on respiratory patterns analysis, data saved as 16-bit PNG files.

Reference Thermal Dataset

In order to avoid experiments being biased towards data collected by us, we decided to use publicly available thermal datasets as well. The IRIS [57] database consists of thermal and visible light images acquired for 30 individuals. For each person about 176-250 images have been recorded, 11 images per various head rotation. Thermal images have been acquired using the Raytheon Palm-IR-Pro camera with a spatial resolution of 320x240 pixels and stored as bitmap (uncompressed image format) graphical files with a depth of 8-bits/pixel. Visible light images have been captured using the Panasonic WV-CP234 camera with a 480TV lines horizontal resolution. Since our research focused on analysis of DL techniques for thermal image processing, we utilized only the thermal subset of the IRIS database (around 4190 thermal images).

Frames from the IRIS dataset can't be used for estimation of vital signs because it contains

only single frames not video sequences, so it's not possible to analyse dynamic changes within facial regions. Also, reference measurements of vital signs were't obtained during IRIS data acquisition. However, IRIS set can be used to verify accuracy and reliability of models aimed at resolution enhancement, what will be presented in Chapter 5. The main difference between our datasets and the IRIS database is that volunteers were asked to remain still while collecting our sequences. During IRIS data acquisition subjects were performing subtle head movements (11 images for each pose per person) and mimicking various emotions: angry, surprised, and happy. Examples of thermal images from the IRIS dataset are presented in Fig. 3.6.

Figure 3.6. Examples of thermal images from the reference thermal IRIS dataset including different head poses and facial expressions of volunteers

For facial areas detection studies, we also evaluated a possibility of transferring a knowledge from a visible light spectrum to a thermal domain. For these experiments, commonly used visible light image enhancement benchmark sets were applied. The Berkeley Segmentation dataset [154] combined with the SPSR data [155] (referred hereafter as BSD+SPSR) was utilized for neural networks training. The Set5 [156] was used for testing.

3.2.3 Emotions Recognition

Additionally, we wanted to evaluate other potential applications that could expand studies on contactless breathing estimation. Specifically, we were interested whether extracted vital signals could carry information about emotions which could be useful for other applications in remote medical diagnostics, e.g. analysis of users' satisfaction or mental health evaluation. Thus, we collected a separate dataset with recorded reference emotional responses to analyse whether similar states can be acquired from estimated vital signs.

Imaging Hardware

A multimodal input was analysed in the conducted emotion invocation study, i.e. heart rate estimated using imaging photo-plethysmography [157] acquired with a standard webcam - the Logitech 9000 Pro camera (30 frames per second at 640x480 resolution) and respiratory rate extracted from thermal sequences collected with the FLIR® Lepton camera. In this way, we were able to evaluate whether there is a correlation between vital signs and different emotional responses. Both invoked and simulated emotions were analyzed in our research. The motivation for our work, also

presented in [158], was based on the fact that the use of bio-signals instead of facial expressions is potentially very useful for emotion recognition task, as suppressing or masking biological responses is very difficult.

Data Collection Procedure

Due to previous successful applications of emotion-stimulating videos for collecting natural emotional responses [159] we decided to follow similar approach for data acquisition. Experiments were performed on a group of 11 healthy volunteers, age 33.7±11.3. Data were collected simultaneously from both visible light and thermal cameras to ensure proper synchronization between emotional responses estimated from extracted heart and respiratory rates. Devices were placed at a distance of approximately 0.5 m form volunteers, who were asked to look towards the camera and remain possibly still. Since the standard approach is to extract breathing signals by analysis of temperature changes in a nostril area, the thermal camera was aiming at a volunteer upward at an angle of 15° to make a nostril region more exposed.

An online questionnaire was used to introduce volunteers to details of the study and guide them through experiments. Taking part in the experiment required selecting a checkbox in the questionnaire with an agreement for data recording, otherwise sequences were not collected. At a first step of the questionnaire we gathered information about participants' age, diseases and ease of getting nervous. The rest of the data acquisition process was divided into two sections. In the first one participants were asked to imitate 4 emotions (neutral, joy, fear, disgust) for 1 minute, with 2-minute relaxation pauses between each emotion. A counter was visible in the questionnaire to help with time measurements. The participants were trying to simulate real emotions, not only facial expressions. As a result we collected 240 seconds of visible light and thermal sequences (60 seconds per each emotion) and 360 seconds of recordings corresponding to relaxation intervals. In this way we were able to determine whether there are differences between estimated vital signs for different emotions. All sequences were saved in original raw data format. Datasets acquired at this step are named Logitech900-ER-simulated and Lepton-ER-simulated for visible light and thermal sequences, respectively. ER abbreviation corresponds to Emotion Recognition study.

During the second part of the data collection process, a series of videos was presented to volunteers in order to invoke a real emotion. Selected vision stimuli included funny scenes from a gym (joy), eating worms (disgust), a dark basement with ghosts (fear) and ill animals (sadness). Similarly to the first part of the experiment, videos aimed at invoking emotions were separated by neutral clips to introduce relaxation pauses, e.g. an empty road, an ocean, snails and clouds. Since emotional response is highly dependent on individual perception, participants were asked to match each clip with emotions that were in their opinion dominant during watching it. Each emotion invoking video lasted 2 minutes, while each neutral video around 1 minute. As a result, we collected 8 minutes of visible light and thermal sequences corresponding to invoked emotions and 4 minutes of visible light and thermal recordings obtained during relaxation pauses (referred thereafter as Logitech900-ER-invoked, Lepton-ER-invoked). Fig. 3.7 and 3.8 present examples of visual light and thermal images collected while mimicking facial expression associated with different emotions and while being exposed to visual stimuli in order to invoke a real emotion.

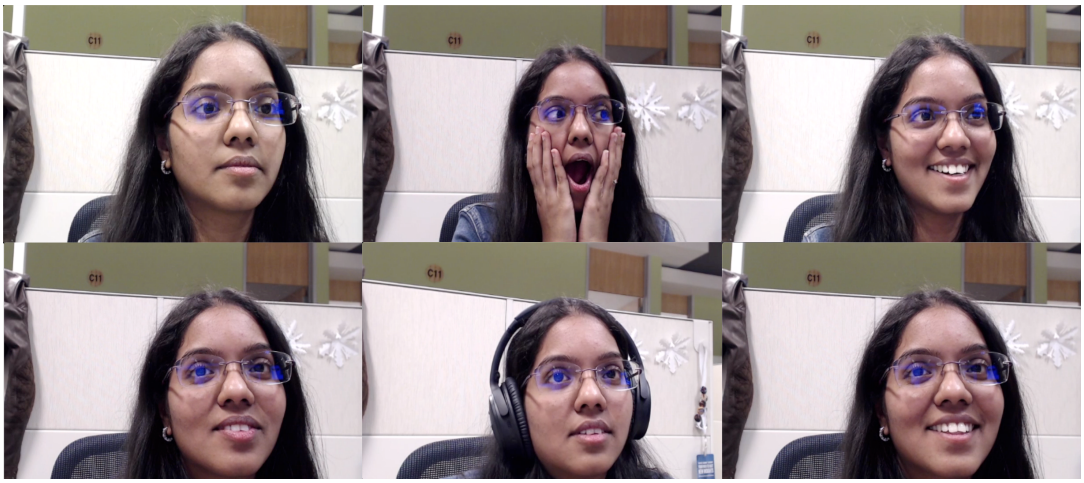


Figure 3.7. Examples of facial expressions in visible light data (Logitech900-ER-invoked and Logitech900-ER-simulated sets). From the left: neutral, fear, joy; top row simulated, bottom row invoked emotions; facial expressions are much more distinct for simulated emotions; author of the dissertation obtained permission for using pictures of selected volunteers in this work

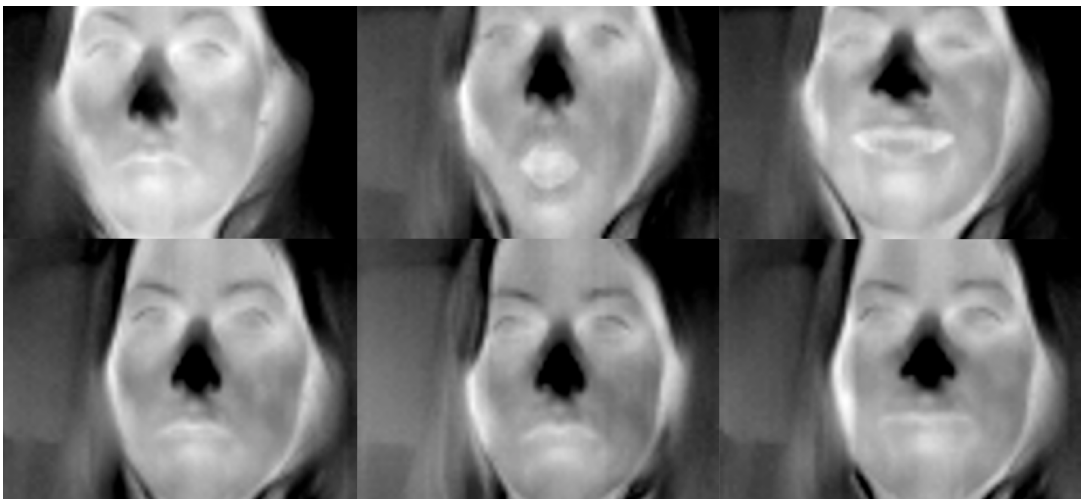


Figure 3.8. Examples of facial expressions in thermal data (Lepton-ER-invoked and Lepton-ER-simulated sets). From the left: neutral, fear, joy; top row simulated, bottom row invoked emotions; as can be seen facial expressions are much more distinct for simulated emotions, in our studies we will evaluate whether vital signs are also changed during various emotional states

3.3 Summary

This chapter overviewed details of thermal datasets collected and utilized by us to evaluate DL models aimed at detecting specific facial areas and enhancing resolution of thermal sequences. Both of proposed model architectures are explained in details in following chapters (Chapter 4 and 5). In this chapter a description of the publicly available IRIS database used as a reference set for verifying accuracy of the designed super-resolution neural network was also provided. Finally, we described the procedure of collecting data for emotion recognition studies that could be potentially useful for remote medical diagnostics, as described in Section 6.3.

Chapter 4

Proposed DL Methods for Facial Features Detection

4.1 Introduction and Overview

As described in Chapter 2, a wide range of methods and algorithms for facial features detection already exist and this subject has been studied in-depth in various applications. Yet, due to recent advances in Deep Learning (DL) techniques and ability to achieve human-like performance using recent neural networks, we are mainly interested in novel deep model architectures for the needs of remote medical diagnostics. In such solutions, thermal imaging is often favoured over visible light images due to privacy concerns and insensitivity to different illumination conditions.

Unfortunately, most of existing DL architectures are designed and tested on visible light images only. The important research question is whether existing neural networks successful in visible light domain can be directly applied to thermal imagery, which has different characteristics, i.e. contrast between adjacent regions is smoother due to a heat flow between objects. Another potential problem with thermal imaging is a limited number of samples that could be used for model training. Taking it into account, in our study we evaluate performance of state-of-the-art neural networks in a task of facial feature detection on thermal data collected by us. In this way, we verify whether similar accuracy could be achieved in other image domains than originally assumed. Secondly, a transfer learning technique, aimed at re-purposing already trained models to a novel task, is examined. Specifically, weights of networks trained on visible light datasets are reused on thermal sequences in order to determine whether it's possible to distinguish other feature representation using kernels trained to extract high frequency features (i.e. edge, corners, lines) present in visible light images.

Additionally, we propose a modification of existing DL classification models to restore features distribution and detect facial areas without the need of providing bounding box annotations during the training step. As a result, the time of data preparation, model training and inference can be significantly reduced. The proposed model architecture is compared against Single Shot Detector (SSD) network, commonly used for similar tasks [138].

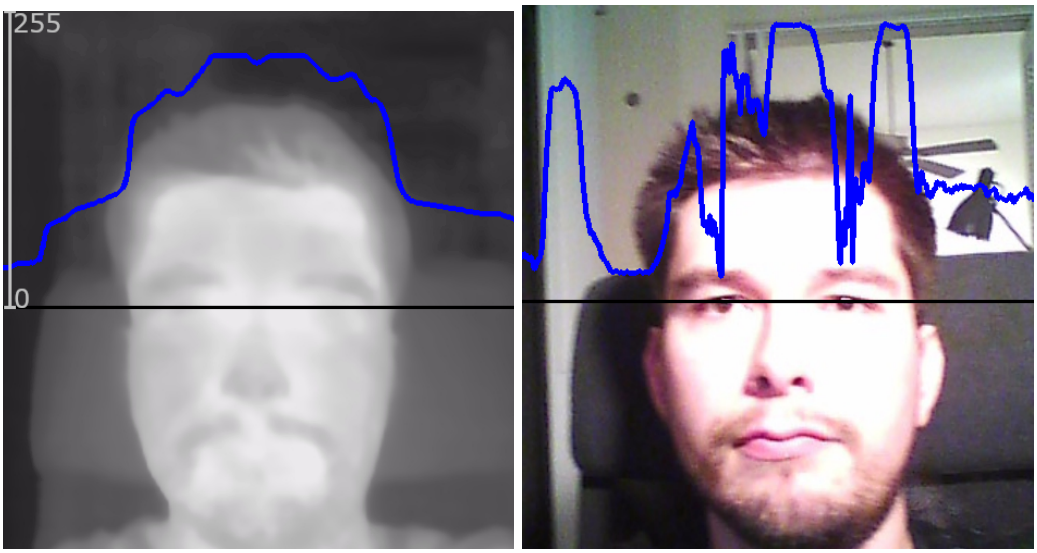
Finally, we provide details of experiments performed with novel DL architecture based on capsules instead of single neurons on our thermal datasets. Achieved results prove that this approach is insensitive to body rotation, what we find very useful for various remote medical diagnostic applications.

4.2 Facial Features Detection

4.2.1 Transfer Learning

Problem Formulation

Although Deep Neural Network (DNN) is a learning method and should be able to generalize well to thermal data, very often a worse accuracy is achieved by directly applying existing Convolutional Neural Network (CNN) trained on visible light images to data from different wavelengths of electromagnetic spectrum. This is caused by different characteristics of thermal imagery. A very important feature of thermal data is smoothness between different components present in the image, as shown in Fig. 4.1. Heat flow in objects leads to decreased values of temperature changes between facial areas, what is represented as a lower gradient (smooth change of pixel values) among adjacent image regions. This aspect of thermal imagery might have a huge effect on prediction accuracy using models learnt to classify visible light images characterized by high frequency features, such as sharp edges between object parts. Thus, training of a network on images from the target domain is a necessity. However, for thermal imagery this is a challenging task due to much smaller sizes of available datasets. This may have a significant influence on model accuracy, because one of the main reason for recent successes of DL is the increased number of samples that we could feed a model with, as explained in Chapter 3.



(a) Thermal image (b) Visible light image

Figure 4.1. Comparison of pixel values (blue plot) at the eye level marked with the black line. Thermal and visible light image taken simultaneously using FLIR® One Gen 2 camera. Dynamics of pixel values in the thermal image is much lower than in the corresponding visible light frame.

Luckily, some solutions for dealing with small database sizes already exist. One of them is transfer learning, firstly described by Thrun in 1996 [160], a technique used for re-purposing a network fully trained on some data to a novel task. Transfer learning approach has already been applied to DNNs in a solution named Deep Convolutional Activation Feature (DeCAF) [161]. Authors of DeCAF proposed to utilize weights of the model trained on ImageNet with 1000 categories in other visual recognition tasks with different classes. The intuition behind transfer learning approach is that kernels from lower level layers are trained to extract features common for various tasks, re-

gardless of target categories, e.g. corners, edges, basic geometric shapes. Hence, weights of lower layers in DL models don't have to be retrained, but directly re-used for other classification task. The re-purposing of a model is achieved by retraining only a fully connected layer (fc), responsible for matching extracted features to output classes and optionally a few top layers before to improve recognition accuracy. The transfer learning process is visualized in Fig. 4.2.

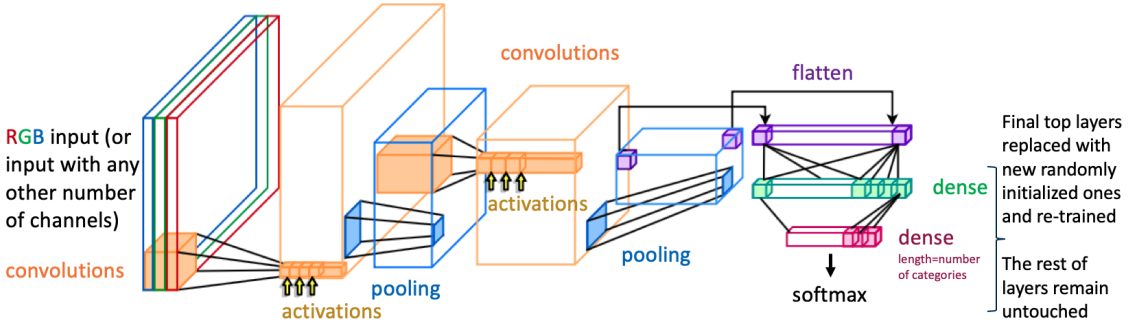


Figure 4.2. Visualization of the transfer learning idea on an exemplary schema of Convolutional Neural Network. Weights of all layers remain untouched and are pre-loaded to a model responsible for solving a novel task. Only final few top layers (usually final fully connected layer) is replaced with a new one that corresponds to a new task (i.e. has different length as number of categories may change) and re-trained.

Lately, some attempts have been also done for applying transfer learning to computer vision applications in medicine. The solution proposed by Wimmer et al. [162] focused on applying CNN originally trained on ImageNet to endoscopic images of duodenum in Celiac Disease analysis to extract feature vectors. A classification step, though, wasn't solved by the neural network, but using Support Vector Machines (SVM) algorithm. Lung pattern analysis was performed by Christodoulidis et al. [163] with a knowledge acquired on six publicly available texture databases and transferring it to lung tissue data.

Methodology

Inspired by existing studies on applying transfer learning to computer vision tasks in medicine, our research aims at evaluating whether models trained on visible light data could be re-purposed to perform facial area and features classification and detection from thermal sequences. To the best of our knowledge, our study was a first attempt to the problem of re-purposing activation maps retrieved from visible light data to a solution based on thermal images.

In our study [141], we evaluated classification model Inception v3 [164], proposed its modification to restore position of facial areas and compared the introduced pipeline to the detection model SSD [138]. Selection of those models was made taking into account their robustness in image recognition tasks. The Inception model makes use of convolutional filters of different sizes in so-called Inception modules to improve perception accuracy. In 2014 Inception won The ImageNet Large Scale Visual Recognition Challenge (ILSVRC), achieving 3.58% top-5 error rate, what is a factor of 5 less than Deep Neural Network predecessor AlexNet [109]. The reason for choosing SSD was motivated by its better accuracy, performance and training simplicity comparing to previous object detection models, as described in Section 2.3.2. Due to relatively small number of parameters and

utilization of a single network for an end-to-end solution, they can be used on embedded platforms, e.g. for autonomous driving [165]. This capability is also of a great interest to us because of target devices for which we design our remote diagnostics solutions, i.e. home-based remote person monitoring platforms or wearable systems for assisted living [166]. Results presented for both Inception and SSD were achieved on large-scale visible light datasets. Our study, though, focus on thermal data processing, which present temperature distribution and its changes over time, and thus have a different representation, as stated before. Therefore, networks have to be adapted to new data to produce correct predictions.

Moreover, images are usually pre-processed before feeding them to models to further improve prediction accuracy. Typically thermal data have a lower contrast than visible light data and exhibit a blurring effect, what makes their interpretability difficult. A commonly used technique for improvement of image contrast is histogram equalization, where we scale data distribution or a part of it from one range to a full image range. Similar operation is performed for conversion of raw thermal data to an output image range, i.e. a full raw data range is mapped to a full output image range. Yet, simple scaling, especially when the output range has fewer bits than original data, can lead to decrease of dynamics and contrast in a face region, as presented in Fig. 4.3. This assumption was verified by us in [141], where we confirmed on data acquired from 26 volunteers (Lepton-IE database, see Section 3.2.1) that automatic scaling of source values reduces the overall contrast of an image (output contrast value of 0.20 ± 0.03).

Figure 4.3. Thermal image of a face (cropped to a facial area) scaled from a full raw data range to a full output image range resulting in a complete loss of facial features visibility

Hence, we applied a pre-processing algorithm to images before feeding data to the model. Specifically we proposed to perform automatic fitting of Gaussian distributions to histogram data. As can be seen in Fig. 4.4, plotted histogram consists of distribution that can be modelled by two Gaussian bell curves. Since experiments were conducted in laboratory room with a controlled temperature, lower than a human body temperature, we assumed that the peak with higher mean corresponds to pixel values of facial areas. Thus, the right-side distribution was used to scale input values (mean $\pm$ standard deviation) to the output image value range. After this operation, image data interesting for us, previously characterized by close pixel values, was equally distributed, resulting in a higher contrast between facial areas, as presented in Figure 4.5.

After data preparation, we adapted models to a new task of thermal images classification and facial features detection. To avoid a need of performing a full network training from scratch, we

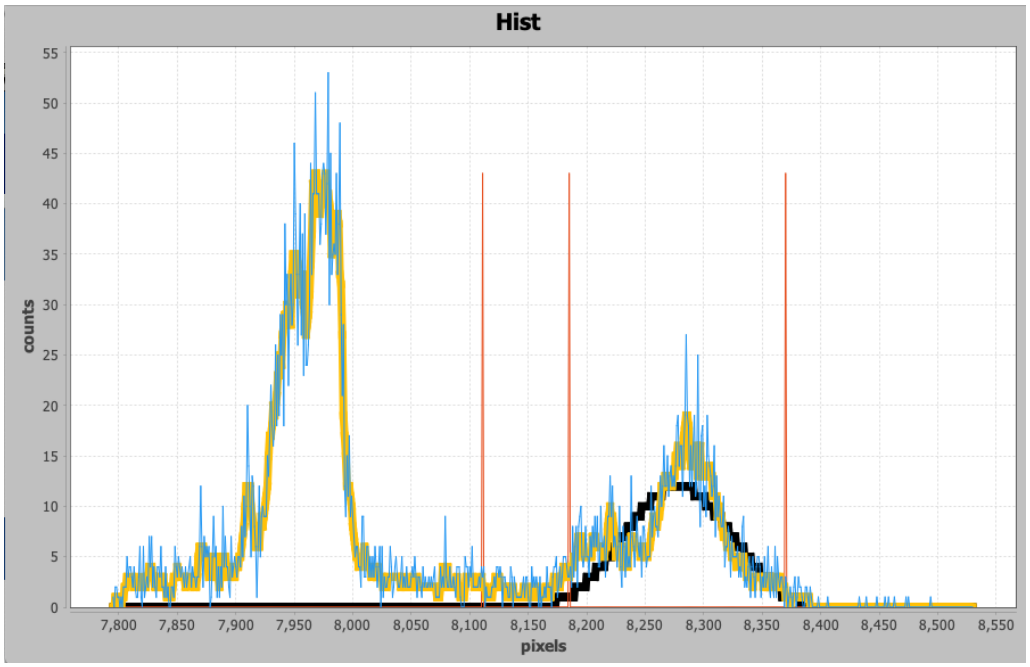


Figure 4.4. Histogram plotted for raw values of a thermal image of a face from Fig. 4.3

Figure 4.5. Image from Fig. 4.3 after applying the proposed by us pre-processing algorithms, where only pixel values of a facial area (right hand side Gaussian bell curve on Fig. 4.4) are re-scaled to an output image range; we can observe that contrast between features was significantly improved reveling specific facial regions

decided to evaluate the transfer learning approach and transfer knowledge learnt from visible light data. The following steps were performed for the Inception v3 model:

1. Feature Extraction - we utilized already trained weights (on ImageNet) of the model to extract feature representations from Lepton-IE dataset.

2. Fine Tuning - The final classification layer (Softmax operation) was replaced in the original model with a new randomly initialized layer responsible for performing classification of facial feature, i.e. eye and nose areas.

The default value of 4000 training steps and learning rate of 0.01 were used for the fine tuning step. Due to the scenario utilized for Lepton-IE dataset collection, which assumed presence of other objects in a background (typical scenario for an elderly person living independently), we decided to perform experiments of facial features detection with this database. In this way, we were able to examine conditions as close as possible to a target solution. As specified in Section 3.2.1, the used dataset contained of 485 images of an 'eye' category and 775 images of a 'nose' category. The re-training process involved an update of weights only for the added classification layer using back propagation algorithm with batch gradient descent optimizer and average cross entropy loss, defined as:

$$L = \frac{1}{N} \sum_{n=0}^{N-1} D_n(S, L) \tag{4.1}$$

where N is a number of samples in a batch, D_n is a cross-entropy function for n^{th} sample, L is a one-hot encoded labels vector, and S is a probability vector for output classes calculated using the softmax function, which for c^{th} class can be defined as:

$$S_c = \frac{e^{y_c}}{\sum_{c=0}^{C-1} e^{y_c}} \tag{4.2}$$

where y_c is the c^{th} class output score, which we want to turn into probability and C is the number of classes. In our study we decided to use a smaller batch size of 100, because it has been proved to lead to better generalization of models.

The same approach was applied by us in the study on face classification from complex background containing objects of temperature and shape similar to facial areas (e.g. measurement devices, lamps, other body parts) [167]. The transfer learning approach was utilized to re-purpose Inception model to binary classifier of thermal images (face vs other category). Furthermore, additional experiments were performed by us to evaluate possible limitations of contactless vital signs monitoring solutions. Specifically, we collected a dataset which allow for estimating influence of movements performed both by subjects and diagnosticians on the accuracy of face classification. Details about the acquired set, referred as Lepton-IE-M, were presented in Section 3.2.1. The model was trained using the same hyperparameters as in [141] and is referred thereafter as Inception-Lepton-IE-M.

To confirm more intense movement content present in scenarios where volunteers were not focused on a specific task or when a camera was placed on a wearable platform, Sum of Absolute Differences (SAD) per pixel was calculated. This metric has been previously proved to be successful in estimating spatial-temporal activities in video inputs [168]. In addition, research on the effect of subjects' movement on accuracy of non-contact breathing rate estimation was also conducted by us to determine how it affects smart home and telemedicine applications that we cover in this work. More details and results of this study are presented in [151].

Results and Discussion

Table 4.1 presents evaluation metrics calculated on the validation part of the Lepton-IE set for Inception and SSD DL models used for initial evaluation of face classification/detection task in thermal images by transferring knowledge from visible light data. Results for the Inception model are calculated for a classification task, our proposal of turning it into detection pipeline at inference step is presented in details in the following section (Sec. 4.2.2), followed by comparison of its performance to a detection network. Number of true positives (face classified as face) and

false positives (other object classified as face) of Inception model re-purposed from visible light data to the thermal Lepton-IE-M set are shown in Table 4.2. True positives calculated for each data collection scenario separately, i.e. S1 - camera mounted on a tripod, volunteers focused on a task, S2 - camera mounted on a tripod, volunteers performing small head movements, S3 - camera mounted on wearable platform, volunteers performing small head movements, as explained in Sec. 3.2.1. In order to verify presence of higher motion content in scenarios 2 and 3, Sum of Absolute Differences (SAD) per pixel metric was calculated (Table 4.3).

Experiments conducted for classification and detection Deep Neural Networks showed that it is possible to re-purpose models trained on huge amount of visible light data to a novel task in a different domain where only limited number of samples is available. Even though representation of thermal facial features differ from those obtained in visible light, both models achieved very high precision values by using transfer learning with only ˜2000 samples (Table 4.1: Inception 99.51%, SSD 91.44%). On the other hand, the recall of the SSD model was much worse due to a big number of false negatives. A possible reason of this result is a huge similarity of facial regions in low resolution images caused by feature blurring and thus low confidence of both classes. We would like to address this problem in the next chapter (Chappter 5) by introducing thermal image enhancement algorithm based on DL.

Table 4.1. Precision, Recall and mean Average Precision on the validation part of the Lepton-IE set for Inception and SSD DL models used for initial evaluation of face classification/detection task

Metric	**Inception (classification)**	**SSD (detection)**
Precision [%]	99.51	91.44
Recall [%]	99.12	36.21
mAP nostril class	1	0.77 (IoU 0.5)
mAP eye class	0.98	0.36 (IoU 0.5)

Table 4.2. True positives and false positives for a face class of Inception model re-trained on the acquired thermal Lepton-IE-M set

Prediction	**Ground-truth**			
	face			other
	S1	S2	S3	
face	89.2%	87.8%	82.5%	0.43%

Analysis performed for potential remote medical diagnostics scenarios confirmed that involuntary motion of volunteers can be reduced by focusing their attention on some tasks (SAD 0.26 for scenario 1 vs 0.28 for scenario 2). This is an important finding for design of contactless vital signs monitoring solution which indicates that collection of measurement should happen during some activities, as it allows for improving accuracy of face classification with DNNs. Number of true positives in scenario 1 (a person focused on reading) was improved by ˜1.5% comparing to the second use case where subjects were performing some small, almost involuntary movements.

Table 4.3. Sum of Absolute Differences (SAD) per pixel for all 3 scenarios (S1, S2, S3) of possible remote medical diagnostics (Lepton-IE-M dataset)

	Subject											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Avg**
S1	0.19	0.18	0.35	0.34	0.26	0.24	0.23	0.22	0.20	0.26	0.36	0.26 ± 0.06
S2	0.22	0.21	0.34	0.38	0.26	0.27	0.29	0.39	0.21	0.29	0.25	0.28 ± 0.06
S3	0.48	0.51	0.41	0.56	0.38	0.45	0.43	0.37	0.30	0.40	0.36	0.42 ± 0.07

In addition, it has been shown that value of the SAD metric is much higher if both diagnostician's and subject's motion is present in collected sequences. This relation has been also confirmed by classification results produced by the Inception model. It can be observed that number of true positives drops significantly for scenario 3, where camera was mounted on wearable eGlasses platform. Although overall system precision for scenario 1 is very high (99.5%), in order to enable various possible remote medical diagnostic solutions (e.g. measurements done in a contactless way by a specialist during annual health checkup or hospital visit), the presence of higher motion content should be taken into account. One reason for lower performance in a presence of motion is a variety of camera angles and body poses that a subject can be in. As a result, CNN which is sensitive to object rotations may produce worse results. Some neural networks allow for mitigating this problem by introducing rotation-invariant architectures. More details about them and results achieved by us using such solutions in thermal spectrum are presented in Section 4.3.

4.2.2 Restoration of Features Distribution

Problem Formulation

The use of transfer learning approach and retraining of the softmax layer allowed for obtaining facial feature classification model in an easy, less time-consuming way when the number of thermal samples was limited. Yet, the problem of getting coordinates for each facial region remained unsolved. The results produced by the retrained Inception model contain only classification details, i.e. probability of each class at the image level. To be able to design an automatic way of vital signs extraction by analysis of color changes within facial regions, there is also a need to obtain position and size of those regions.

Various studies on facial region detection from thermal images have been already conducted, as explained in Chapter 2. Yet, many of them are based on hand-crafted representations what might be a limitation of applying them in real life problems due to difficulties with defining universal sets of features and sensitivity to changing measurement conditions. Moreover, some accurate facial areas detection methods [26] use geometric details which often require the presence of face boundary in a frame. In case of sequences cropped and zoomed to a middle part of face only (e.g. while using frontal camera in smartphones), this boundary may not be visible affecting information about anthropometric measurements and relations between facial regions. In such cases, DL-based solutions are very useful as they allow for detecting specific areas regardless of the presence of other objects. Since our proposed approach is based on DNN, the imitation of a lack of facial boundary in a frame is not a concern. The standard DL-based solution to the task of facial areas

localization is to use object detection networks. With the progress of DL, this problem has been studied in depth and various detection architectures already exist, as described in Section 2.3.2. Hence, in our study, we propose a different, novel approach to object detection by utilization of classification models modified during inference. Since spatial distribution of features is restored from CNN at a run time, there is no need for training model with region generation steps and thus for providing datasets with bounding box annotations. As a result, the dataset preparation step is simplified. Also, our proposed algorithm can be applied to any already trained CNN, not only to the Inception model, providing very useful information about locations of classes what may be utilized in various other applications as well, e.g. pedestrian detection in autonomous driving [165] or person identification in smart home/office environment [169].

Methodology

To better explain the proposed modification of the classification model flow, we will start by revisiting the architecture and basic building blocks of CNNs. In the simplest setting, CNN can be visualized as a pyramid, as presented in Fig. 4.6. We start by feeding an input to the first layer. In general let's assume the input is represented as image data, so it has some width, height and 1 or 3 channels depending on color space settings (e.g. grayscale has 1 channel, while RGB has 3 channels). The input is then convolved with n number of filters, producing a deeper output representation (the depth corresponds to the number of applied filters n). At the same time the spatial dimension can be reduced by using stride above 1 or *valid* padding, parameters described in details in Section 2.3.2, producing activation maps of smaller output size (S_{out}), defined as:

$$S_{out} = (S_{in} - S_k + 2p)/s + 1 \tag{4.3}$$

where S_{in} is the input size, S_k is the kernel size, s is a stride and p is a padding, equal 0 for *valid* padding and 1 for *same* padding. Another option for spatial dimension reduction is a pooling operation applied just after convolution.

Decreasing of width and height of representations at each step is not a necessity and various other approaches have been already proposed, including downsampling later in the network to increase network accuracy [170]. Here though, we focus on the most standard version of CNN architecture, where at each hidden layer the depth is increased, while reducing the spatial dimension. After the pyramid of convolution and pooling operations, a fully connected layer is usually utilized, which maps extracted features to specific categories. Thus, the output is represented as a vector of a length equal to number of categories, producing predictions at the image level, with the complete loss of information about features distribution.

A key concept in turning classification models into a region detection pipeline lies in the removal of the final pooling layer. In case of the Inception network, after passing data through all hidden layers, the representation of a size 8×8×2048 (height x width x number of features) is produced. The final pooling turns it into a 1-dimensional vector of features (length 2048) which are mapped to output classes. Prediction is performed by selecting the maximum value from each of 2048 slices of a size 8x8. The intuition behind it is that the maximum value carries the most significant information for the image, so it's best to use it for classification. However, in this way we loose localization information. Instead, we propose to preserve all 64 values (8 rows x 8 columns) from each of 2048 slices and map it to specific regions of the image in order to produce class predictions in various locations. At first, we feed an image into the model and run it through all layers except the final pooling layer to produce the 8x8x2048 representation. The output feature map is then

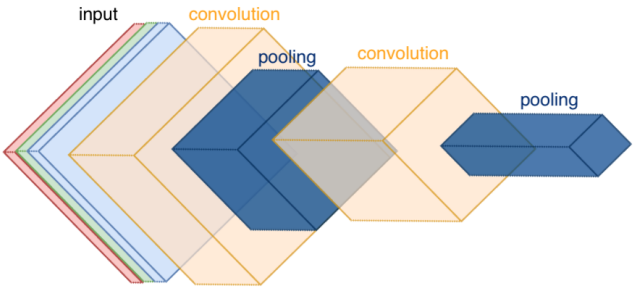


Figure 4.6. CNN can be visualized as a pyramid, after each convolution a depth of representation is increased, while a spatial size is reduced due to applying pooling operation, stride and padding

divided into 64 vectors of a length 2048. At the same time, we divide the original image into an 8x8 grid with each cell corresponding to one of the produced vectors. The vectors are then passed through the already trained final classification block (fully connected layer and softmax operation) to output probabilities of all classes at the cell level. For each location i, j (row, column), the output probability for class c is defined by the softmax function:

$$S_c^{i,j} = \frac{e^{y_c^{i,j}}}{\sum_{c=0}^{C-1} e^{y_c^{i,j}}} \tag{4.4}$$

In this way, we know which category has the highest probability at each grid location (i, j for $i \subseteq < 0, 7 >$ and $j \subseteq < 0, 7 >$). Neighbouring cells with the same output category can then be merged to form bigger regions based on produced cell-level predictions. Fig. 4.7 presents face area divided into the grid with cells corresponding to vectors acquired from the feature map before final pooling and cells activated by each class.

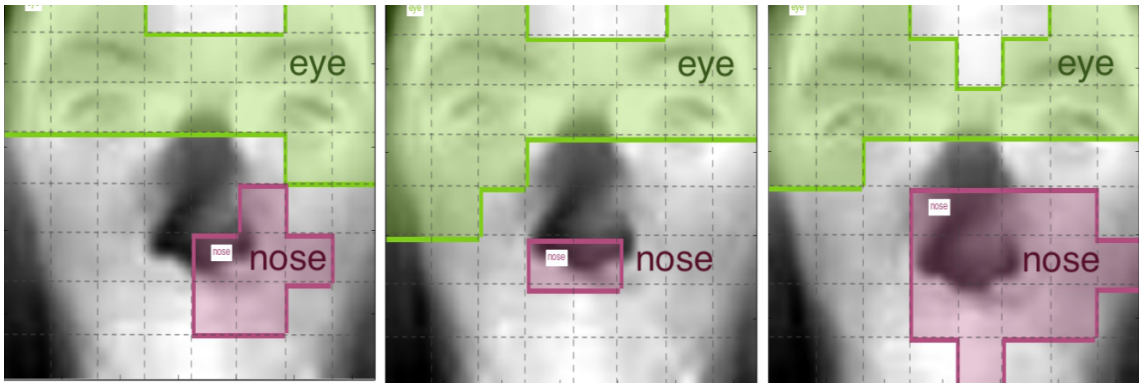


Figure 4.7. Examples of images with cells activated by nose and eye classes; constructed grid of cells correspond to feature map vectors acquired from the representation before final pooling

Highlighted areas were then post-processed in order to form final facial regions. Nostril and eyes regions were constructed from merged cells by taking top left and bottom right corners of boundary cells. Single (isolated from other groups) cells were skipped. The modified flow of the Inception model is presented in Fig. 4.8. The proposed approach is also described using following function in programming language Phyton:

Listing 1. The proposed modification of the classification inference flow

```
import tensorflow as tf

def function calculate_cell_predictions(image):
    """Calculates prediction for each cell in the input image grid.

    Args:
        image: input image data loaded into numpy array

    Returns:
        Dictionary with predictions for each cell
        [row][column] = vector of a length corresponding to number of categories with calculated probabilities.
    """

    with tf.Session() as sess:
        # Load trained graph from protobuf format
        load_CNN_graph(file_path)

        # Get necessary tensors
        input_tensor = tf.get_default_graph().get_tensor_by_name("input:0")
        extracted_features_tensor = tf.get_default_graph().get_tensor_by_name("mixed_10:0")
        softmax_tensor = tf.get_default_graph().get_tensor_by_name("softmax:0")

        # Obtain feature maps before final pooling
        feature_maps_8x8x2048 = sess.run(extracted_features_tensor, feed_dict={input:image})


        cells_probabilities = {}
        for cells in feature_maps_8x8x2048:
            for row in range(0, len(cells)):
                for column in range(0, len(cells[x])):
                    cell = cells[row][column]
                    cell = tf.reshape(cell, (1, 1, 1, 2048))
                    # pool_3:0 is an input to the softmax
                    cell_prob = sess.run(softmax_tensor, feed_dict={"pool_3:0":cell})
                    cells_probabilities[row][column] = cell_prob[0]

        return cells_probabilities
```

In order to evaluate our proposed approach, we compared it to the SSD model, described in Section 2.3.2. SSD is frequently used for various computer vision applications because of high accuracy and utilization of a single network for end-to-end detection pipeline. SSD, as the detection DNN and similarly to our modified classification model, utilizes CNN to extract feature maps during feed-forward pass. The difference lies in the rest of the network, where an auxiliary structure is added to produce detections. This additional step requires data to have bounding box annotations to learn the presence of objects at different locations and adjust scale and ratio of produced locations to match the ground-truth bounding boxes.

Since the SSD network is also a DL model which requires a lot of data to learn how to extract proper predictions and the size of our thermal database with annotated facial areas is limited (the Lepton-IE set contains only 1260 train, 160 test, and 187 validation images), we utilized the transfer learning approach, similarly as in the classification task. All weights of the SSD model trained on visible light data except the final prediction layer were loaded into the network aimed at performing thermal facial features detection. The train set was utilized to adjust weights of final layers, at first randomly initialized to re-purpose the model to our application during 30000 training steps using RMSprop optimizer [171], unpublished algorithm first proposed by Geoff Hinton in his 'Neural Networks for Machine Learning' course. All other hyperparameters were set to their default values,

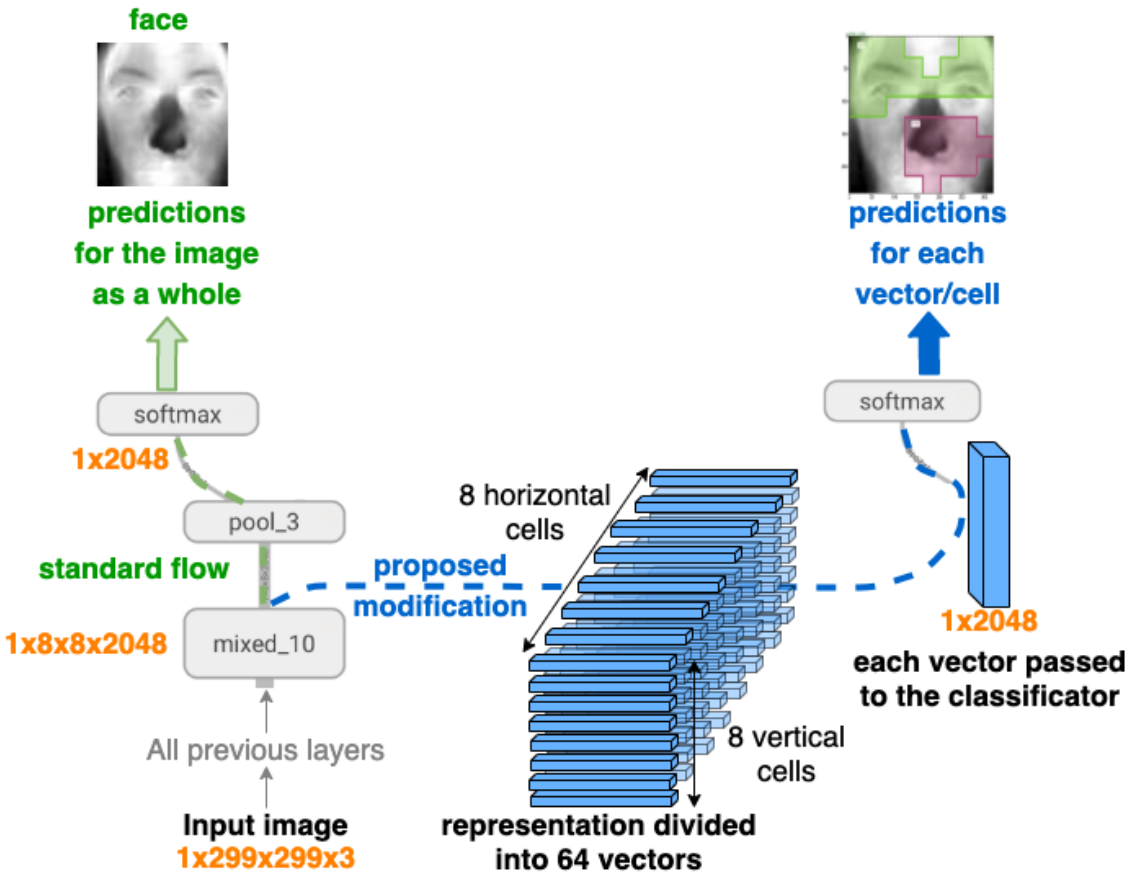


Figure 4.8. Proposed modification of the inference flow in order to restore distribution of features from the classification model and detect facial areas

as specified in a TensorFlow model implementation [1].

Our proposed modification of the Inception model was also examined on the dataset collected with possible scenarios of remote medical diagnostics in mind (Lepton-IE-M set used for evaluation of motion influence). At this step, we utilized the Inception network already trained in our experiments focused on evaluating possibility of transferring knowledge between image domains, described in the previous section (i.e. Inception-Lepton-IE-M model was used). The model already optimized on our thermal dataset was modified during the inference to restore distribution of features and mark facial areas. In addition, we trained a second binary Inception classifier capable of determining nostril areas from other objects. This network was optimized in the same way as the Inception-Lepton-IE-M model using nose regions extracted from face images as one category and frames of other objects as a second category. After localizing faces with the proposed modified Inception flow, nostril areas were marked in detected areas using the same approach. Robustness of the proposed solution was verified by calculating Root Mean Squared Error (RMSE) of average pixel intensities in detected areas vs. regions marked manually by an expert. For correctly detected areas, this error should be minimal. In order to verify this assumption we also compared RMSE of average pixel intensities in a static location of face and nose (i.e. the same coordinates of facial regions for the whole video sequence instead of adjusting them for each frame) with areas marked manually by an expert. In this way, we were able to evaluate if motion has influence on signals

[1] https://github.com/tensorflow/models Accessed: 2017-05-01

obtained from detected areas, i.e. in a case of a significant motion, areas important for signal extraction may move out of the static region leading to big changes in average color values used for vital signs analysis [172].

Results and Discussion

Evaluation metric calculated on the validation part of the Lepton-IE set for SSD network and the modification of Inception model flow proposed to restore distribution of facial features and perform detection of eye/nose areas are presented in Table 4.4. Both pipelines were re-trained from networks initially tuned on visible light images to verify possibility of transferring knowledge between domains, as explained in the previous part of experiments. Comparison is performed by analysis of Intersection over Union (IoU) between areas detected with neural networks and regions marked manually by an expert.

Table 4.4. IoU calculated for the proposed thermal feature detection method and the reference DL architecture SSD

IoU detected areas vs manual annotations	**Proposed method (modified Inception flow)**	**SSD (all results)**	**SSD (False Negatives not considered)**
Eye area	0.53 ± 0.15	0.32 ± 0.38	0.84 ± 0.23
Nostril area	0.60 ± 0.18	0.55 ± 0.42	0.86 ± 0.05

Table 4.5. Time of a single inference and training pass for the proposed modification of Inception model flow and the reference object detection SSD model

Platform	**Proposed method (modified Inception flow)**	**SSD**
Inference - single pass [ms], batch size=1		
Intel® Xeon® E5-2697v2	139 ± 23 (2% util.)	596 ± 39 (4% util.)
NVIDIA® DGX-1™ Station	62 ± 7 (5% util.)	531 ± 62 (4% util.)
Training - single pass [ms], batch size=32		
Intel® Xeon® E5-2697v2	9513 ± 102 (45% util.)	7358 ± 180 (55% util.)
NVIDIA® DGX-1™ Station	201 ± 8 (50% util.)	167 ± 9 (50% util.)

Time of processing a given batch size of images for SSD network and the modification of Inception model flow proposed to restore distribution of facial features and perform detection of eye/nose areas is collected in Table 4.5. Inference time is presented as a forward pass, while the training time includes both forward and backward pass. Batch size (number of samples processed simultaneously in the same network pass) was set as 1 for inference taking into account target applications based on real-time processing of collected sequences.

Facial regions detected with evaluated Deep Neural Networks from the Lepton-IE dataset are presented in: Fig. 4.9 - bounding boxes constructed from highlighted grid cells produced with the proposed method and Fig. 4.10 - output areas of the SSD model. Nostrils were marked with red, eyes with green and ground-truth annotations with blue colors. In a case of the Lepton-IE-M set, experiments were conducted for nose and whole face regions. Results shown in Fig. 4.11 present produced locations of those areas.

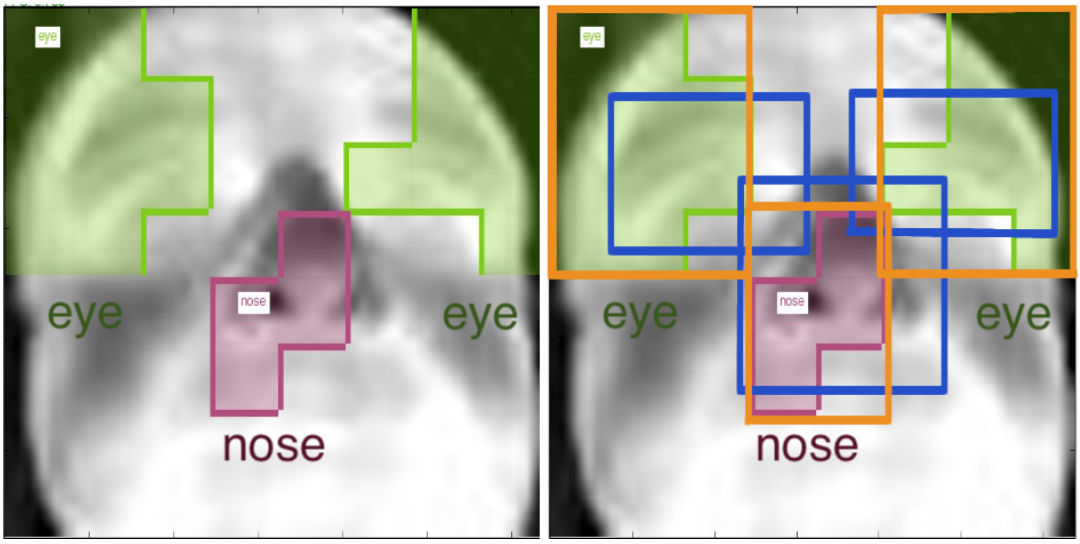


Figure 4.9. Facial areas detected with the proposed method based on the modified inference flow of the classification model; left: image before bounding box post-processing; right: final constructed boxes are marked in orange, blue areas show ground-truth annotations

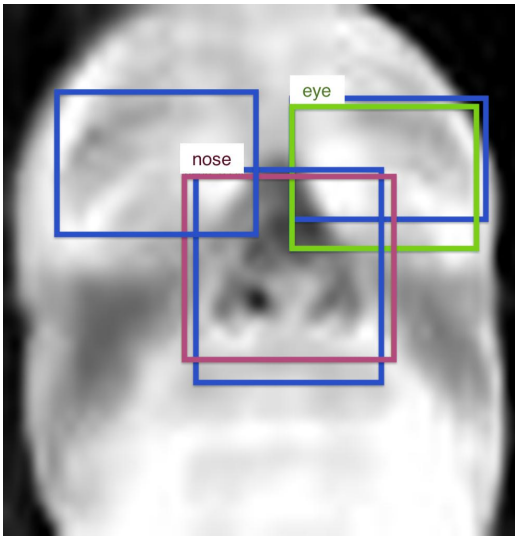


Figure 4.10. Facial areas detected with SSD model; detected eye and nose regions are marked in pink and green, blue boxes represent ground-truth annotations; please note lack of the right eye detection (false negative) - this problem was very often experienced by us for SSD model trained on thermal data

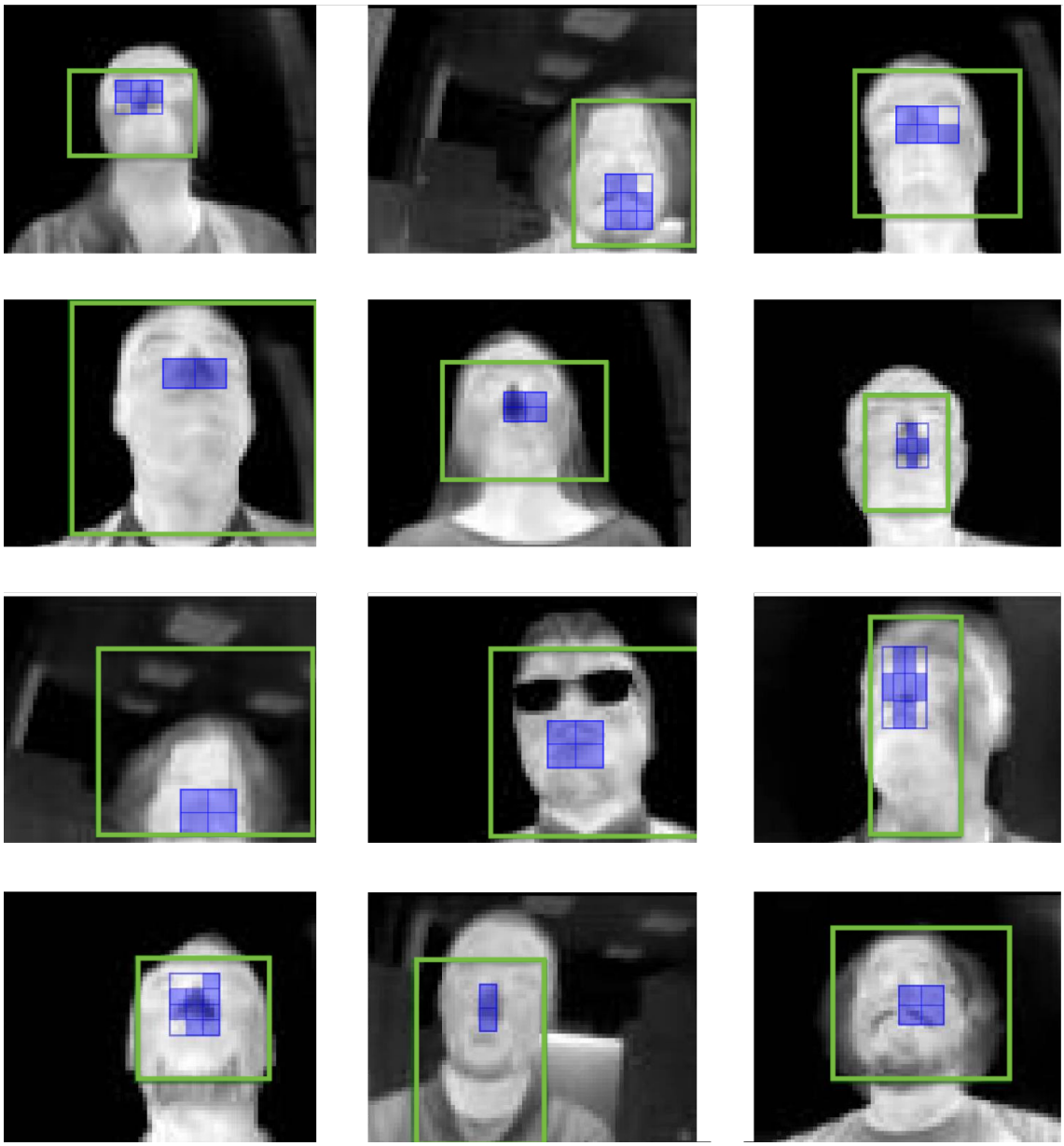


Figure 4.11. Face (green) and nostril (blue) areas detected with the proposed DL-based method; even for a higher motion content (camera placed on eGlasses) when only a part of a face was visible, the model was still able to properly detect specific regions (e.g. 1^{st} image in a 3^{rd} row)

Our study focuses on contactless monitoring of human subjects in remote medical diagnostic scenarios. Since vital signs are constructed by aggregating pixels intensities within detected facial regions, we evaluated how motion affects signal constructed by averaging pixel values in face and nose areas. For this, we utilized two types of regions: detected with the proposed model and static region marked manually in a first frame of each volunteer's sequence and applied to all remaining frames. Location marked manually in each frame by an expert was used as a reference. Then, the average value of pixel intensities within each of those 3 areas was calculated and analysed. A final Root Mean Squared Error for each volunteer was produced by taking mean value of Root Mean Squared Errors of each frame, i.e. errors between average value of pixels in the reference region and corresponding average pixels value for detected/static locations. Results are presented in Table 4.6.

Table 4.6. Root Mean Squared Error (RMSE) of average pixel values in the static (same location for the whole sequence) and detected areas compared against areas marked manually by an expert; results normalized by a color range; tests performed on the Lepton-IE-M dataset

RMSE for facial area						
Subject	**Detected location**			**Static location**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
1	0.100	0.118	0.090	0.026	0.080	0.071
2	0.145	0.069	0.047	0.143	0.056	0.057
3	0.097	0.173	0.075	0.014	0.070	0.170
4	0.080	0.129	0.057	0.055	0.156	0.107
5	0.169	0.094	0.159	0.320	0.032	0.079
6	0.078	0.088	0.018	0.028	0.044	0.105
7	0.028	0.048	0.051	0.009	0.060	0.121
8	0.286	0.095	0.080	0.130	0.037	0.089
9	0.157	0.150	0.087	0.031	0.011	0.126
10	0.079	0.118	0.080	0.035	0.034	0.087
11	0.086	0.113	0.084	0.101	0.044	0.040
Avg.	0.119 ± 0.069	0.109 ± 0.035	0.075 ± 0.035	0.081 ± 0.092	0.057 ± 0.038	0.096 ± 0.036
RMSE for nostril area						
1	0.238	0.336	0.197	0.257	0.210	0.260
2	0.307	0.224	0.132	0.347	0.231	0.248
3	0.295	0.285	0.168	0.349	0.294	0.181
4	0.318	0.106	0.149	0.335	0.107	0.116
5	0.049	0.096	0.162	0.071	0.031	0.170
6	0.152	0.136	0.081	0.195	0.151	0.082
7	0.104	0.143	0.176	0.067	0.145	0.167
8	0.090	0.115	0.095	0.206	0.036	0.152
9	0.152	0.066	0.083	0.130	0.148	0.093
10	0.079	0.102	0.106	0.092	0.022	0.107
11	0.153	0.149	0.102	0.068	0.058	0.055
Avg.	0.176 ± 0.098	0.160 ± 0.085	0.132 ± 0.041	0.192 ± 0.115	0.130 ± 0.090	0.148 ± 0.066

The presented study aimed at proposing a novel method for modification of deep classification neural networks at the inference time to restore spatial distribution of features and thus detect facial areas that could be used for estimation of vital signs. Calculated values of the IoU metric (Table 4.4) show that areas detected with the state-of-the-art SSD network are more precise but only if false negatives are not taken into account (IoU for nostrils ˜0.86 vs. 0.60 for the proposed method). However, if all outputs are considered, the proposed method produces almost twice as good results for eye regions and 5% better results for nostrils than SSD. This is caused by the fact that SSD is very sensitive to false negatives. In many input images, facial regions were not detected by SSD model at all. As a result, we believe this method might not be appropriate for remote medical diagnostics solutions, where detection of those areas is crucial to extract vital signs. A possible reason for such results is similarity of facial areas and their low contrast in small resolution thermal sequences, as presented in Fig. 4.12. We would like to verify this assumption by increasing resolution of acquired sequences using DL. Details about those experiments will be presented in the next chapter.

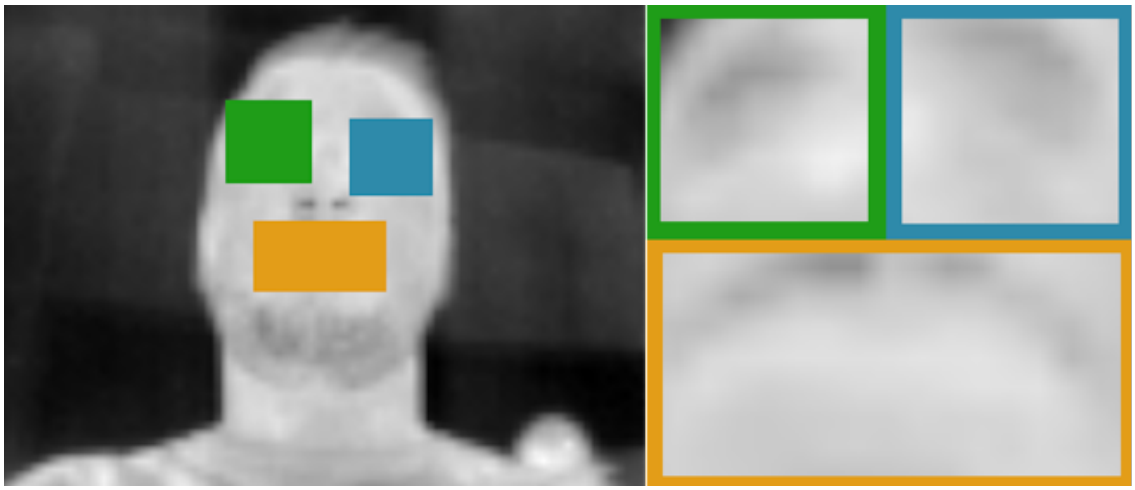


Figure 4.12. Enlarged facial features from thermal image acquired with the low resolution Lepton camera (80x60)

Another important finding is that IoU was higher for nostril area than for eye area in case of both models. We believe this could be caused by the use of imbalanced dataset, as ˜30% more samples were used for the nostril class. This result is especially valid for SSD model, which turned out to be sensitive to unequal distribution of classes within the used set. SSD predictions were shifted toward the most common class, producing almost twice as high precision for nostril region than for eyes. The proposed network modification turned out to be less prone to the class imbalance problem. Yet, a significant limitation of the introduced method is a resolution of detected regions, restricted by the number of cells in the constructed grid (corresponding to the size of the feature map thus dependant on number of filters applied in all network layers). In some cases the location of the detected area is correct, but it may be significantly smaller than the cell size or only a small part of a facial region may be present is a cell, while most of it may occupy other cells. As a result, more cells than required may be highlighted, leading to detection of inaccurate boxes, which contain other objects or background elements as well. To mitigate this problem, a post processing algorithm should be utilized to adjust borders of detected regions.

On the other hand, the proposed method posses some advantages over the examined SSD model

which include simplicity of network training and fast inference time. First of all, feature distribution can be restored at a run-time allowing for utilization of any already trained classification model. There is no need for providing datasets with bounding box annotations for model optimization, what is important in thermal image domain due to a limited number of publicly available thermal image sets. Secondly, calculated image processing time (Table 4.5) proves the robustness of the introduced method for live video stream processing. The proposed classification model modification achieves ˜16 FPS on NVIDIA® DGX-1™ Station and ˜7 FPS on Intel® Xeon®, almost 10 and 4 times more respectively than SSD on the same platforms. Since the used thermal camera is capable of acquiring 9 FPS, our solution is suitable for real time processing (for NVIDIA® DGX-1™ Station, for Intel® Xeon® some frames could be skipped to match 7FPS). In addition, it has been shown that utilization of compute resources is minimal (below 5% in all cases). Thus, the future work can be twofold. We could either focus on improving processing time of single stream increasing performance or simultaneously serve multiple models/subjects without impacting latency, e.g. in a centralized health monitoring station.

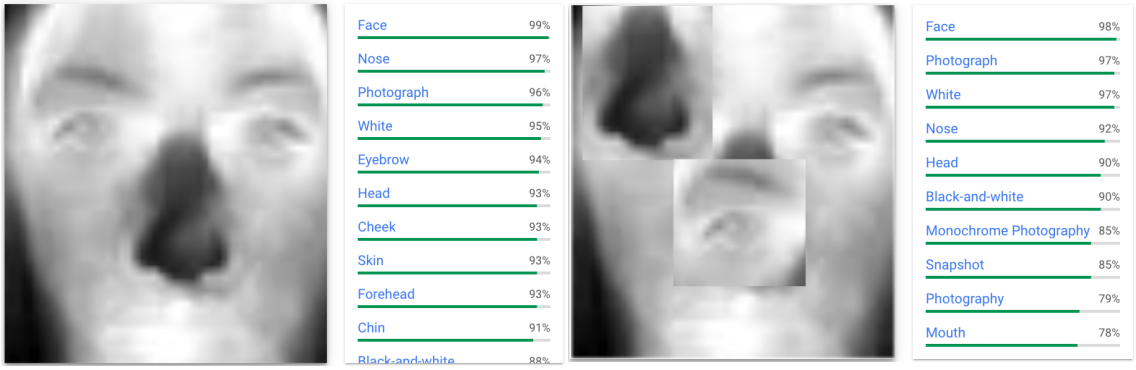
Comparison of RMSE values between static and dynamically detected areas showed that the size of the region has influence on results. In case of face region, errors were smaller when the static location was used. Yet, for nose areas the results were opposite. Although volunteers were asked to remain still, the presence of some involuntary movements is still possible. For bigger areas, this may not influence results, but for smaller regions, it may result in interesting feature moving entirely beyond the detected location. One solution to this problem is to detect dynamically interesting regions in the first frame and monitor the motion content in the remaining ones, e.g. using SAD metric as showed in the previous section (Section 4.2.1). If a bigger movement is detected, locations could be re-adjusted using CNNs to improve accuracy of signals used for vital signs estimation.

4.3 Novel Architectures Insensitive to Body Rotations

Problem Formulation

Although results achieved for facial areas detection and transferring the knowledge from visible light to thermal image domain were very promising, solutions presented so far were based on the assumption that a person is looking towards the camera being as still as possible. We also proved that more intense movement content affects the accuracy of areas detection [167] and consequently the accuracy of non-contact vital signs estimation [151]. However, in order to make remote diagnostics solutions convenient and unnoticeable by subjects during their daily activities, it's essential to provide algorithms insensitive to various body poses, e.g. tilting head, or lying down. Scenarios, that do not impose specific behaviors on users, may lead to more trustworthy results, as subjects behave more naturally often without noticing that the diagnosis is in progress.

Most of current object detection oriented solutions utilize CNNs due to their human-like performance and capabilities to significantly outperform other machine learning techniques based on hand crafted features, such as Viola-Jones algorithm [32]. Yet, convolutions have been also proved to be sensitive to various data distortions, such as displacement of features or image rotations. CNNs only look for the presence of features, they do not determine if spatial relations between them are preserved, as presented in Fig. 4.13. Some solutions to deal with this problem have been already proposed, including DL-based pipelines, e.g. Deep Dense Face Detector, which classifies a face in different orientations [173] or facial landmark alignment [139], as well as pose estimation



(a) Thermal image of a face

(b) Thermal image with displaced facial features

Figure 4.13. Visualization of CNN limitation - lack of spatial relation between learnt features; image with displaced features is still classified as a face, as CNN only looks for a presence of features, not relations between them; categories produced with https://cloud.google.com/vision/

techniques [174] and methods based on the use of additional sensors [175]. On the other hand, such solutions require additional memory and computational resources, what we would like to avoid due to the target platforms that we address, such as smart home or wearable devices (e.g. eGlasses [150]). That's why algorithms optimized for resource-constrained inference are often preferred in scenarios considered by us. Also, because of person's data privacy, thermal imaging is usually a better choice as it doesn't reveal sensitive details of captured objects contrary to RGB data.

To the best of our knowledge, our work is a first attempt to apply a novel rotation invariant NN based on capsules to thermal data. This evaluation is an important research in order to confirm that architectures successful in visible light domain characterized by high frequency features can be used for processing data from other lengths of electromagnetic spectrum.

Methodology

The same dataset as in the case of transfer learning and facial detection experiments was used to compare capsule network and CNN (the Lepton-IE set acquired by us, see Section 3.2.1). Facial images were extracted from recorded sequences, producing a set of 3256 samples (every ~6^{th} frame was preserved to ensure uniqueness of images). Since collected images were the same as in our facial areas detection study [141], there was a need to increase the visibility of specific face regions. The contrast enhancement was solved with automatic fitting of Gaussian distributions to image histogram and re-scaling of the values corresponding to the face area to the full image range, the same as described in Section 4.2.1. To make the model more robust and suited for real-life applications, we created additional 5 negative categories and collected samples for them using the same Lepton camera module (computer mouse – 2855 images, projector - 2968, keyboard - 3086, back of a head - 3083, hand – 3083; Fig. 4.14).

A testing set was created by randomly selecting 159 images from the face category. Original images were kept as a baseline set for models evaluation. Then, distortions and image modifications were introduced to all selected images in a post-processing phase to simulate various possible scenarios of remote medical diagnostics (e.g. lying, tilting a head, etc) and verify robustness of models to displacement of facial regions. Specifically, created categories include: (159 images in each category) random displacement of facial features, random displacement of image quaters, rotation 90^0, rotation 180^0, rotation 45^0, as presented in Fig. 4.15. Selected and modified images

were saved as a separate set and used only for a testing purpose, models were trained on original (not distorted) images (remaining images after selection of 159 test samples).

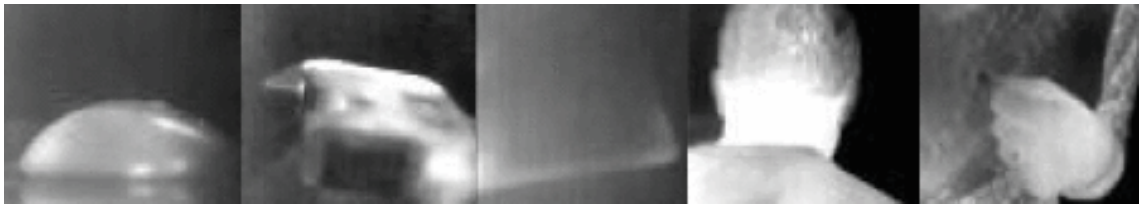


Figure 4.14. Examples of images collected for negative categories; from the left: computer mouse, projector, keyboard, back of a head, hand

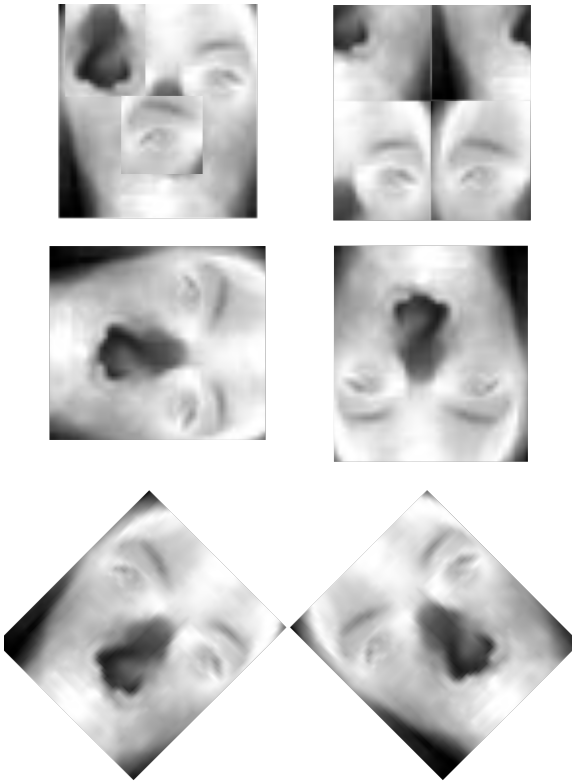


Figure 4.15. Distortions and modifications applied to the Lepton-IE set to simulate possible scenarios of remote medical diagnostics and compare robustness of Capsule and Convolutional Neural Networks in those use cases; from the left top row: random displacement of facial features, random displacement of image quaters, rotation 90^0, rotation 180^0, rotation 45^0

The main difference between CNNs and the capsule network is that the latter divides each model layer into neurons grouped together into so-called capsules. Due to the use of the group of neurons, input and output of each layer component is represented as a vector instead of single scalar values, as in CNNs. Then, the Iterative Routing by Agreement (IRA) mechanism is used to find the route between capsules which would lead to the best prediction. In previous studies, IRA has been proved to be more effective than max pooling used in CNNs, as it looks for all correlations, not only the most active feature within the given window. IRA is based on an iterative approach, which at first passes the output O of a capsule i in a layer n (O_i^n) to inputs I of all m capsules

in the following layer ($I^{n+1}_{1...m}$). For all m capsules in the layer $n+1$, total input to each capsule ($I^{n+1}_{1...m}$) is calculated as a sum of prediction vectors $u_{1...k}$ from all k capsules in the preceding layer weighted by the coupling coefficients c. Prediction vectors ($u_{1...k}$) are calculated as the preceding layer's capsule's output (O^n_i) multiplied by a weight matrix (W). For the capsule i in the layer n and the capsule j in the layer $n+1$, the total input to the capsule j is defined as:

$$I^{n+1}_j = \sum_{i=0}^{i=k} O^n_i W_{i,j} c_{i,j} \tag{4.5}$$

Coupling coefficients are determined by IRA. The prediction vector which produces the largest scalar product with the output of the next layer (O^{n+1}_i) is selected (u_z) and the coupling coefficient of the corresponding capsule (O^n_z) is increased in the top-down adjustment process in order to indicate its higher relevance in calculating the output predictions. The probability of the object represented by capsule j being present in the input is then given by a length of output vector of this capsule (j):

$$O^{n+1}_j = \frac{\left\|I^{n+1}_j\right\|^2}{1+\left\|I^{n+1}_j\right\|^2} \frac{I^{n+1}_j}{\left\|I^{n+1}_j\right\|} \tag{4.6}$$

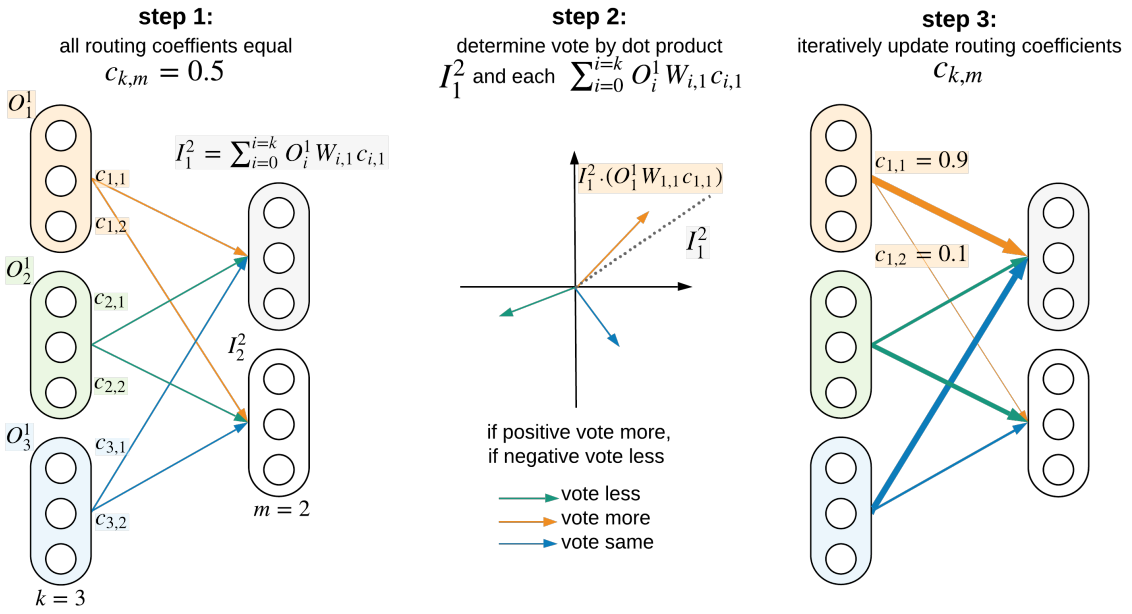


Figure 4.16. Iterative update of routing coefficients applied in Capsule Networks

The flow of the IRA approach is illustrated in Fig. 4.16. Step 2 shows the process only for the first capsule in the second layer. For training of the capsule network we used default regularization and hyper-parameters, as specified in [176]. To compare different variants of IRA, we tested 1, 3, 4 and 7 routing iterations in each case during 50 epochs of training.

Results and Discussion

Table 4.7 presents comparison of accuracy achieved on the test subset of the Lepton-IE set by Inception and Capsule Network with different number of routing iterations. Solutions considered by us should be convenient and ensure constant monitoring of subjects. Thus, it is very important to ensure that face areas are never missed even at the cost of some false predictions. In order to provide better insight into performance of both networks and take number of false negatives into account, we also calculated recall of examined models. Produced results together with confidence score for face class are collected in Table 4.8. Examples of body positions in possible remote medical scenarios are presented in Fig. 4.17, which includes visible light images and transformed thermal data from the Lepton-IE dataset to visualize each of considered remote medical diagnostic scenario. Accuracy achieved by capsule and convolutional-based networks in each use case is presented in the Fig. 4.17 caption.

As presented in Table 4.7, the accuracy of convolutional and capsule-based networks (regardless of a number of routing iterations) are on pair if original data with faces oriented in one direction (vertically) are used. However, the rotation invariance of Capsule Network can be observed by analysis of recall results calculated for modified images. Last three rows of Table 4.8 present potential remote medical diagnostic scenarios for which accuracy of face classification should not be impacted. Those cases include rotations by various angles which may simulate different body poses during data acquisition, e.g. lying down (see Fig. 4.17). We can observe that Inception, which uses convolution operations, is very sensitive to such modifications, leading to a significant decrease of accuracy, e.g. for rotation of 45^0 the recall value is below 10%, while the corresponding results for capsules are above 90%, and in the best case are close to 100%. This proves significant advantage of using Capsule Network in applications where objects can have different orientations.

Similar analysis performed for modifications based on displacement of facial feature/image parts showed that performance of both networks is not sufficient. Those modifications led to distortion of faces appearance, so theoretically fewer faces should be detected, causing the recall values to drop. This result is true for random displacement of image quaters for both Inception and Capsule Networks. Yet, random displacement of facial features hasn't resulted in a decrease of accuracy, still predicting most of samples as faces. We believe that this may be again caused by low resolution of acquired thermal images and high similarity of facial features. Since cost of higher quality thermal sensors is still high, in the next chapter we will describe details of our novel Deep Neural Networks proposed for generating super-resolved thermal data, so that lower resolution cameras can lead to similar classification accuracy as in case where better sensors are used.

Table 4.7. Comparison of Inception and Capsule networks accuracy on the test subset of the Lepton-IE database [%]

Inception	**Capsule Network**			
	routing iterations			
	1	3	4	7
98.91	99.92	99.85	99.88	99.66

(a) Rotation 180^0 - accuracy for proposed capsules 100%, CNN 9%

(b) Rotation 90^0 - accuracy for proposed capsules 78%, CNN 20%

(c) Rotation 45^0 - accuracy for proposed capsules 99%, CNN 84%

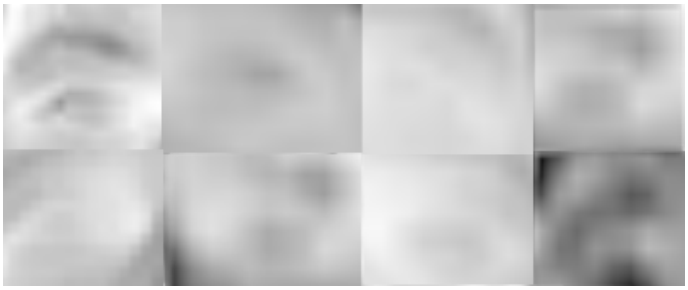
Figure 4.17. Examples of potential body positions in remote medical diagnostics solutions

Table 4.8. Comparison of Inception and Capsule networks on the test subset of the Lepton-IE database [%]; top line: recall for a face class; bottom line: average confidence value of True Positive samples of a face class [avg. % ± stdev]

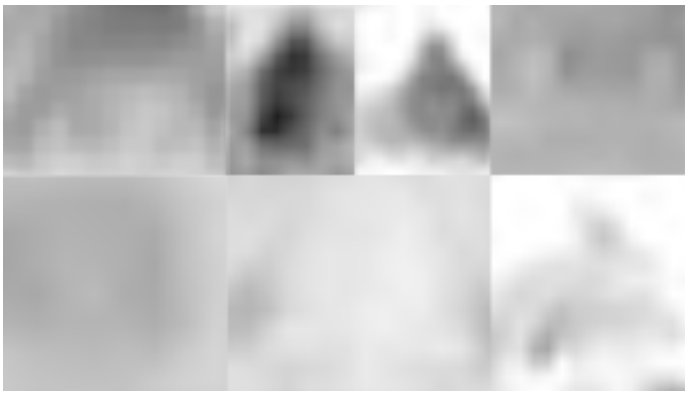
	Inception	**Capsule Network**			
		routing iterations			
		1	3	4	7
baseline	94.68	100	100	100	100
	96.04 ± 7.79	91.20 ± 1.35	79.64 ± 3.36	65.47 ± 4.73	58.93 ± 5.60
random displacement	87.34	99.28	100	99.36	100
of facial features	92.87 ± 7.66	89.94 ± 2.63	79.4 ± 3.36	62.00 ± 5.87	54.44 ± 6.80
random displacement	48.10	54.43	99.36	48.30	56.32
of image quaters	77.66 ± 19.20	79.2 ± 8.23	74.5 ± 5.54	45.43 ± 11.14	39.32 ± 10.58
rotation 90^0	20.25	78.48	67.72	70.88	82.91
	69.37 ± 14.78	76.64 ± 7.89	57.27 ± 9.87	42.92 ± 12.08	38.55 ± 9.66
rotation 180^0	83.54	99.36	96.83	96.83	98.73
	85.79 ± 14.32	86.80 ± 4.10	67.92 ± 7.90	54.64 ± 9.29	47.97 ± 9.86
rotation 45^0	9.49	100	99.36	91.14	99.36
	63.25 ± 9.41	83.42 ± 3.35	56.67 ± 6.82	43.08 ± 7.44	45.71 ± 6.76

4.4 Problems

Results of experiments performed for deep classification and detection models on thermal images proved that even with a limited amount of data, high prediction accuracy can be achieved by transferring knowledge from other image domains. Thermal datasets collected by us contain much fewer samples than publicly available visible lights sets, for which DNNs achieve human-like performance. However, utilization of already adjusted weights and retraining of only the final classification component allowed for preserving high accuracy (above 90%) for face and facial areas classification tasks.



(a) Eyes areas



(b) Nose areas

Figure 4.18. Facial features extracted from low resolution thermal images; one can note lack of high frequency features and blurriness which makes it hard to distinguish facial regions, e.g. in eye images only eyebrows are distinguishable

In addition, we showed how to provide automatic, rotation invariant solution for face classification in possible scenarios of remote medical diagnostics by utilization of novel DL architecture based on capsules. Due to the use of a single end-to-end neural network-based pipeline, the need for additional pose compensation techniques with e.g. facial landmarks [139] was eliminated making the solution more suitable for resource-constraint devices. Yet, although the proposed capsule-based solution outperformed CNN on images at various rotations, it has been observed that image distortions have more significant effect on prediction accuracy. In the case of completely deformed faces constructed by facial features displacement, both CNN and capsule model still recognized them as facial areas, what is incorrect. We believe that a likely cause of this result lies in the low accuracy of acquired sequences (80x60 pixels). Worse image resolution might have led to wrong generalization due to the blurring of features. For many samples from the Lepton-IE set, it is difficult to distinguish specific face areas even for a human (see Fig. 4.18). Thus, it may turn out that the model performs better if images of higher resolution are provided.

On the other hand, the cost and physical size of higher resolution thermal sensors is still quite high comparing to visible light cameras. Considering target platforms and applications that we would like to address in this work, the goal is to deliver telemedicine solutions for low resolution thermal images by analysing them and improving their quality using Artificial Intelligence and at the same time avoiding the use of more sophisticated thermal devices. Taking it into account, we propose to apply Super Resolution (SR) algorithms to acquired sequences in order to determine if DL-based thermal image enhancement can lead to more exact detection of facial areas and, as a result, higher accuracy of contactless vital signs extraction.

4.5 Summary

In this chapter we introduced novel methods and techniques which utilize DL models in order to perform face and facial areas detection from thermal images. Performed evaluation proved that even with the limited amount of data it's possible to achieve high prediction accuracy by utilizing the knowledge from visible light spectrum using transfer learning technique. Moreover, an in-depth analysis of motion influence on face area classification accuracy was performed taking into account various possible scenarios of contactless vital signs estimation. Sequences with higher motion content led to worse accuracy, missing some true positives. Yet, we proved that motion can be reduced by focusing person's attention on some tasks and thereby improving system precision.

Additionally, we proposed an innovative neural network architecture aimed at detecting facial regions without the need of annotating data with bounding boxes, what led to a significant decrease of the inference time and improved the ease of model training. Also, it turned out that the introduced detection pipeline outperforms existing object detection model due to being less sensitive to false negatives and class imbalance problem, achieving IoU above 0.5 for all evaluated facial areas in thermal images of a small spatial size (80x60). Conducted experiments and achieved results support the first thesis formulated in the presented doctoral dissertation which stated that architecture of Deep Neural Network designed for classification of visible light images can be modified in such a way that distribution of extracted features will be recreated enabling detection of facial areas from low resolution thermal data. Since some previous studies on computer vision-based remote medical diagnostics indicated the need for automatic localization of body parts and regions (i.e. pulse estimation from a forehead [44]) , we believe that the proposed method can be very useful for those applications by automating their pipelines. In addition, some standard medical procedures can be speed up, improving quality of healthcare, e.g. the proposed detection model can be used for person and object identification from graphical markers [153]. Due to the small physical size of the used thermal camera sensor, it can be embedded in existing smart home/building infrastructure and used for collection of human-related data in a non-disruptive way.

Finally, we presented details of capsule network and evaluated its performance on thermal sequences to prove its robustness in different image domains, not only visible light data, as originally assumed. Limitations of presented methods (such as influence of low image resolution on classification results) were identified. Although this chapter introduced a simple automatic image pre-processing step, which led to improvement of facial feature representation and thus easier network training (due to the presence of more distinct features), we believe that other techniques, e.g. Super Resolution could be very beneficial for further accuracy increase. Studies on thermal image enhancement and its applications are presented in details in following chapters (Chapter 5 and 6).

Chapter 5

Proposed Model for Thermal Images Resolution Enhancement

5.1 Introduction and Overview

Latest technological advances have led to increased availability of affordable, higher resolution thermal cameras. FLIR®, the world leader in thermal imaging infrared cameras, have recently introduced Lepton family sensors characterized by a small physical size (e.g. 10.5 × 11.7 × 6.4 mm) and much lower price comparing to other thermal acquisition devices (i.e. 200 USD). As a result, enablement of various thermal imaging monitoring solutions, including remote medical diagnostics, becomes more feasible and various studies in this area have already been proposed, e.g. contactless monitoring of respiratory activity [148]. In spite of a more rapid progress in thermal cameras development, our studies showed that image resolution is still not satisfactory and may affect detection accuracy (see Section 4.3). One possible cause of this result is the blurring of features present in thermal images and as a result their high similarity. Low resolution of facial features could be a problem for determining Region of Interest (RoI) which represents respiration-related color pixel changes due to the use of kernels learnt to extract high frequency components [141]. Thus, resolution improvement using image processing techniques is of a high interest to us, as it may have a positive effect on areas detection accuracy, while using the same imaging device.

The idea of enhancing image resolution, known as Super Resolution (SR), has been widely studied and various techniques have been already proposed. Due to the progress in Deep Neural Network (DNN) development, different architectures have been created, improving state-of-the-art methods. Yet, most of proposed solutions are still focused mainly on a visible light spectrum and direct application of those models to thermal data may not be satisfactory. Thus, in this chapter we provide in-depth analysis of existing SR solutions and their applicability to thermal domain. Then, we introduce a novel DNN architecture, designed by us with thermal features characteristic in mind. To the best of our knowledge, this is the first attempt to design model specifically dedicated to thermal image enhancement by addressing more distant dependencies between interesting components caused by the heat flow in objects. The proposed model is verified and compared with other SR models on a wide set of thermal datasets. In addition, we also verify possibility of transferring knowledge by utilizing SR models pre-trained on visible light data. Conducted analysis include e.g. determining the influence of temporal frames averaging and various bit depths on Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM).

5.2 Super Resolution

5.2.1 Objective

The aim of the Super Resolution (SR) method is to restore High Resolution (HR) data from corresponding Low Resolution (LR) inputs. Specifically, if an output is restored from a single image, the approach is called Single Image Super Resolution (SISR). In general, SISR is very challenging, because various outputs can be produced from a single input The general formulation of Super Resolution task is a solution to the ill-posed inverse problem, defined as:

$$X = (Y \circledast K) \downarrow_s + n \tag{5.1}$$

$$\widehat{Y} = SISR(X) \tag{5.2}$$

where K is the degradation operator, $\downarrow_s$ is down-scaling operation with a given scale s and n is additional noise. Y is the original HR input and $\widehat{Y}$ is the restored HR data we want to infer from LR input X by applying SISR model in order to achieve (in the best case) $\widehat{Y} = Y$. To alleviate this ill-posed problem of SISR, a prior knowledge is usually utilized to constrain the solution space. The prior can be obtained by e.g. a) interpolating pixels values [177, 178, 179]; b) exploiting internal structure of an image based on self-similarity of recurring image sub-parts [180], or 3) analysing correspondence between a pair of images with two different resolutions, i.e. example-based algorithms [181].

One of the first method, which dates back several decades, was based on bicubic scaling [177]. Other techniques, where the prior knowledge is learnt using interpolation algorithms were proposed by Zhang and Wu [178] using edge-guided information or by Romano et al. [179], who utilized local sparsity-based modeling. Yet, this group of Super Resolution methods, known as interpolation-based SISR utilize generic smoothness assumptions and do not discriminate between edges and object parts. As a result, restored images may become blurred as all components are treated similarly, what is a serious constraint of interpolation-based techniques.

Thus, more and more studies are being directed towards example-based solutions which satisfy the prior knowledge by preserving consistency between LR and HR data pairs. The intuition behind the success of example-based approach is that image pixels have less variability than a set of random variables. During the training phase, we teach our algorithm how to restore image details using corresponding image regions seen at lower resolution. Even though features are more blurred and distorted in LR inputs, it's easier for the model to learn those relations than if random data is used.

Since the access to a perfectly synchronized acquired in the same conditions pair of LR and HR images is usually limited, in a standard approach LR input (X) is generated synthetically from collected HR image (Y) using downscaling operation ($\downarrow$) with a scale s, e.g. using bicubic interpolation:

$$X = Y \downarrow_s \tag{5.3}$$

The LR image can either have a spatial size smaller than the HR sample (in this case its resolution is *truly* lower) or it can be up-scaled again ($\uparrow$) after downscaling (scaling twice with inverse scales, e.g. 1/2 and 2):

$$X = Y \downarrow_s \uparrow_s = Y \uparrow_{\frac{1}{s}} \uparrow_s \tag{5.4}$$

In the latter approach, the generated LR image has the same spatial dimension as the original HR pattern but its quality is degraded, as shown in Fig. 5.1. Then, the goal is to restore degraded image

features and components, so that the output of SR is as similar as possible to the original HR data. For image with a *truly* lower resolution, apart from feature restoration the applied algorithm has to use an up-scaling operation in order to generate a higher scale image version. In this case, the standard Convolutional Neural Network (CNN) pyramid architecture is not applicable, as pooling and strides are common downsampling operations that would lead to decrease of a size instead of increasing it. For such tasks a transposed convolution is usually utilized. Transposed convolution is also known as deconvolution, what in fact is an incorrect name, as deconvolution is a signal filtering operation to compensate for calculated convolution and transposed convolution is in fact a convolution but with a transposed kernel matrix. However, the information lost after pooling or stride may still not be completely recovered. Due to higher complexity of solutions based on deconvolution, in this work we focus on enhancement of upscaled LR images, so that the size of the image is already bigger and the only task is the restoration of features (pooling operation is completely skipped and stride is set to 1).

Specifically, the solution S (e.g. neural network) characterized by parameters θ is applied to the generated LR input X, restoring the HR image $\widehat{Y}$. Parameters θ are adjusted in such a way that the restored output is as close to the original HR data Y as possible. Since this is a regression problem, the most frequently used estimator is the Mean Squared Error (MSE) metric, which calculates distances (errors) between restored and original pixel values and squares them producing the average over a set of all pixel errors. The goal is to find parameters θ^*, for which MSE reaches its minimum:

$$\begin{aligned} \theta^* &= argmin_\theta (Y - \widehat{Y})^2 \\ & argmin_\theta (Y - S_\theta(X))^2 \end{aligned} \tag{5.5}$$

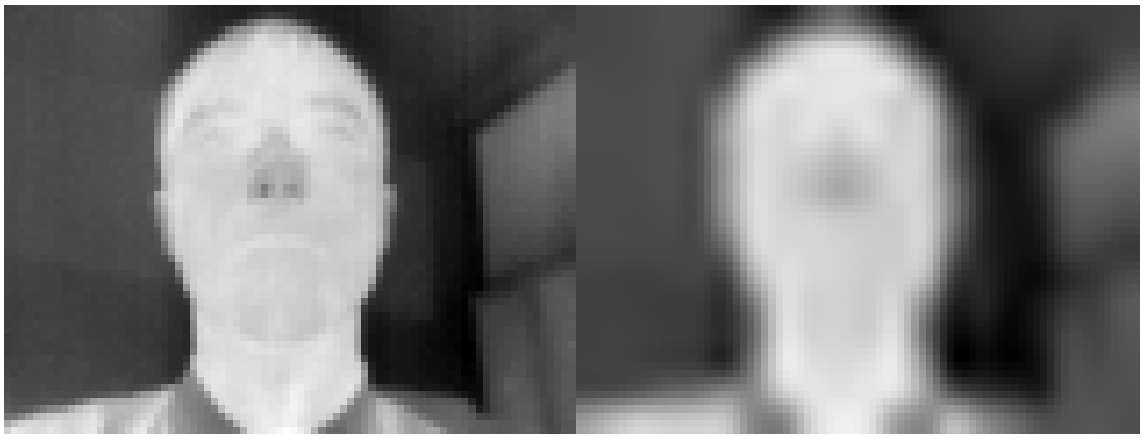


Figure 5.1. Example of thermal image of a face (on the left) and its corresponding LR version generated by downscaling and then upscaling of an original image by a factor of 4

5.2.2 Evaluation of Resolution Enhancement Methods

As stated in the second thesis formulated in the presented dissertation, the aim of one of the studied problems is to propose a novel thermal image enhancement method which allows for increasing quality of collected sequences and thereby accuracy of remote medical diagnostic solutions. In order to elaborate on the proposed solution, it's important to evaluate it and compare

results with other existing Super Resolution approaches. Two image quality metrics are usually used for this purpose in image enhancement tasks.

The first metric, PSNR, is a ration between an original data and image that we want to evaluate, e.g. a compressed frame, restored output, etc. The result of PSNR is specified in decibels. The higher the value of PSNR, the better the quality of evaluated data. PSNR is calculated based on the Mean Square Error (MSE), which is an inverse measure of enhancement quality, i.e. the lower the error, the better the quality. MSE indicates the cumulative squared error between two images (I_1 and I_2) of a size rows x columns, as:

$$MSE = \frac{\sum_{r,c \in rows,columns} (I_1(r,c) - I_2(r,c))^2}{rows * columns} \tag{5.6}$$

PSNR represents a peak of MSE error:

$$PSNR = 10log(\frac{R^2}{MSE}) \tag{5.7}$$

where R corresponds to maximum fluctuation of input data, e.g. for 8-bit images R equals 255. As can be deduced, PSNR specifies correlation between maximum power of a signal and power of a noise leading to degradation and decrease of representation quality.

The second metric, SSIM, is used for evaluation of perceived quality of digital images and videos. Equivalently to PSNR, SSIM allows for evaluating similarity between two inputs. Specifically, a restored (previously compressed or processed in other way) input is compared with its corresponding original version. Since SSIM uses perception information, it allows for obtaining additional details to absolute errors produced by PSNR. Apart from evaluating decrease of quality as observed changes of structure, SSIM also takes into account luminance and contrast components. Thus, combination of PSNR and SSIM allows for getting complete information about performance of image enhancement models. SSIM for 2 inputs I_1 and I_2 is defined as:

$$SSIM(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2} + c_1)(2\sigma_{I_1,I_2} + c_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + c_1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + c_2)} \tag{5.8}$$

where c_1 and c_2 are constants used for stabilization, defined as $c = (kR)^2$, k equals 0.01 for c_1 and 0.03 for c_2. μ is the average and σ is the variance for a given input. σ_{I_1,I_2} is the covariance of two inputs.

5.2.3 Existing neural network-based Super Resolution Methods

The Super Resolution problem is a classical computer vision task and has already been studied in-depth for various applications, including medical solutions e.g. computer-aided diagnosis (CAD) for determining breast tumor [182] or reconstruction of computed tomography (CT) data [183]. As already mentioned, example-based method deal better with some limitations of interpolation techniques, e.g. a tendency to generate overly smooth edges, especially for higher scale values. Thus, approaches which utilize relations between a pair of images to train Super Resolution machine learning algorithms have become more popular and they are also the main focus of our study. This section overviews examples of such learning-based solutions. Specifically, we present state of the art neural networks for which a pair of corresponding LR and HR images is used to optimize network parameters θ such that the error between restored and the original image will be minimal, as defined by Eq. 5.5

Visible Light Data Processing

One of the first studies on utilization of neural networks for image enhancement focused on image denoising, which solves a slightly different problem than Super Resolution as it tries to recover a clean image from a noisy input X. For denoising, the ill-posed inverse problem of Super Resolution, defined by Eq. 5.2, can be re-formulated to a simplified formula:

$$X = Y + n \tag{5.9}$$

since the degradation operation K is skipped. Even before the reinvention of Deep Learning (DL) in 2012 with AlexNet [109], CNNs have been used for image denoising [184]. Yet, as pointed out by the authors, achieving better performance with increased number of layers may become computationally intensive and back then not all efficient DL regularization techniques were known. Image restoration with CNNs has also been described by Eigen D. et al. [185], where a neural network with two hidden layers was used. Comparison with previous non-convolutional techniques showed increase of PSNR by $\sim 0.9dB$. Other, more classical machine learning algorithms have been also adapted to remove noise from image data. Burger H. et al. [186] proposed to learn LR-HR mapping with multi layer perceptron (MLP), showing that the large training set (>150k images) is essential for achieving a decent performance. MLP has been also used on wavelet coefficients instead of image data in the study presented by Zhang S. and Salari E. [187]. The motivation for using wavelet coefficients lies in the fact that they can be treated as strong image priors what may lead to better accuracy.

Recent techniques take advantage of more sophisticated neural network architectures. The overview of various DNN-based SR techniques has been provided in [188]. Here, we describe in details some key examples of such solutions. Cui Z. et al. [189] proposed a method based on non-local self-similarity search (NLSS) process and collaborative local auto-encoders (CLA) stacked on top of each other in order to perform a graduate upscaling of LR inputs. Due to the use of multiple stacked CLAs, the model is referred to as Deep Network Cascade (DNC). The main drawback of the proposed auto-encoder-based pipeline is the requirement to perform a separate optimization of NLSS and CLA.

To overcome the disadvantage of the DNC pipeline, Dong et al. [190] proposed to obtain SR output through a single end-to-end mapping that can be learnt by a single neural network. The introduced convolutional-based architecture, known as Super Resolution Convolutional Neural Network (SRCNN) turned to be a breakthrough in image enhancement techniques. SRCNN was shown to form all operations usually performed in neural network-based Super Resolution task with convolutions. The only pre-processing performed before CNN is image interpolation to the size of the HR data (as specified by Eq. 5.4). The prepared input is then fed into 3 convolutional layers performing feature extraction (FE), non-linear mapping (NLM) and reconstruction (R), respectively. At the feature extraction step, each LR patch is passed through convolutional filters to represent them in a form of feature maps. Then, in the non-linear mapping phase each extracted vector is mapped to another representation which conceptually is equivalent to a high-resolution (HR) patch. Finally, the output is reconstructed by aggregating all produced HR patches. Although task solved by each of those layer is different, it has been proved that representing all of them as a convolutional layer is sufficient for accurate image restoration. Since then, the idea of using CNNs for SR was constantly developed and led to design of various state-of-the-art solutions. Additional experiments performed in SRCNN study, showed that the increased number of layers can lead to a further accuracy improvement, what became a base for next CNN-based SR models. At first, Kim

et al. [191] made use of the VGG-net classification model [110] to solve SISR, introducing Very Deep Super Resolution (VDSR) network. In addition, the residual skip connection was applied from the input to the output in order to correlate LR and HR data. The use of adjustable gradient clipping, the strategy that helps with exploding gradient problem in DNNs, allowed for increasing learning rates and at the same time convergence speed.

Further modifications proposed by authors of [191] included incorporation of recursive supervision to increase the network depth, without introducing new parameters and reduce the problem of vanishing gradients [192]. The vanishing gradient problem is often present in a neural network training process, especially for deeper architecture, where the gradient may be vanishingly small for subsequent layers in the backpropagation step, leading to none or insignificant updates of weights. The SR model presented in [192], called Deeply Recursive Convolutional Network (DRCN) led to the improvement of PSNR by $\sim 1dB$ comparing to SRCNN. Recursive supervision is the approach of applying the same convolutional operation multiple times in the non-linear mapping subnetwork (name convention as defined by SRCNN), producing D predictions, all supervised during training. The final output is produced as a weighted sum of outputs from all those D recursions. Additional contribution to the recursive structure, made by Tai et al. [193], was based on the huge success of residual network (ResNet [113]) in the image classification task. In ResNet, the use of residual blocks with skip connections at all levels of the feature extraction step allowed for a significant increase of the network depth while eliminating the vanishing gradient problem. Thus, more complex representations were extracted, significantly increasing prediction accuracy. Tai et al. inspired by those findings, proposed to apply a structure similar to ResNet to their Super Resolution CNN, called Deep Recursive Residual Network (DRRN) [193]. Similarly to DRCN, DRRN proposed to utilize recursions, but adopted additional residual connections with shared weights both in a global and local manner. However, experiments performed by authors showed that the best performing architecture didn't use recursion at all and the final model consisted of residual blocks only.

A separate group of image enhancement techniques is based on generative models. Generative Adversarial Network (GAN) [194] have recently gained a lot of popularity in various computer vision tasks, also in medicine, e.g. for synthesis of missing PET and MRI data [195]. The idea of this architecture is to use two separate neural networks. The first one, generator, has to create an image as similar as possible to the real data. Then, the goal of a second network, discriminator, is to determine whether the input image is real or generated. It turned out that this approach can also be successfully applied to SISR task, synthesizing crucial image details that might be lost because of downscaling, especially when big scale factors are used.

The SRGAN network proposed by Ledig C. et al. [196] allowed for successful restoration of detailed components of the HR image, setting a new state-of-the-art performance even for bigger scaling factors (4x). Other examples of GANs used for image Super Resolution include EnhanceNet [197] aimed at automated texture synthesis instead of pixel values reproduction and SFTGan [198] - the GAN network with Spatial Feature Transform modules based on affine transformation coefficients. Hinton's Deep Belief Network (DBN), generative model designed to learn the entire input representation at each layer, has been also successfully used for SISR [199], showing an inference time speed-up of 8 times for 2D data and of 200 times for 3D data, opening new possibilities for various medical applications that usually operate on 3D volumes. Other examples of generative models include autoregressive networks which produce a value for a current pixel using its preceding left and top neighbour (so-called Pixel-CNNs). This approach has been utilized for SISR in the network proposed by Dahl et al. [200].

Thermal Data Processing

The main drawback of utilizing described models in our research on thermal image processing is the fact that they were designed and tested only on visible light images. Although conducted experiments prove their effectiveness and high accuracy of image restoration, direct application of those solution to the thermal domain may not be sufficient due to a different characteristic of thermal data. At the same time, only a limited number of studies focus on designing DL models for the thermal domain. Most of existing methods for thermal data enhancement utilize machine learning approaches, requiring the manual selection of a prior knowledge, what frequently is difficult to achieve, as described in Chapter 2.3. The approach proposed in [201] utilized the Huber cost function for defining the difference between HR and LR frames in combination with the bilateral Total Variation as the prior. Other study, introduced by Kiran Y. et al. [202], utilized a simple regression technique, where Matrix Value Regression Operator was used to learn mapping between LR and HR patches. The work proposed by [203] made use of the rigid transformation matrix to determine pixel shifts and rotations in order to produce geometric relation of LR images to the reference data. After that, LR images were projected on HR grid and used for interpolation of unknown data. An interesting approach to thermal SR was described by Almasri F. and Debeir O. [204], who proposed to solve the problem of thermal resolution enhancement using fusion of thermal and visual modalities. The method was compared with SRCNN and VDSR, showing the improvement in PSNR metric on a dataset containing visible light data and their thermal counterparts. Another solution which utilize visible light data as a guidance for thermal image restoration was proposed by Chen X. et al. [205]. Both of those methods though, were based on a strong correlation between visible light and thermal data, what requires the use of two synchronized sensors and acquisition of two video streams simultaneously. Applications analysed by us require a special attention to data privacy aspects and acquisition of visible light data is often undesirable in such cases, e.g. in remote healthcare. Thus, the goal is to design a model which utilizes only thermal sequences, what makes the fusion-based network inapplicable for our target solutions. To address the problem of the lack of high frequency features in data reconstructed from LR thermal inputs, Zhang X. et al. [206] combined a compressive sensing technique aimed at image resolution enhancement with DNN applied on the produced output to reduce the noise. The results showed the advantage of this approach over SRCNN, yet the method hasn't been compared with more recent Super Resolution architectures.

Another novel work on super-resolving thermal images was conducted by Kuang X. et al. the same year as our study. The authors proposed a SISR technique based on combined CNN and GAN models designed for thermal image enhancement [207]. The introduced pipeline was tested on visible light dataset and thermal images collected from internet. We encounter two main disadvantages associated with the presented study. First of all, no details about thermal images, e.g. bit-resolution, data format, data acquisition process were provided, what makes it hard to evaluate the proposed model. Moreover, although authors stated that existing residual and encoder-decoder-based Super Resolution networks are not efficient on infrared images, they did not provide any quantitative comparison with state-of-the-art models. The lack of this comparison makes it difficult to determine whether the proposed CNN-GAN solution is more efficient in thermal domain, contrary to our work [208] which provides the extensive benchmark evaluation of our model and a set of existing SR networks, performed on publicly available thermal dataset (IRIS) and on thermal images acquired by us. The neural network architecture introduced by us is presented in the following section.

5.3 Proposed Network for Thermal Data Enhancement

Our work is one of the first attempts to the improvement of thermal image resolution with Super Resolution Deep Neural Network specifically designed for infrared data without the use of visible light features. The model introduced by us consists of a wider receptive field in the feature extraction subnetwork. As a result, more distant relations between adjacent facial regions are taken into account while building feature vectors representing HR image patches. This design is essential for addressing the blurring effect of thermal images.

5.3.1 Problem Formulation

To better understand the novelty of the introduced thermal data-oriented neural network architecture, let's start with a formulation of the basic CNN-based SR network. neural network S, characterized by parameters θ, produces a restored HR image $\widehat{Y}$ from a corresponding LR input X (generated from the original HR data Y, as specified in Eq. 5.4) by performing a following sequence of operations:

$$\widehat{Y} = F_r(F_{nlm}(F_{fe}(X))) \tag{5.10}$$

where $F_{FE/NLM/R}$ are subnetworks which realize feature extraction, non-linear mapping and reconstruction tasks, respectively. As showed in the pioneer work in this area [190], all those subnetworks can be represented as convolutional layers with trainable weights $W_{FE/NLM/R}$ and biases $B_{FE/NLM/R}$ matrices:

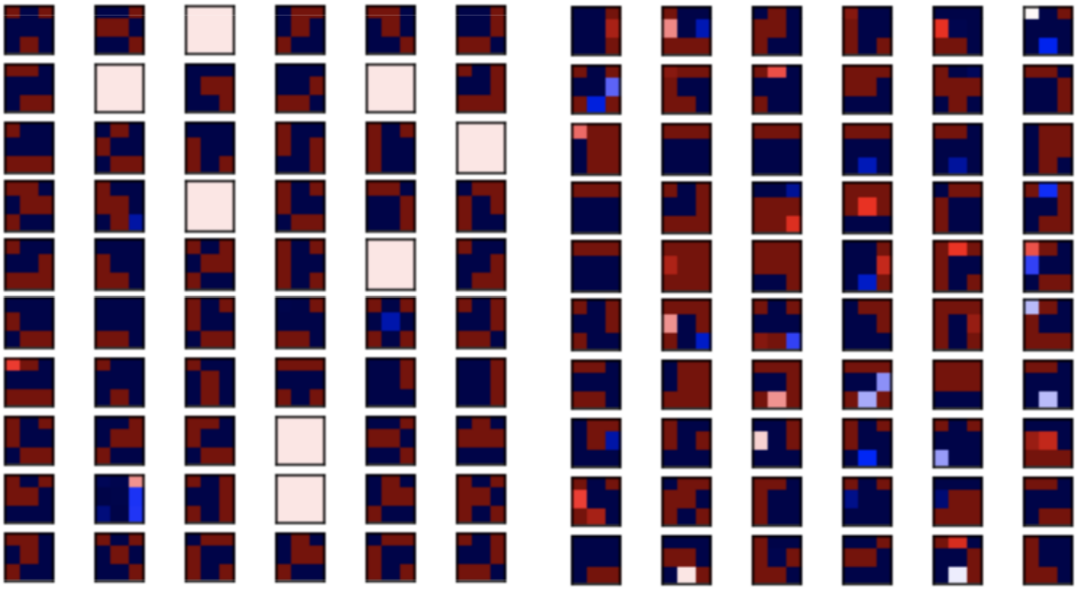
$$\widehat{Y} = W_r \circledast (\sigma(W_{nlm} \circledast (\sigma(W_{fe} \circledast X + B_{fe})) + B_{nlm})) + B_r \tag{5.11}$$

where σ is the activation function applied after each convolution to introduce nonlinearities and the symbol $\circledast$ denotes the convolution operation. Eq. 5.11 presents a general idea of CNN-based Super Resolution model with one convolutional layer in each subnetwork. Thus, $W_{FE/NLM/R}$ correspond to $k_{FE/NLM/R}$ kernels, each of a size $width_{FE/NLM/R}$ x $height_{FE/NLM/R}$ x $channels_{FE/NLM/R}$, producing $k_{FE/NLM/R}$ output feature maps after each step. Weights and biases form together parameters θ ($\theta = \{W_r, W_{nlm}, W_{fe}, B_{fe}, B_{nlm}, B_r\}$) adjusted during network optimization in such a way that the reconstructed output $\widehat{Y} = S_\theta(X)$ is as similar to the original HR data Y as possible. Our study utilizes the supervised learning, so original HR images $Y_{1...N}$ (where N is the number of samples) are used as labels for corresponding reconstructed samples $\widehat{Y}_{1...N}$ and the distance between them is defined by the cost function represented as Mean Squared Error (Eq. 5.5) averaged across all samples:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (((Y_i - \widehat{Y}_i)^2) = \frac{1}{N} \sum_{i=1}^{N} ((Y_i - S_\theta(X_i))^2) \tag{5.12}$$

The goal is to minimize the cost function, i.e. favor higher PSNR, which is directly related to MSE, as specified in Eq. 5.7.

According to previous studies, better performance is achieved with deeper architectures [192, 193], thus it's beneficial to extend the basic form of CNN-based SISR described by Eq. 5.11 by introducing more layers. On the other hand, increased number of parameters may have a negative influence on the ease of network training and solutions may suffer from the vanishing gradient problem. Our study addresses both problems by providing a deeper SISR model with weights shared across both residual and recursive blocks. In addition, the proposed network is specially



(a) Filters of a single convolutional block applied to an input image

(b) Filters after 3 residuals (each consisting of 2 convolutions) with weights shared among residuals

Figure 5.2. Examples of filters learnt by a model aimed at solving Super Resolution task

designed with thermal data characteristic in mind, proving its advantages over other state-of-the-art models introduced for visible light images. The introduced DNN architecture is explained in details in the following section.

5.3.2 Proposed Network Architecture

Since our study focuses on evaluation of face hallucination solutions in thermal domain, the main motivation for designing a novel DL model is to address different characteristics of visible light and thermal images and reflect this discrepancy in the neural network architecture itself. Main difference between images acquired for various ranges of electromagnetic radiation lies in the information visible in this range. Heat flow in objects leads to the equalization of temperature values at the boundaries of adjacent regions what is reflected by a low contrast between them in the constructed thermal image, as shown in Sec. 4.2.1 (Fig. 4.3). As a result, a relatively small receptive field, e.g. 3x3 after one convolutional layer in the feature extraction subnetwork of SRCNN may not capture important distant dependencies between facial features within specific image regions. This assumption can be proved by visualizing features extracted from thermal data using a single convolution vs. a sequence of 6 convolutions. Fig. 5.2 presents examples of kernel weights learnt after a various number of applied convolutional layers. It can be easily observed that the more filters is used, the more complex features are extracted, what we believe is crucial for mitigating the problem of more distant dependencies between facial regions in thermal imaging.

Taking it into account, the core idea of the introduced SR CNN is based on the widening of the receptive field. However, as proved by previous studies [113] simple stacking of more layers is not efficient and the network may suffer from the vanishing gradient problem. Therefore, our model utilizes residual blocks, similarly to other SR CNN-based solutions [193, 209]. Yet, contrary to them, we propose to apply residual blocks both in supervised recursions in the non-linear mapping subnetwork as well as at the feature extraction step. In this way, we revisit the problem of combining

recursions and residuals in a single SR CNN. We believe that previously this combination turned out to be unsuccessful [193] due to the incorrect placement of those blocks. Our work [208] provides in-depth evaluation of SISR architecture variants that differ in both location and number of each block used for thermal image enhancement.

Residual block used in our model consists of two convolution operations, followed by batch normalization and activation function σ. In DL solutions, a commonly used activation function is ReLU (Rectified Linear Unit), defined as:

$$y(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \tag{5.13}$$

The popularity of this function is caused by the fact that it's nonlinear, so can be used for solving complex problems, while its derivative is relatively simple, what makes the backpropagation easy. The application of ReLU has been found very effective for various DNNs. Therefore, ReLU was selected as the activation function at all steps in our network. We also proposed to use batch normalization in residual units of our model [208]. This choice was motivated by the research conducted by Ioffe S. and Szegedy C. [210], which proved that mitigating the problem of covariate shift (distribution of inputs change with the change of network parameters optimized during training) by normalizing not only network inputs by also outputs of subsequent network layers is very efficient in improving the ease of network training.

In case of the second convolution in the proposed residual block, the activation is applied after summing up the main branch with the shortcut connection. The shortcut connection is a side branch, skipping convolutional layers. A single residual block used in the proposed architecture is presented in Fig. 5.3.

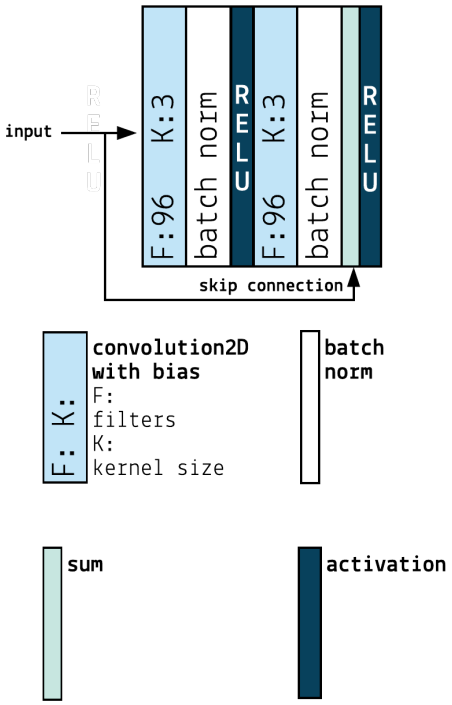


Figure 5.3. Residual block used in the proposed super-resolution neural network

Output from the e^{th} residual block ($F_{fe}^{(e)}$) can be defined as:

$$F_{fe}^{(e)} = \begin{cases} \sigma(g_{fe}^{(e)}(I_{fe}, W_{fe}^{(e)}) + I_{fe}),\ e = 1 \\ \sigma(g_{fe}^{(e)}(F_{fe}^{(e-1)}, W_{fe}^{(e)}) + I_{fe}),\ e \in (1, E\rangle \end{cases} \tag{5.14}$$

where I_{fe} is the input to the feature extraction subnetwork, calculated as the result of applying initial convolution on the generated LR input X:

$$I_{fe} = W_{0\,fe} \circledast X \tag{5.15}$$

E is the number of residual blocks used at feature extraction step and g is a residual block which consists of convolution operations described before, i.e.:

$$g_{fe}^{(e)} = g(x, W_{fe}^{(e)}) = W_{fe}{}_{conv2}^{(e)} \circledast (\sigma(W_{fe}{}_{conv1}^{(e)} \circledast x)) \tag{5.16}$$

The output from the last residual block of the feature extraction step is fed to the second subnetwork aimed at performing non-linear mapping using the recursive approach. Contrary to the DRCN network [192], we further widen the receptive field by using additional residual blocks inside recursions. Thus, the output from d^{th} recursion ($F_{nlm}^{(d)}$) is simultaneously the output after the last (U) residual block ($F_{RES}^{(d,U)}$) within this recursion (d), defined as a sum of the last residual mapping $g_{RES}^{(d,U)} = g(F_{RES}^{(d,U-1)}, W_{RES}^{(d,U)})$ and the input to this recursion I_{nlm}^{d}:

$$F_{nlm}^{d} = F_{RES}^{d,U} = \begin{cases} g_{RES}^{d,U}(F_{RES}^{(d,U-1)}, W_{RES}^{(d,U)}) + I_{nlm}^{(d)}, U > 1 \\ g_{RES}^{d,U}(I_{nlm}^{(d)}, W_{RES}^{(d,U)}) + I_{nlm}^{(d)}, U = 1 \end{cases} \tag{5.17}$$

where d is the current number of the recursion (out of all D recursions), U is the number or residuals in each recursion, residual mapping $g_{RES}^{(d,U)}$ is denoted in the same way as residuals in the FE subnetwork, i.e. by Eq. 5.16 and input to the recursion d is specified as:

$$I_{nlm}^{(d)} = \begin{cases} \sigma(W_{0_{nlm}}^{(d)} \circledast F_{fe}^{(e=E)}), d = 1 \\ \sigma(W_{0_{nlm}}^{(d)} \circledast F_{nlm}^{(d-1)}), d \in (1, D > \end{cases} \tag{5.18}$$

As can be deduced from Eq. 5.17, at the non-linear mapping step we repetitively apply the same set of operations, producing D outputs. Please note that for a single residual block in the recursion ($U = 1$), the input to the residual mapping comes from the input to the whole recursion block ($I_{nlm}^{(d)}$) instead from the previous residual. Each of outputs produced at the non-linear mapping step represent some level of HR data reconstruction.

In the reconstruction subnetwork another convolution operation was applied to the constructed feature maps:

$$F_{nlm}^{(d)} = \sigma(W_{0r} \circledast F_{nlm}^{(d)}), d \in < 1, D > \tag{5.19}$$

HR output is very similar to LR input except some fine, high frequency details. Thus, correlating them together lead to very good restoration capabilities of the model as adjusting only some components of an image is easier than generating them from random distributions. Taking it into account, we added skip connections from LR input X to each of the result from the non-linear mapping step:

$$F_{nlm+X}^{(d)} = F_{nlm}^{(d)} + X, d \in < 1, D > \tag{5.20}$$

Then, all of those outputs were passed to the reconstruction subnetwork, convolved with reconstruction weight kernel (W_r) and used to form the final HR output ($\widehat{Y}$) as the weighted average of

all predictions:

$$\widehat{Y} = F_r = \sum_{d=1}^{D} w^{(d)} \sigma(W_r \circledast F_{nlm+X}^{(d)}) \tag{5.21}$$

where $w^{(d)}$ is the weight associated with d^{th} recursion, and F_r is the output from the reconstruction subnetwork that is equivalent to the reconstructed HR data $\widehat{Y}$. For simplicity of mathematical formulations biases were skipped in all equations.

With the increased network depth, another problem appears. The number of parameters and hence the model size rapidly grows, what leads to a more difficult optimization process. To deal with the issue of the increased number of parameters, we proposed to use shared weights for all residuals within both *fe* and *nlm* subnetworks:

$$\begin{gathered} \forall_{\ e\in<1,E>} W_{fe}^{(e)} = W_{fe} \Rightarrow W_{fe}{}_{conv2}^{(e)} = W_{feconv2} \ and \ W_{fe}{}_{conv1}^{(e)} = W_{feconv1} \\ \forall_{\ u\in<1,U>} \ \forall_{\ d\in<1,D>} W_{RES}^{(d,u)} = W_{RES} \Rightarrow W_{RES}{}_{conv2}^{(d,u)} = W_{RESconv2} \ and \ W_{RES}{}_{conv1}^{(u,d)} = W_{RESconv1} \end{gathered} \tag{5.22}$$

except initial convolution in recursions, i.e. $W_{0_nlm}^{(d)}$. As a result, we only used 3 unique sets of weights at the feature extraction step (one initial convolution W_{0fe} and 2 convolutions in residual blocks with weights shared across all residuals: $W_{feconv1}$ and $W_{feconv2}$), 2 unique sets at the non-linear mapping step shared between all residuals and all recursions ($W_{RESconv2}$ and $W_{RESconv1}$), D sets of weights for the initial convolution in recursions ($W_{0_nlm}^{(d)}$), and one weight matrix for the recursion subnetwork (W_r). For example, the total number of weight matrices for the model with 9 recursions is 15, while the depth of the model might be much bigger, e.g. 27 levels ($U = 9, D = 9, E = 9$). This significant reduction of the parameters number led to the generation of a compact model (9MB), suitable for resource-constraint devices that are often used in remote medical diagnostic applications.

Simplified equations describing the model, taking into account the weight sharing idea, can be formulated as follows.

From Eq. 5.14 and Eq. 5.22:

$$F_{fe}^{(e)} = \begin{cases} \sigma(g_{fe}^{(e)}(I_{fe}, W_{fe}) + I_{fe}),\ e = 1 \\ \sigma(g_{fe}^{(e)}(F_{fe}^{(e-1)}, W_{fe}) + I_{fe}),\ e \in (1, E\rangle \end{cases} \tag{5.23}$$

From Eq. 5.17 and Eq. 5.22:

$$F_{nlm}^{d} = F_{RES}^{d,U} = \begin{cases} g_{RES}^{d,U}(F_{RES}^{(d,U-1)}, W_{RES}) + I_{nlm}^{(d)}, U > 1 \\ g_{RES}^{d,U}(I_{nlm}^{(d)}, W_{RES}) + I_{nlm}^{(d)}, U = 1 \end{cases} \tag{5.24}$$

The network was optimized using the distance between the regressed HR image values and the original HR data. Following the general SR CNN loss function (Eq. 5.12 and predictions produced form applied recursions Eq. 5.21), the final loss can be defined as:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \sum_{d=1}^{D} w^{(d)} \sigma(W_r \circledast F_{NLM+X}^{(d)}))^2 \tag{5.25}$$

where N is the number of samples in each batch used for model training.

Since the model proposed by us uses residuals for calculating embeddings at the feature extraction step and for supervised recursive non-linear mapping we call it Deeply Residual Embedding and Supervised-recursion (DRESNet).

5.3.3 Comparison with Reference Models

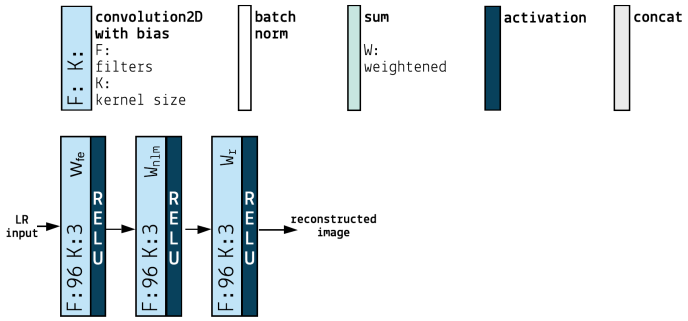
To evaluate the model introduced by us for thermal image enhancement we selected two DL SR CNNs with architectures closest to it: Deeply Recursive Convolutional Network (DRCN) [192] and Deep Recursive Residual Network (DRRN) [193]. The graphical overview of the proposed model and chosen state-of-the-art solutions which are based on similar image enhancement ideas, but designed for visible light data are shown in Fig. 5.4. All weights captions in the DRESNet model are consistent with symbols introduced in the mathematical formulation of the network. Additionally, for all networks, the idea of weights sharing is visualized by using the same weight labels for blocks with the same weights.

As can be deduced from the presented comparison of networks, the proposed model utilizes recursive approach, similarly to DRCN. All produced recursions are supervised, what helps to deal with the overfitting. Yet, there is an essential difference between those architectures introduced by us to improve accuracy of thermal face hallucination using CNNs by better fitting thermal data characteristic. This key idea was based on achievements in the image classification domain which proved that deeper architectures can lead to better performance [39]. Thus, soon after the introduction of SRCNN [190], which revolutionized SR techniques, the research was directed towards models with more layers. In our case, the use of more layers lead to widening of the receptive field and thus obtaining knowledge about facial features by making predictions from more distant relations between them. The use of convolutional filters with the size bigger than 1x1 lead to increased size of receptive field with every additional convolution. Yet, simple stacking of more convolutional layers is not beneficial and may cause gradient to vanish. Thus, we followed a similar approach as ResNet [113] and introduced residual blocks to recursions. Since weights between residuals are shared, as previously explained, the number of parameters remain constant, what is beneficial for network optimization and the model size.

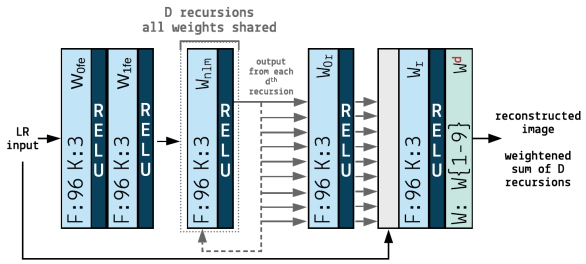
An attempt for combining residuals with recursions in a single network has been already made in the DRRN model [193], presented in Fig. 5.4. However, according to the results presented by its authors, the best accuracy was achieved for the configuration B1U25, meaning that only 1 recursion with 25 residual blocks was used. Thus, we treat DRRN as the residual, non-recursive model, while the DRESNet network, introduced by us, successfully applies residuals and recursions in a single CNN by examining different configurations of the model, i.e. various numbers and placements of each block, described in details in the following section. Secondly, contrary to DRRN, we propose to share weights between all residual blocks across all recursions. DRRN uses weight sharing, but only across residuals, each recursion utilizes a unique set of weights. Therefore, the number of parameters in the non-linear mapping subnetwork of DRESNet is reduced D times comparing to the same subnetwork in DRRN (where D denotes number of recursions). As proved by results presented in [208] and described in Section 5.3.5, introduced modifications allowed for outperforming DRCN and DRRN on examined thermal datasets.

5.3.4 Performed Experiments

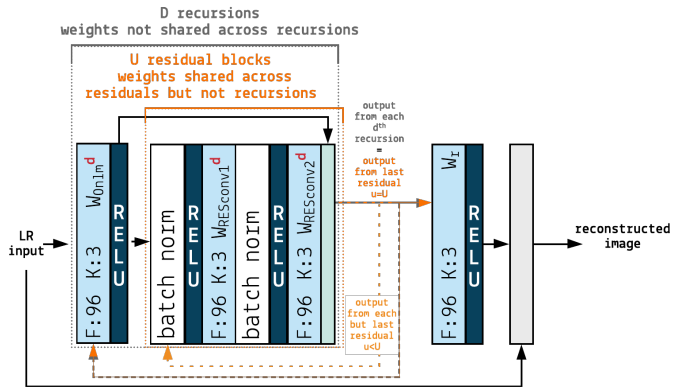
Since the goal of enhancing thermal data is to generate higher resolution sequences that would allow for improving accuracy of facial areas detection and non-contact respiratory rate estimation from detected regions, the proposed method was evaluated on thermal datasets collected by us with this application in mind (i.e. SC3000-ADRA, Lepton-ADRA 3.2.2). At first the model was trained and evaluated on higher resolution images (320x240) from the SC3000-ADRA set. LR data



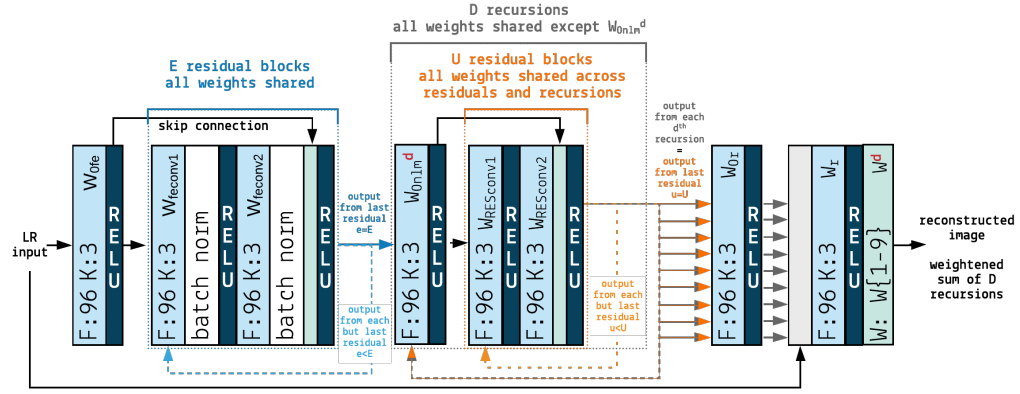
(a) SRCNN



(b) DRCN



(c) DRRN



(d) DRESNet (ours), please note the use of more convolutions in the feature extraction part, in this way the receptive field was widened to take into account more distant relations between facial features in thermal images caused by image blurring (Enlarged version of the proposed model in appendix)

Figure 5.4. Comparison of latest CNN-based SR models evaluated in the study with the proposed network introduced specifically for thermal image enhancement; for all models the same number of filters were used for a fair comparison; upper index 'd' marked in red denotes blocks for which weights are not shared across residuals/recursions

was generated using a scaling factor of 2. Then, to confirm achieved results, similar analysis was performed for lower resolution sequences from the Lepton-ADRA set (80x60). In the case of the Lepton-ADRA dataset, we wanted to determine if extremely low resolution images will still allow for proper image restoration. Thus, LR data was generated using scale 2 and 4 producing images of 40x30 and 20x15 resolution, respectively. After that, to avoid being biased by dataset collected by us, we evaluated the proposed SR DL model and compared it with other state-of-the-art models on the publicly available reference datasets, described in details in Section 3.2.2. In this case, we analysed three scenarios:

1. Models trained on thermal images from the IRIS dataset and evaluated on the IRIS dataset
2. Models trained on visible light data used for the SR task and evaluated on the IRIS dataset
3. Models trained and evaluated on visible light data usually used for the SR task.

In the case of scenario 2, we were able to determine if high frequency components present in visible light images, that the model learns to restore, carry valuable information in thermal data as well. Scenario 3 allowed us to compare the accuracy of the proposed network against existing solutions using commonly utilized test set and determine whether DRESNet is also effective for visible light data, which other models were designed for. Generating super-resolved visible light sequences might also have a huge potential for remote diagnostics, e.g. non-contact heart rate extraction [44]. Although those applications are not the subject of this work, we would like to examine them in future studies. Details of conducted experiments on each dataset are presented in following sub-sections.

SC3000-ADRA: Experiments with Different Windows of Temporal Averaging

Usually datasets used for object detection contain single images only, however, our data was collected as video sequences in order to enable extraction of vital signs. Thus, there is a potential of utilizing this feature for improvement of facial features representation. Temporal averaging of neighbouring frames can be potentially used to remove random noise from collected data. The possible source of the noise include acquisition device noise or influence of the environment. By calculating the average of frames in a given window, those random changes could be smoothed, preserving only important features. In order for this assumption to be true, data acquisition process has to assume that participants stand still during sequence collection. Otherwise, not only noise data but also facial features may become blurred. Although during our data acquisition procedure, subjects were asked to remain still, it's difficult to avoid involuntary motion completely. Some of our previous studies were focused on performing analysis of motion influence on respiratory rate analysis [151], where we proved that attention focusing tasks (e.g. silent reading during data acquisition) have positive effect on reducing motion artifacts. The error for extracted breathing rate during silent reading was 0.27 breaths per minute (bpm), around 5 times smaller than when person was reading aloud. During saying sequences audibly, subjects had to catch breath and most of them were trying to make a proper intonation what had influence on performed motion (it was shown that motion content is higher in such sequences using the Sum of Absolute Differences metric).

In this work, we wanted to estimate if temporal averaging helps with enhancement of thermal sequences by preserving important details and blurring random data, even in scenarios where higher motion content may be present. To perform such experiments, an average of W subsequent

frames was produced and compared against results achieved for a single frame. The window size was selected taking into account frame per second (FPS) parameter of the thermal sensor and an average value of the respiration rate for an adult which is equal ˜12 breaths per minute (bpm) in rest. Three different scenarios were tested:

1. W=1 - average operation not applied
2. W=7 - relatively small window to avoid influence of motion artifacts and respiratory events
3. W=30 - window size equal to number of frames collected during a 1-second period
4. W=90 - window size which covers each respiratory event, as for a respiratory rate of 12 bpm inhalation/exhalation occurs every 2-3 seconds.

The same averaging operation was performed for 8 and 16-bit data, resulting in 1296 images in each of 8 subsets. Some initial images for each volunteer were skipped due to body movements before getting familiar with the procedure, from the rest of samples every 90 (the biggest window) frame was used as a middle frame within each temporal averaging window. Subsets were named using the following convention: bitwidth{8/16}-scale{S2}-window{W1/W7/W30/W90}. Prepared sets were split into train, test and validation parts (70:15:15 proportion), to distinguish those sets from each other, a proper name representing the purpose of the set was appended to the name, i.e. bitwidth{8/16}-purpose{train/val/test}-scale{S2}-window{W1/W7/W30/W90}. As previously explained in the SR objective subsection (Sec. 5.2.1), networks learn how to restore image by passing LR frames through convolutional layers, enhancing it and comparing restored data with original HR images. Therefore, after generating image sets, we had to prepare inputs for DL model training. LR images were modeled by downscaling and upscaling original HR frames with a scale 2, as defined in Eq. 5.4. Examples of original HR images and corresponding LR samples for all window sizes (8 bit-width data) are presented in Fig. 5.5. Fig. 5.6 shows images which represent difference between the middle frame in each window and the calculated average frame. One can note that with the bigger window size, body features become more emphasized.

The proposed SR model and state-of-the-art networks used as a reference (DRCN, DRRN) were optimized using training subsets. The performance was being evaluated after each training epoch using validation subsets. Models were trained on each generated dataset separately using NVIDIA® DGX-1™ Station.

Conducted experiments focused on evaluating how many recursions and residuals in the introduced DRESNet lead to best results in terms of PSNR and SSIM metrics and whether it's more beneficial to apply residuals inside recursions (what turned out to be unsuccessful in DRRN) or before them in the feature extraction part (to address the problem of bigger distances between interesting components in thermal images). Taking it into account, the number of each block was randomly selected in order to find the most optimal configuration of the model. Following the nomenclature introduced earlier (Section 5.3.2), let's denote E as the total number of residuals in the feature extraction subnetwork, D as the total number of recursions at the non-linear mapping step and U as the number of residuals within each of recursion. Thus, tested configurations can be represented as Dx.Ey.Uz, where x, y, z is the number of each D, E, U blocks, selected from subsets $D =< 1 : 9 : 2 >$, $E =< 1 : 9 : 2 >$, $U =< 1 : 9 : 2 >$ ($< startvalue : endvalue : step >$). Hyperparameters tuning was performed using the random search approach on the configuration similar to DRRN, i.e. E=0, D=1, U=9. For a fair comparison, the same hyperparameters were applied for training all selected configurations, which were optimized using Adam optimizer [211]

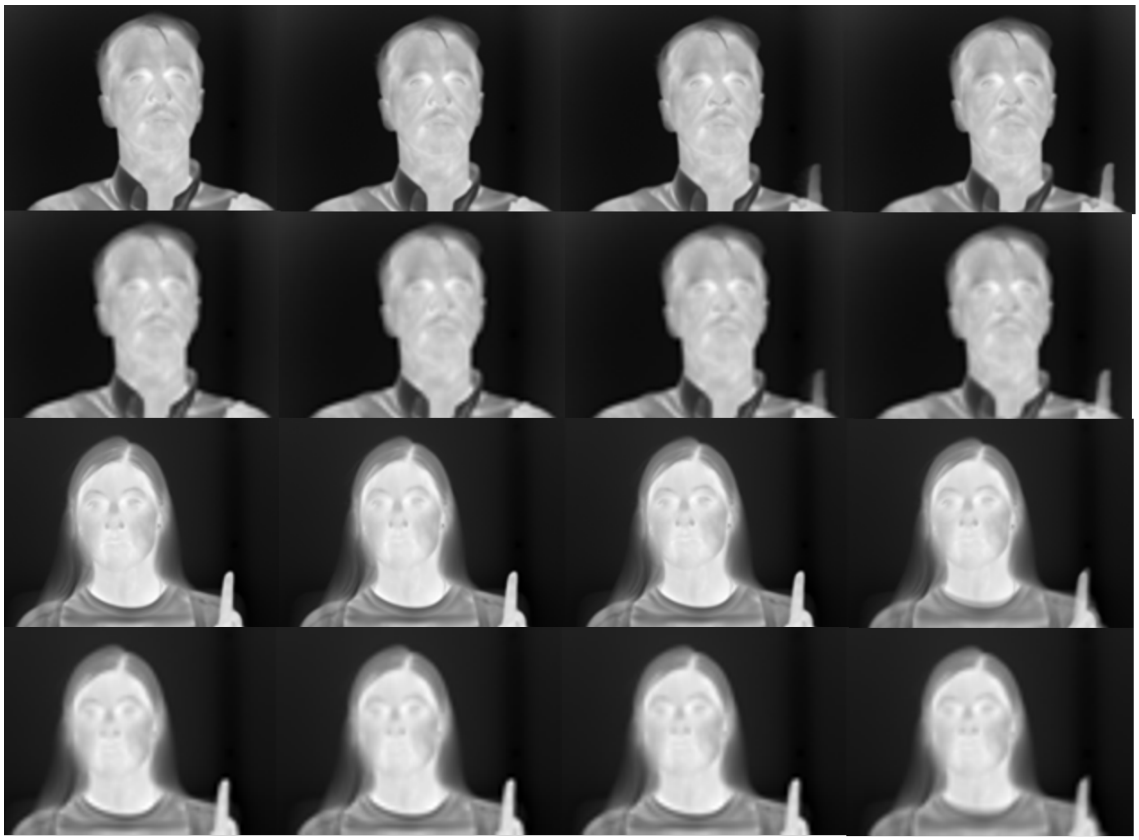


Figure 5.5. Examples of images from SC3000-ADRA-8; from the left: W1, W7, W30, W90; 1^{st} row volunteer 1 HR image, 2^{nd} row volunteer 2 LR, 3^{rd} row volunteer 2 HR, 4^{th} row volunteer 2 LR

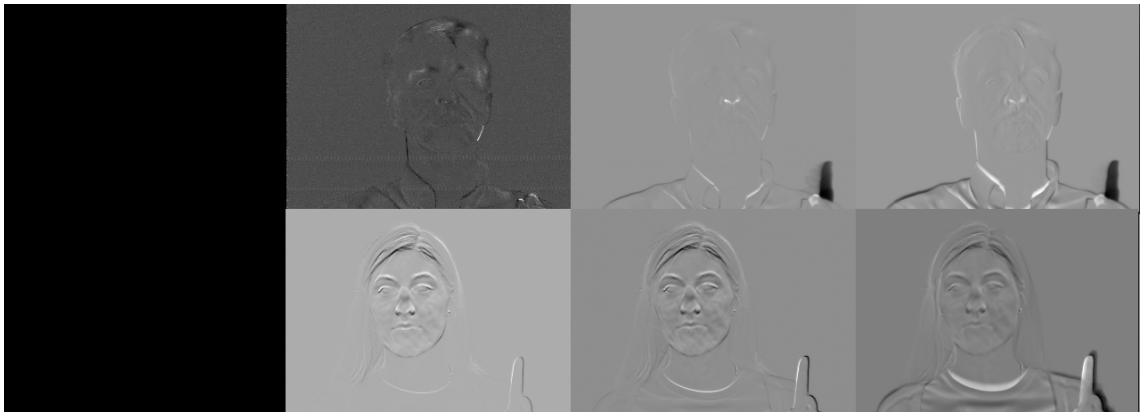


Figure 5.6. Difference between calculated average frame and the middle frame in each window; 1^{st} row volunteer 1, 2^{nd} row volunteer 2

applied to the MSE loss function, defined by Eq. 5.12 in the backward propagation. The initial learning rate was set to 10^{-2} and then reduced by an order of magnitude after 5 subsequent epochs during which the validation error was not decreasing. In addition to the loss calculated for each current batch, a value of the loss for previous batch was also taken into account in the parameters optimization. This technique, known as momentum, has turned out to be beneficial in improving accuracy of NN training. In our case, the momentum factor was set to 0.9, meaning that past gradi-

ents are included in the current update step with the weight of 0.9. Another regularization applied by us aimed at decaying weights values and thus limiting them from becoming too large. This was achieved by multiplying weights by 0.999 after each network update. Following [191], training data were cropped to patches of a size 41x41 with a stride 21. All tested network configurations used convolutional layers with 96 filters of a size 3x3 each. Weights of convolutional kernels were initialized using He algorithm [212]. The training was stopped once the learning rate was smaller than 10^{-5}.

After all introduced DRESNet configurations converged, evaluation of resulting models was performed on the test subset of the SC3000-ADRA dataset. Fig. 5.7 presents the values of the PSNR for various DRESNet configurations. As can be seen, the best results were achieved for residuals applied before recursions ($D1, U = 0, E1$). It turned out that the best PSNR and SSIM results were achieved for 3 residuals in the feature extraction subnetwork and 9 recursions without residuals at the non-linear mapping step. This finding proved the reason for unsuccessful combination of residuals and recursions in DRRN as their model introduced residuals into recursions. We showed that better results are achieved if the receptive field is at first widened at the feature extraction step and then produced feature representations are converted to HR patches using recursions with the simple stack of convolutions. Thus, in further experiments we utilized DRESNet with the D9.E3.U0 configuration, presented in Fig. 5.8.

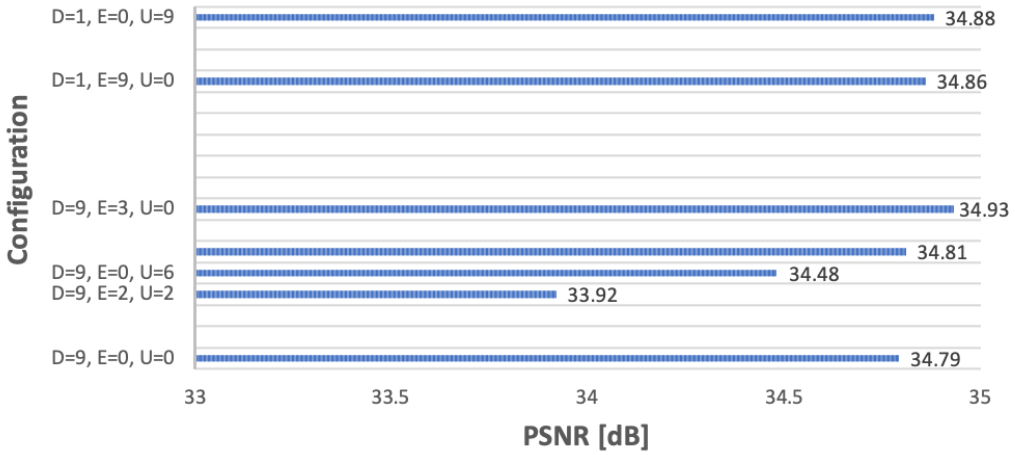


Figure 5.7. Exemplary PSNR values for different DRESNet configurations

For DRCN and DRRN model training, TensorFlow implementation [1], recommended as an alternative version of the original model code was used. All hyperparameters were set to their default values, suggested by authors of those architectures, i.e. [192, 193]. Both networks were configured with their default setting, i.e. 9 residuals without recursion in DRRN and 16 recursions in DRCN. Yet, the number of filters in convolutional layers was modified to be consistent across all tested models. Since, DRESNet uses 96 kernels for all convolutions, the same setup was used for DRRN and DRCN. This modification was motivated by the fact that increased width of convolutions (i.e. more filters) leads to better accuracy [190]. Thus, if networks differ in the number of filters, results may be biased, leading to incorrect conclusions about the introduced architecture, number and placement of residual/recursion blocks. Taking it into account, all tested SR models were using the same number and size of kernels.

[1] https://github.com/LoSealL/VideoSuperResolution Accessed: 2018-10-10

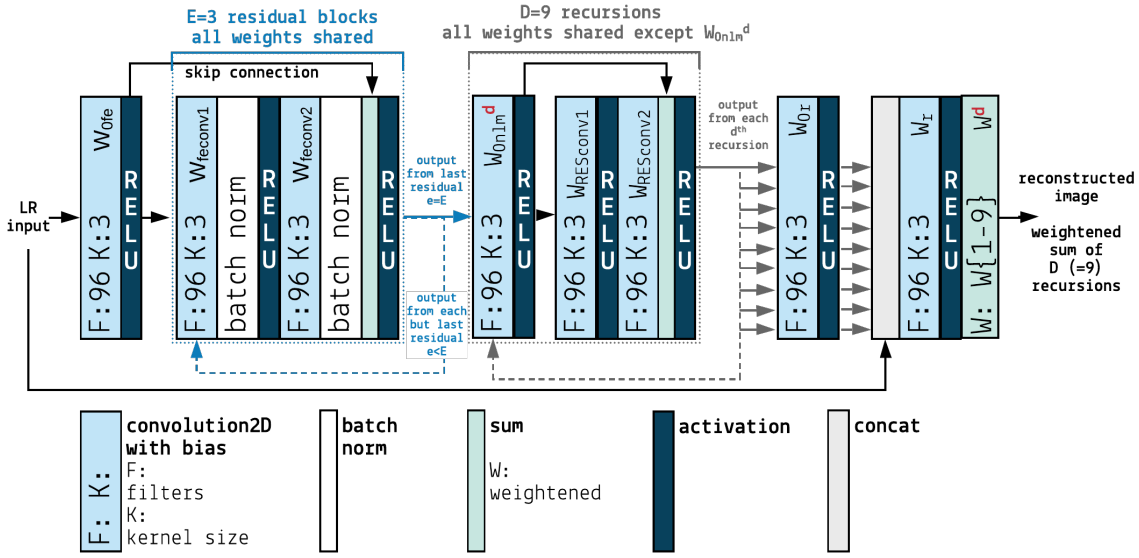


Figure 5.8. Final architecture of the proposed DRESNet model leading to the highest image quality metrics (D=9, E=3, U=0); please note that non-linear subnetwork consists of recursive blocks without residuals, instead residuals are placed in the feature extraction subnetwork what leads to better restoration accuracy; upper index 'd' marked in red denotes weights used in a specific block that are not shared across recursions

Similarly to DRESNet, DRRN and DRCN were evaluated on the test subset of the SC3000-ADRA dataset, using PSNR and SSIM image quality metrics. For experiments we used all generated sets of images, i.e. samples produced for different window sizes used in temporal averaging. We also examined different bit-width of data to determine whether preserving original information (instead of performing lossy conversion from original raw format to PNG images with lower bit resolution) helps with restoration of facial components. Specifically, as explained in Section 3.2.2, 14-bit IR digital images obtained from Lepton camera were stored as both 8 and 16-bit PNG files and compared in performed experiments. Results of the benchmark evaluation performed for all three SR CNNs are presented in Section 5.3.5.

Lepton-ADRA and SC3000-ADRA: Experiments with Different Scaling Factors

After conducting initial experiments with data collected using FLIR® SC3000 sensor [208], we wanted to determine if our solution would work for images with much lower spatial resolution, i.e. 80x60 vs 320x240. Hence, our next study [213] focused on enhancing thermal sequences acquired with the FLIR® Lepton camera. Details about this dataset and parameters of the used thermal sensor were presented in Section 3.2.2. Verification of the proposed novel NN SR architecture on a second database allowed for limiting the possibility of results being biased towards only one set of images.

In addition to using a lower resolution camera, we decided to evaluate the accuracy and robustness of DRESNET on images with extremely small spatial size by using bigger scaling factors that in the previous work [208]. Specifically, we not only applied a scale factor of 2 to simulate resolution degradation, but also a scale factor of 4. As a result, LR inputs to the model, produced by downscaling original HR frames, had sizes of 40x30 and 20x15 for scales 2 and 4, respectively in a case of the Lepton-ADRA set and 160x120 and 80x60 for scales 2 and 4, respectively in a

case of the SC3000-ADRA set. Examples of produced 20x15 Lepton images (8-bit data) upscaled back to the original size of 80x60 and corresponding original HR frames are presented in Fig. 5.9. The blurring and loss of facial features is clearly visible in generated LR images. The goal of the proposed SR model is to restore them, so that the produced outputs are as similar to the original frames as possible.

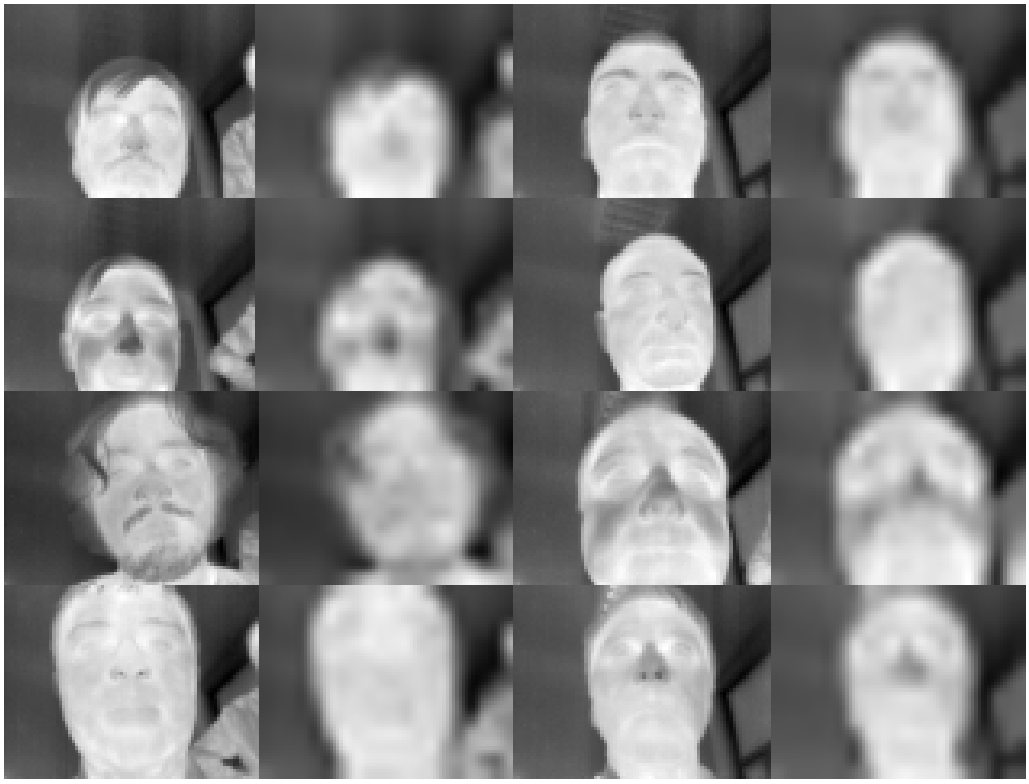


Figure 5.9. Examples of HR samples from the Lepton-ADRA-8 set and corresponding LR images generated with bicubic interpolation using scale 4; different interpolation techniques were used in the dissertation for visualization of low resolution data purposes only, images of original resolution were used in all experiments

Since previous experiments revealed that the best DRESNet configuration (residuals in the feature extraction part instead of inside recursions at the non-linear mapping step) leads to the best values of image quality metrics, we decided to utilize the same configuration in further studies on the influence of scaling factor. At first, the constructed set was divided into two parts: I) data used for training the proposed SR model (from first 15 volunteers), II) data (from the remaining volunteers) which after enhancement with the trained SR network could be used for studies on the influence of face hallucination on accuracy of respiratory rate estimation, described in details in the next chapter (Chapter 6). The latter set was also used to evaluate the accuracy of the proposed DRESNet network using PSNR and SSIM metrics.

To ensure high variability of data, so that the model will learn correct predictions and calculated evaluation metrics will be meaningful, 20 images per volunteer were randomly selected for all three parts of the dataset (total of 300 images). The same data split (15 volunteers for training, remaining for testing) and random selection of frames was done for both 8-bit and 16-bit PNG images. LR inputs for the model were prepared by downscaling and upscaling all images from the constructed sets with a scale factor 2 and 4. After this step, 8 data subsets were created: 8-bit and 16-bit scale

2 and scale 4 images used for SR model training (each set with 300 images - 20 images from 15 volunteers); 8-bit and 16-bit scale 2 and scale 4 images on which the model was evaluated (each set with 320 images - 20 images from 16 volunteers). The nomenclature of those sets is specified as bit-width{8/16}-purpose{train/test}-scale{S2/S4}, e.g. for 8-bit data used for evaluation, created by downscaling images with a scale 4 the name of the subset is Lepton-ADRA-8-test-S4, for the same configuration but used for training, the name is Lepton-ADRA-8-train-S4.

For the SC3000-ADRA dataset, similarly to the Lepton-ADRA set, 4 images from each of first 15 subjects were used for DRESNet optimization to obtain high variability of data. The rest of sequences (from remaining 25 volunteers were utilized for RR estimation studies, described in the next chapter, Sec. 6.3.2). The reason for selecting 4 images per person instead of 30 was motivated by memory limitations of hardware used for training and since SC3000 frames have higher resolution than Lepton ones, only limited number of samples were within this limit. We deduced that this number of images is sufficient for CNN-based SR model training, as each frame is either way further divided into smaller (41x41) overlapping patches using a stride of 21. Data preparation was the same as in case of Lepton-ADRA set, resulting in 8 sets named SC3000-ADRA-bit-width{8/16}-purpose{train/test}-scale{S2/S4}.

The proposed DRESNet model was trained on each set separately, using 10% of samples for validation, which was performed after each training epoch in the same way as for experiments on various averaging window sizes (Sec. 5.3.4). Also, the same hyperparameters as in those previous experiments were utilized for model training (i.e. 41x41 patches extracted using stride of 21, Adam optimizer with momentum set to 0.9, weight decay of 0.999, initial learning rate 10^{-2} decreased by an order of magnitude every 5 epochs for which validation error was not decreasing). Trained models were later used to enhance corresponding test sets, i.e. model trained on Lepton-ADRA-8-train-S2 was used to generate HR samples from the Lepton-ADRA-8-test-S2 set, etc. For all generated HR frames, evaluation was done by calculating values of PSNR and SSIM metrics. Achieved results are presented and discussed in the following section (Section 5.3.5).

Reference Public Datasets: Experiments with Deblurring

The last experiment performed by us involved the use of publicly available thermal dataset of facial images - the IRIS database. All details, such as size of the set, data format, etc. were presented in details in Chapter 3. By using public data, we allow readers to verify their own solutions against results achieved with the use of the proposed novel SR CNN dedicated to processing of thermal images. Eventually, we would also like to make datasets collected by us publicly available. Currently we are working on preparing data structure and selecting a storage for collected sequences. Instructions on how to access datasets will be placed in our repository containing all data associated with our research on thermal image Super Resolution [1].

The IRIS dataset was randomly divided into training, validation, and test sets (8:1:1 split) and used to optimize SR CNNs models. In this experiment DRESNet was compared against SRCNN, DRCN, and DRRN using scaling factor of 2. In addition we also evaluated pixel-2-pixel Tensorflow Generative Adversarial Network (GAN)[2], referred thereafter as p2p, based on the image to image translation idea [214]. It's important to note that p2p network solves a separate problem from the Super Resolution task. Deblurring aims only at mitigating a problem of alleviating effect of convolution, while super-resolved images are also reversed versions of down-sampled data. Since

[1] https://github.com/akwasnie/Super-Resolved-Thermal-Imagery
[2] https://github.com/ceshine/pix2pix-deblur Accessed: 2018-12-29

both of those task are exclusive, we wanted to compare them and evaluate whether deblurring also helps with thermal data enhancement.

All models were trained on the NVIDIA® DGX-1™ Station. For DRESNet, DRCN, and DRRN the same training procedure as in case of our thermal datasets was applied. SRCNN was optimized using TensorFlow version[1] of its original implementation and default hyperparameters [190]. As explained earlier, the bigger filter width may increase network accuracy and thereby lead to false assumption about other aspects of the network architecture. Thus, to ensure fair comparison with other SR CNN models, the number and size of filters in convolutions was set to the same value (96 filters of a size 3x3). For GAN p2p model training the default setting was applied.

After optimizing all DL networks on the training subset of the IRIS dataset, they were evaluated on the test part of the same database. Yet, since most of the models, except ours, were designed for visible light data, an interesting research question is whether the proposed by us novel SR solution could perform equally well on visible light frames as on thermal sets. Taking it into account, all models trained on the thermal dataset IRIS were also evaluated on the visible light Set5.

Furthermore, we studied the effect of utilizing knowledge learnt by models from visible light data on thermal data resolution increase. Since the characteristic of images from different domains varies, in theory worse results would be achieved by directly applying models trained on data from one spectrum to the other. We wanted to verify this assumption by training all SR CNN networks on the combined BSD and SPSR visible light databases and then calculating achieved metrics on images from two different domains: I) test subset of the IRIS thermal set II) Set5 containing RGB pictures acquired in visible light. Set5, BSD and SPSR are visible light benchmarking sets commonly used for validating SR algorithms, previously described in Chapter 3, Section 3.2.2.

5.3.5 Results and Discussion

Image quality metrics: Peak Signal-to-Noise Ratio and Structural Similarity Metric calculated for images from SC3000-ADRA-{8/16}-test-S2-{W1/W7/W30/W90} sets generated with Super Resolution models are presented in Tables 5.1 and 5.2, respectively.

Table 5.1. Experiments with different averaging window sizes: Peak Signal-to-Noise Ratio (for facial regions and frames as a whole) for generated 8 and 16-bit LR images and enhanced with DRCN, DRRN or DRES(Net) (our) SR models (red - first best, blue - second best for each region and within each averaging window separately)

	SC3000-ADRA-8-test-S2-W1				**SC3000-ADRA-16-test-S2-W1**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.90 ±0.10	29.86 ±1.84	41.01 ±1.81	43.87 ±1.58	27.90 ±0.11	51.29 ±0.11	52.12 ±0.34	53.06 ±0.39
face	27.92 ±0.10	30.28 ±1.86	40.73 ±1.65	44.20 ±1.78	27.90 ±0.05	51.31 ±0.09	52.11 ±0.53	53.88 ±0.56
nose	27.93 ±0.21	30.36 ±1.52	41.72 ±1.55	44.98 ±1.86	27.89 ±0.14	51.28 ±0.13	52.13 ±0.30	53.38 ±0.63

[1] https://github.com/tegg89/SRCNN-Tensorflow Accessed: 2018-12-29

frame	27.91 ±0.16	31.49 ±2.37	43.07 ±1.06	47.49 ±1.28	27.90 ±0.11	51.29 ±0.11	52.12 ±0.39	53.36 ±0.61
	SC3000-ADRA-8-test-S2-W7				**SC3000-ADRA-16-test-S2-W7**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.88 ±0.10	30.14 ±2.73	41.94 ±1.97	44.72 ±1.70	27.89 ±0.10	51.33 ±0.26	52.12 ±0.37	53.39 ±0.46
face	27.90 ±0.03	30.66 ±2.30	41.22 ±1.50	44.33 ±1.64	27.90 ±0.02	51.34 ±0.16	51.95 ±0.47	53.98 ±0.48
nose	27.89 ±0.14	30.36 ±2.54	41.28 ±1.71	44.57 ±1.62	27.91 ±0.15	51.32 ±0.24	52.03 ±0.26	53.11 ±0.51
frame	27.89 ±0.10	31.15 ±2.67	43.18 ±1.01	47.45 ±1.21	27.90 ±0.11	51.33 ±0.23	52.05 ±0.38	53.47 ±0.58
	SC3000-ADRA-8-test-S2-W30				**SC3000-ADRA-16-test-S2-W30**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.91 ±0.11	29.44 ±2.47	42.11 ±2.16	44.85 ±2.00	27.90 ±0.02	51.33 ±0.14	52.12 ±0.32	53.44 ±0.63
face	27.90 ±0.02	30.59 ±2.36	41.68 ±1.72	44.49 ±1.73	27.90 ±0.02	51.35 ±0.12	51.95 ±0.39	54.28 ±0.60
nose	27.89 ±0.13	30.78 ±2.93	42.59 ±1.65	45.54 ±1.93	27.91 ±0.11	51.31 ±0.14	52.09 ±0.30	53.90 ±0.77
frame	27.90 ±0.10	31.51 ±2.92	43.76 ±1.18	47.69 ±1.28	27.90 ±0.09	51.33 ±0.14	52.07 ±0.34	53.78 ±0.75
	SC3000-ADRA-8-test-S2-W90				**SC3000-ADRA-16-test-S2-W90**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	27.89 ±0.10	30.90 ±3.32	43.77 ±2.29	46.55 ±2.08	27.91 ±0.15	51.24 ±0.21	52.22 ±0.46	53.95 ±0.82
face	27.90 ±0.03	31.64 ±2.81	42.84 ±1.07	46.00 ±1.77	27.90 ±0.03	51.26 ±0.18	52.02 ±0.55	54.64 ±0.77
nose	27.90 ±0.11	32.34 ±3.14	43.82 ±1.86	46.54 ±1.80	27.90 ±0.13	51.26 ±0.19	52.13 ±0.44	54.13 ±0.81
frame	27.89 ±0.09	31.81 ±3.09	44.62 ±1.19	49.02 ±1.25	27.90 ±0.10	51.25 ±0.20	52.14 ±0.49	54.18 ±0.85

Table 5.2. Experiments with different averaging window sizes: Structural Similarity Index (for facial regions and frames as a whole) for generated 8 and 16-bit LR images and enhanced with DRCN, DRRN or DRES(Net) (our) SR models (red - first best, blue - second best for each region and within each averaging window separately)

	SC3000-ADRA-8-test-S2-W1				**SC3000-ADRA-16-test-S2-W1**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.71 ±0.28	0.88 ±0.04	0.98 ±0.01	0.99 ±0.00	0.71 ±0.28	0.85 ±0.04	0.97 ±0.01	0.99 ±0.01
face	0.64 ±0.27	0.93 ±0.01	0.98 ±0.00	0.99 ±0.00	0.64 ±0.27	0.91 ±0.01	0.98 ±0.01	0.99 ±0.00
nose	0.53 ±0.39	0.92 ±0.04	0.99 ±0.01	0.99 ±0.00	0.53 ±0.39	0.90 ±0.03	0.99 ±0.01	0.99 ±0.00
frame	0.64 ±0.32	0.89 ±0.06	0.98 ±0.01	0.99 ±0.01	0.64 ±0.32	0.92 ±0.01	0.98 ±0.01	0.99 ±0.01
	SC3000-ADRA-8-test-S2-W7				**SC3000-ADRA-16-test-S2-W7**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.73 ±0.25	0.96 ±0.01	0.98 ±0.01	0.99 ±0.01	0.73 ±0.25	0.94 ±0.0.05	0.98 ±0.01	0.99 ±0.01
face	0.63 ±0.26	0.96 ±0.01	0.98 ±0.00	0.99 ±0.00	0.63 ±0.26	0.96 ±0.01	0.98 ±0.01	0.99 ±0.00
nose	0.50 ±0.41	0.96 ±0.01	0.99 ±0.01	0.99 ±0.01	0.50 ±0.41	0.96 ±0.01	0.99 ±0.01	0.99 ±0.01
frame	0.64 ±0.32	0.89 ±0.06	0.98 ±0.01	0.99 ±0.01	0.64 ±0.32	0.94 ±0.01	0.98 ±0.01	0.99 ±0.01
	SC3000-ADRA-8-test-S2-W30				**SC3000-ADRA-16-test-S2-W30**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES
eye	0.70 ±0.25	0.96 ±0.01	0.98 ±0.01	0.99 ±0.01	0.70 ±0.25	0.87 ±0.06	0.98 ±0.02	0.99 ±0.01
face	0.61 ±0.26	0.96 ±0.01	0.98 ±0.00	0.99 ±0.00	0.61 ±0.26	0.93 ±0.01	0.99 ±0.01	0.99 ±0.00
nose	0.50 ±0.36	0.98 ±0.01	0.99 ±0.01	0.99 ±0.00	0.50 ±0.36	0.91 ±0.03	0.99 ±0.01	1.00 ±0.00
frame	0.62 ±0.30	0.90 ±0.06	0.98 ±0.01	0.99 ±0.01	0.62 ±0.30	0.93 ±0.01	0.98 ±0.01	0.99 ±0.01
	SC3000-ADRA-8-test-S2-W90				**SC3000-ADRA-16-test-S2-W90**			
region	bicubic	DRCN	DRRN	DRES	bicubic	DRCN	DRRN	DRES

eye	0.75 ±0.21	0.97 ±0.01	0.99 ±0.01	0.99 ±0.00	0.75 ±0.21	0.96 ±0.03	0.99 ±0.01	0.99 ±0.01
face	0.64 ±0.24	0.98 ±0.00	0.99 ±0.00	0.99 ±0.00	0.64 ±0.24	0.97 ±0.01	0.99 ±0.01	0.99 ±0.00
nose	0.48 ±0.37	0.98 ±0.01	0.99 ±0.01	0.99 ±0.00	0.48 ±0.37	0.96 ±0.02	0.99 ±0.00	1.00 ±0.00
frame	0.65 ±0.29	0.90 ±0.06	0.98 ±0.01	0.99 ±0.01	0.65 ±0.29	0.93 ±0.01	0.98 ±0.01	0.99 ±0.01

Examples of thermal images from SC3000-ADRA-8-test-S2-W1 set are presented in Fig. 5.10. Facial area in each image was enlarged to better show features representation in original data. Then, a size of a region with chosen facial feature was further increased to visualize difference of feature quality after applying various SR networks vs. LR image generated with bicubic interpolation. Enlarged regions of original HR data were not presented intentionally, as a goal was to visualize differences between LR and SR patches. As can be seen, visually there are no big differences between applied models, but some improvement of features representation comparing to LR data is visible. Comparison of eye and nose areas extracted from LR image (SC3000-ADRA-8-test-S4-W1 set) generated with bicubic interpolation and its corresponding enhanced version (processed with the proposed DRESNet network) are presented in Fig. 5.11 and 5.12, respectively.

Table 5.3 shows comparison of image quality metrics (PSNR and SSIM) calculated on both datasets for different image scaling factors. Presented values include results for 8 and 16-bit LR images generated with bicubic interpolation and then enhanced with the proposed thermal super-resolution Deep Neural Network.

Relation between PSNR calculated for the reference thermal dataset IRIS and the number of residuals applied to enhance thermal image resolution in feature extraction subnetwork of the proposed SR model (DRESNet) is presented in Fig. 5.13.

Table 5.3. Experiments with different scaling factors: PSNR and SSIM for sequences downscaled with bicubic interpolation using scale of 2 and 4 and then enhanced with the proposed SR DL model (red - first best, blue - second best for each dataset, separately for SSIM and PSNR).

Dataset	**Method**	**Bit resolution**	**Evaluation Metrics**	
			PSNR	**SSIM**
Lepton-ADRA-test-S2	**bicubic**	8 bits	41.82 ± 0.55	0.96 ± 0.01
		16 bits	63.82 ± 0.73	0.99 ± 0.01
	DRESNet	8 bits	43.21 ± 0.33	0.97 ± 0.01
		16 bits	72.92 ± 4.47	0.99 ± 0.01
Lepton-ADRA-test-S4	**bicubic**	8 bits	39.91 ± 0.46	0.81 ± 0.11
		16 bits	61.31 ± 0.69	0.99 ± 0.01
	DRESNet	8 bits	42.18 ± 0.24	0.95 ± 0.01
		16 bits	67.34 ± 5.83	0.99 ± 0.02

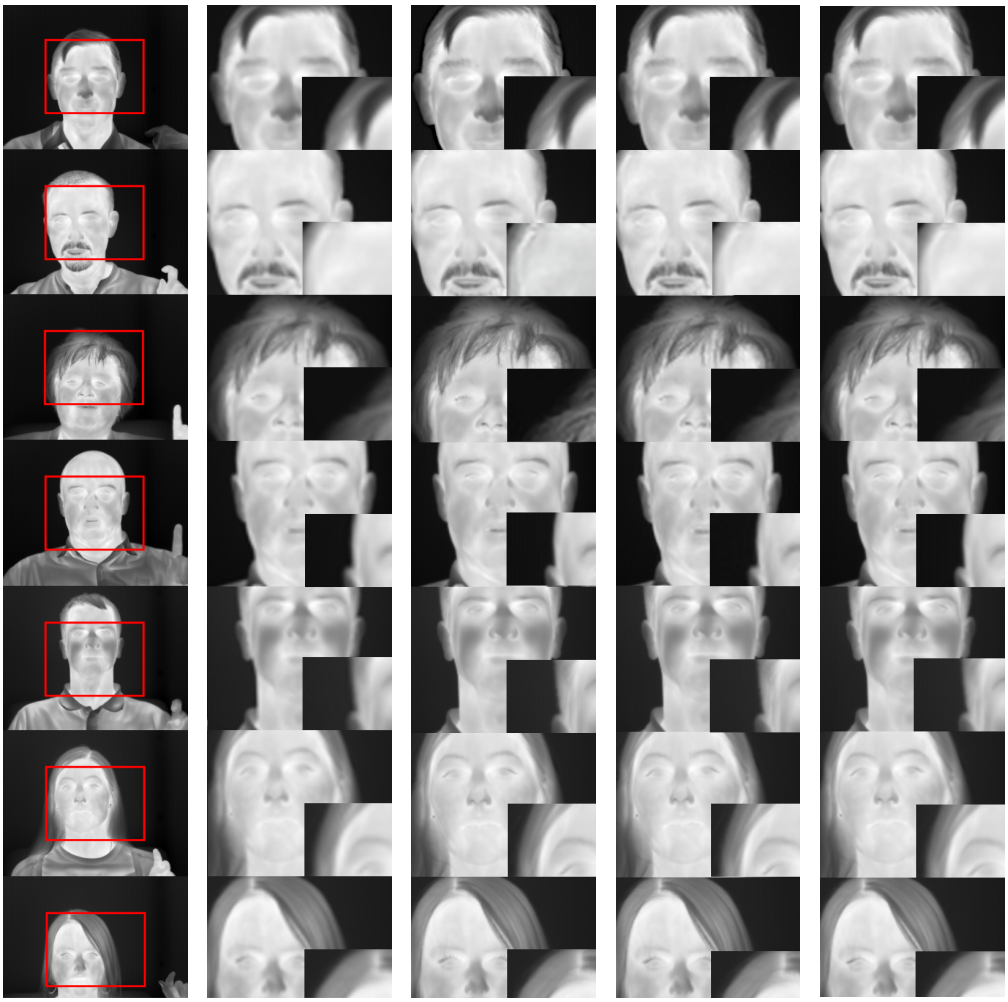
SC3000-ADRA-test-S2	**bicubic**	8 bits	42.69 ± 3.36	0.81 ± 0.22
		16 bits	68.98 ± 1.02	0.99 ± 0.01
	DRESNet	8 bits	43.61 ± 0.18	0.96 ± 0.01
		16 bits	70.05 ± 0.91	0.99 ± 0.01
SC3000-ADRA-test-S4	**bicubic**	8 bits	41.36 ± 2.30	0.79 ± 0.21
		16 bits	65.74 ± 0.95	0.99 ± 0.01
	DRESNet	8 bits	43.97 ± 0.22	0.96 ± 0.01
		16 bits	66.50 ± 0.93	0.99 ± 0.01

Results of experiments with reference thermal and visible light data are presented in Tables 5.4, 5.5, 5.6. Image quality metrics (Peak Signal-to-Noise Ratio and Structural Similarity Index Metric) collected in those tables allow to evaluate whether 1) performance of the proposed thermal image enhancement model is sufficient also for other publicly available databases to ensure it's not biased toward our sets - Table 5.4; 2) features optimized on visible light images are transferable to Super Resolution task applied to thermal data leading to satisfactory restoration accuracy, as models were trained in visible light domain and then utilized for thermal image enhancement - Table 5.5; 3) the proposed model is also applicable to visible light image enhancement, even though its structure was designed with thermal image representation in mind - Table 5.6. In addition, we also present results of deblurring algorithm and compare it with the proposed SR model in Table 5.4, since those tasks are exclusive and we believe they should produce similar results. Qualitative results of applying tested and proposed SR Deep Neural Networks, as well as deblurring algorithm on the same source image from the reference thermal database IRIS are presented in Fig. 5.14.

Table 5.4. Experiments with reference thermal dataset and deblurring algorithm: Peak Signal-to-Noise Ratio (top row) and Structural Similarity Index Metric (bottom row) for the IRIS test subset downscaled and then upscaled with a scale 2, all models trained on the IRIS training subset (red - first best, blue - second best)

bicubic	p2p-deblur	SRCNN	DRCN	DRRN	DRESNet
29.4 ± 0.13	30.10 ± 1.12	29.61 ± 0.36	31.33 ± 0.78	34.01 ± 0.41	34.93 ± 0.55
0.78 ± 0.02	0.83 ± 0.01	0.82 ± 0.02	0.88 ± 0.01	0.88 ± 0.01	0.893 ± 0.02

In-depth benchmark evaluation performed for various datasets proved that the proposed super-resolution CNN leads to the increase of thermal image quality expressed by PSNR and SSIM metrics. Comparison with other state-of-the-art solutions used for image enhancement (see Table 5.1 and 5.2) showed that the use of more blocks in the feature extraction subnetwork and thus widening of the receptive field is beneficial in case of thermal data, where dependencies between interesting components are bigger. The proposed DRESNet model outperformed other solutions by a large margin. In the best case by 21.13 dB comparing to bicubic interpolation, 17.21 dB to DRCN and 4.4 dB to DRRN. Calculated SSIM values showed advantage of the proposed architecture over most of other models, except DRRN. It may be caused by the fact that additive noise is better



(a) Original HR data (b) Generated LR data (c) Enhanced with DRCN (d) Enhanced with DRRN (e) Enhanced with DRESNet

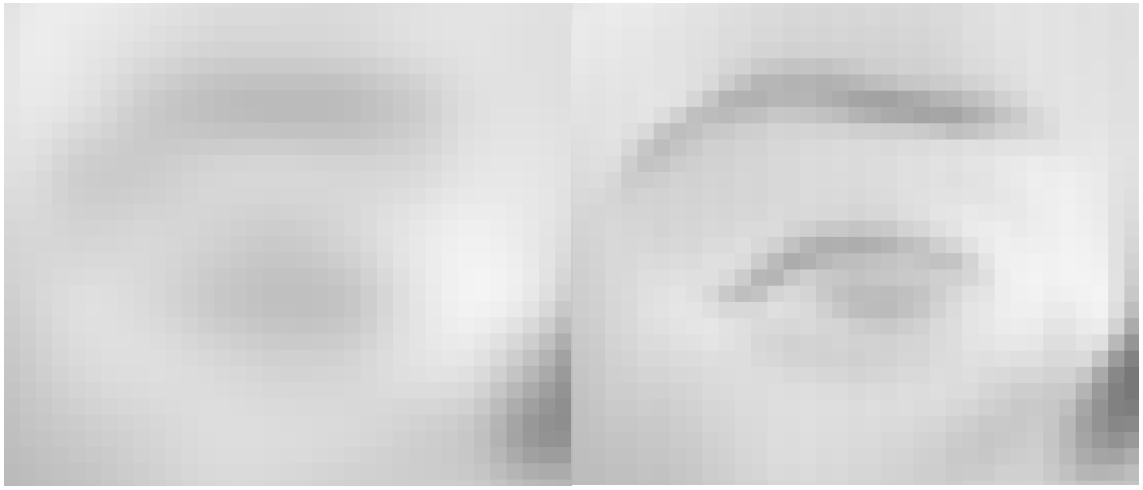
Figure 5.10. Examples of original HR thermal images from SC3000-ADRA-8-test-S2-W1 set, LR samples generated with bicubic interpolation scale 2 and their enhanced versions produced using evaluated SR models

Table 5.5. Experiments with reference thermal dataset and transfer of knowledge from visible light images: Peak Signal-to-Noise Ratio (top row) and Structural Similarity Index Metric (bottom row) for the IRIS test subset downscaled and then upscaled with a scale 2, all models trained on the BSD+SPSR training subset (red - first best, blue - second best)

bicubic	SRCNN	DRCN	DRRN	DRESNet
29.4 ± 0.13	29.59± 0.32	29.63 ± 0.27	33.67 ± 0.37	34.29 ± 0.89
0.78 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.87 ± 0.02	0.89 ± 0.02

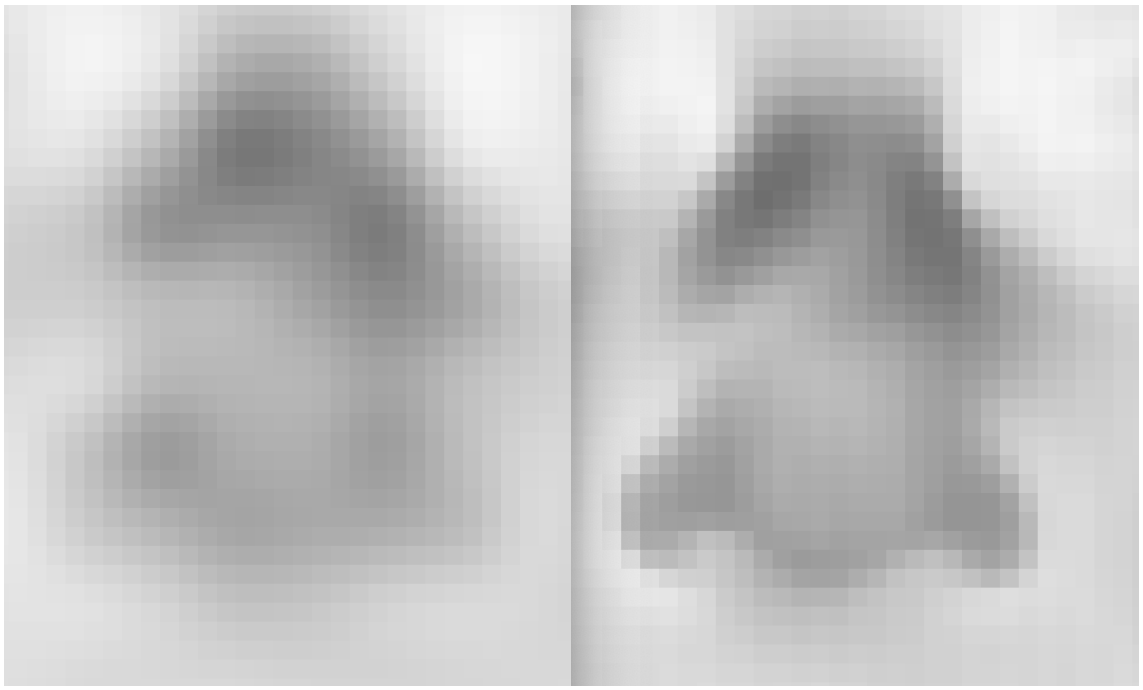
reflected by PSNR metric and even in case of its small values, it may become exceedingly prevalent in very small resolution data. This relation is not that clear in SSIM values.

Analysis of various window sizes used for averaging frames turned out to be beneficial for further enhancement of low resolution inputs. The highest values of image quality metrics were obtained



(a) scaled down (4x) with bicubic interpolation (b) after applying the proposed DRESNet model

Figure 5.11. Example of extracted eye area



(a) scaled down (4x) with bicubic interpolation (b) after applying the proposed DRESNet model

Figure 5.12. Example of extracted nose area

Table 5.6. Experiments with visible light dataset: Peak Signal-to-Noise Ratio (top row) and Structural Similarity Index Metric (bottom row) for the Set5 visible light data downscaled and then upscaled with a scale 2, all models trained on the BSD+SPSR training subset (red - first best, blue - second best)

bicubic	SRCNN	DRCN	DRRN	DRESNet
30.2 ± 0.13	32.35 ± 0.32	33.63 ± 0.27	35.50 ± 1.89	35.84 ± 1.86
0.82 ± 0.02	0.87 ± 0.02	0.89 ± 0.02	0.93 ± 0.05	0.94 ± 0.05

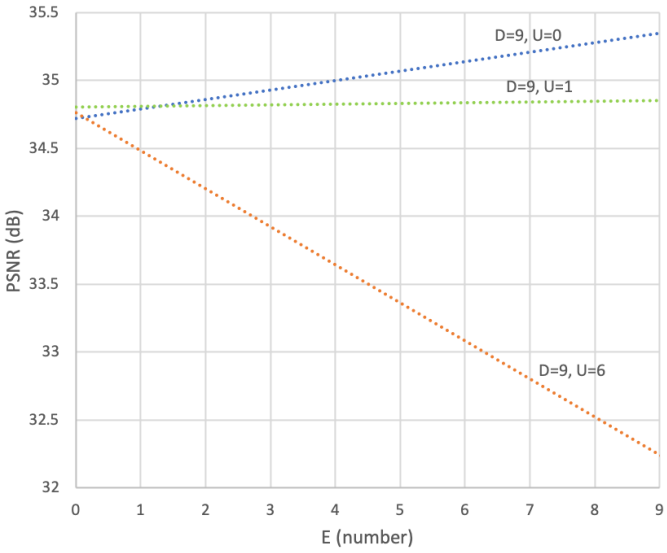
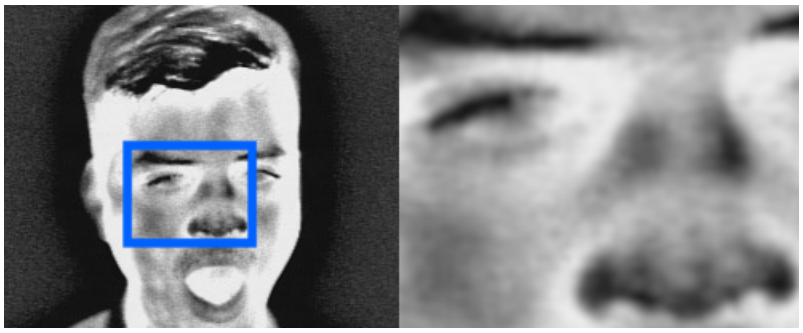


Figure 5.13. Relation between Peak Signal-to-Noise Ratio (PSNR) and the number of residuals in the feature extraction subnetwork (E) at a given number of recursions (D); configuration with residuals placed inside recursions was also visualized to show that this setup is unsuccessful, as proved in previous experiments (Fig. 5.7)

for big averaging windows (90 adjacent frames) in case of all tested Super Resolution models. This result may indicate the need of applying additional pre-processing steps or focusing volunteers' attention on some tasks (as showed in the previous chapter (Sec. 4.2.1)) in order to reduce motion. Yet, it may be still difficult to completely avoid involuntary movements during data acquisition. In such cases, an interesting research problem would be to make use of differential images (see Fig. 5.6) that would expose object features while blurring redundant background details.

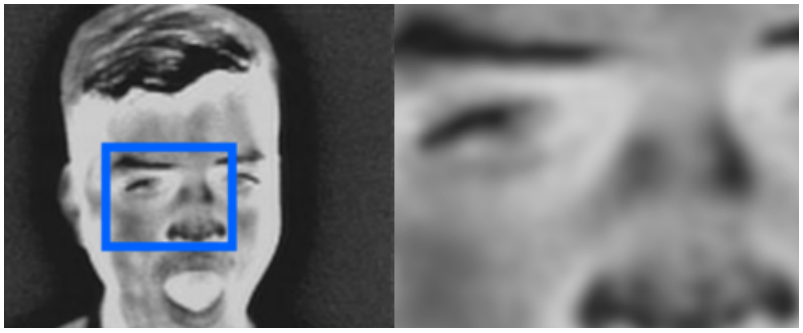
Moreover, as can be seen in Tables 5.1, 5.2, and 5.3, the need of using original bit resolution is an important finding of the presented work. We can observe gain of PSNR of at least 10% for sequences converted to higher bit resolution (16 bits) from original 14-bit format. Lossy compression to 8-bit PNG format led to decrease of image quality. In many cases, the difference of PSNR metric between 8 and 16-bit data was even higher, e.g. for window of a size 1, eye area the difference was ˜25%. Higher bit resolution of images turned out to be beneficial for other models, as well. The difference of PSNR between DRCN and DRESNet was reduced 5 times (from 15.36 dB to 2.97 dB) if 16-bit sequences were used (window size 90). Similarly for DRRN, we noticed the reduction of PSNR difference by around 10% comparing to 8-bit sequences. This finding proves requirement for one of tasks defined in our study - the need of creating thermal image database with preserved original raw data format. Thus, facial sets collected by us (see Chapter 3) may become a useful reference for further studies on thermal image processing and we are planning to make them publicly available.

In addition, experiments performed for different scaling factors (Table 5.3) showed that SC3000-ADRA set downscaled 4 times and then enhanced with Deep Neural Network produced even higher PSNR and SSIM values than for scaling factor of 2. This result proves the high robustness of the proposed SR model and confirms that it's possible to learn efficient image restoration function that would allow for improving image resolution even for inputs with a very small spatial sizes and low quality. Therefore, theoretically low-cost thermal sensors can be used in remote medical diagnostics leading to accuracy equivalent to the one achieved with more advanced cameras.



(a) Original HR data

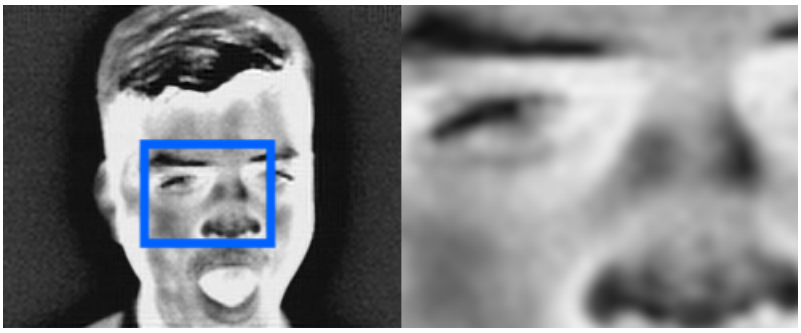
(b) LR data, scale 2, PSNR 28.45



(c) Enhanced with SRCNN, PSNR 29.16

(d) Enhanced with p2p deblur, PSNR 29.62

(e) Enhanced with DRCN, PSNR 31.86



(f) Enhanced with DRRN, PSNR 34.05

(g) Enhanced with DRESNet, PSNR 34.25

Figure 5.14. Results of applying selected SR methods and deblurring algorithm (pix2pix) on the same source image from IRIS set and calculated PSNR metric

Based on the quantitative analysis of different SR networks on reference thermal and visible light databases, it was confirmed that increased number of convolutions in the feature extraction subnetwork lead to the higher values of PSNR. Our proposed model achieved the highest values of PSNR and SSIM metrics on IRIS (Table 5.4 and 5.5) and RGB Set5 (Table 5.6) databases, outperforming other state-of-the-art solutions. Also, our study showed that residuals can be successfully used with recursions in a single CNN network. Previous attempts of combining these blocks in DRRN [193] did not lead to the higher accuracy and the proposed solutions ended up using either recursions or residuals, but not both of them at the same time. Our extensive benchmark evaluation proved that the best results are achieved if residuals are placed before recursions, not inside them as in DRRN. PSNR results of block placement analysis are also visualized in Fig. 5.13. As can be observed, the increase of PSNR value is the highest for residuals applied before recursions (D=9, U=0, increasing E). Even if only one residual is used in non-linear mapping subnetwork (U=1) together with recursive blocks (D=9), PSNR values remain almost constant. This is an important finding for designing SR Deep Neural Networks. The proper placement of specific blocks should be carefully analysed to improve the image enhancement process.

Furthermore, experiments performed for networks trained on visible light images showed that in the case of SR algorithms transfer knowledge from visible light spectrum is not sufficient. All models trained on RGB images (BSD+SPSR set - Table 5.5) led to worse results of the PSNR metric than models trained from scratch on thermal data (IRIS set - Table 5.4). This may indicate that features learnt from visible light images are sufficient for performing image classification, as shown in Chapter 4, but it's not possible to restore proper representation of thermal regions which lack in high frequency components present in visible light spectrum data.

Results produced by the GAN p2p network aimed at performing deblurring operation showed

that its ability to restore thermal features is worse than for convolutional-based models by almost 5 dB comparing to the architecture proposed by us. Since GAN models are more difficult to train [39], the limited amount of training samples may be a reason for overfitting and lack of generalization to new samples, leading to worse image quality metrics. Taking it into account we would like to perform more experiments with GAN-based super-resolution and deblurring models in order to evaluate possibility of applying them to image domains other than visible light spectrum.

5.4 Problems

Apart from ideas for accuracy and performance improvement, discussed in the previous Section, another issue associated with the selection of network structure was identified by us. Although the achieved results are promising, we are aware of some limitations of the introduced thermal image enhancement solution. First of all, the configuration of neural network was chosen from a limited set of parameters in our studies [213, 208]. However, for such complicated architectures this approach may not be efficient due to the confined number of possible combinations of utilized building blocks. Thus, we would like to apply evolutionary algorithms that would generate the most promising configuration by evolving the proposed neural network. Also, there are other techniques that could help with model training and increasing its performance, such as gradient clipping or data augmentation.

Furthermore, according to results achieved by state-of-the-art Super Resolution models on visible light data, image quality of data produced using GANs is better than if CNNs are applied. The novel SR structure proposed by us was based on CNNs due to ease of training them comparing to generative models. Yet, more and more studies have been recently focused on improving stability of GANs [215], thus it's important to analyse whether existing solutions are suitable for thermal image processing and compare results achieved by them with the proposed DRESNet. Some publications already considered application of GANs to thermal image enhancement task [207], but they lack a quantitative comparison of the proposed model with other SR networks. We would like to perform such analysis in the future work.

Details about neuroevolution approach and explanation how it could be used for generation and training of SR CNN model as well as the initial analysis of suitability of existing GAN models to task of enhancing thermal images and other future work directions are provided in Chapter 7.

5.5 Summary

In this Chapter, we explained the objective of Super Resolution task and provided in-depth overview of existing Super Resolution algorithms with the main focus on Convolutional Neural Network based solutions. Presented DL models were analyzed in order to identify blocks and structures that could be potentially useful for thermal image processing, taking into account differences in representation of images between different domains.

After performed analysis, a novel Deep Neural Network designed for enhancing thermal image sequences was introduced. Contrary to other state-of-the-art models, the architecture of the proposed solutions was selected with the characteristic of thermal images in mind. Specifically, we proposed to widen a receptive field, so that the model will be able to learn more distant relations between facial region features that are observed in thermal imaging. Most of other solutions were

designed for visible light images for which this problem is not valid since objects are represented by high frequency components that can be easily extracted with smaller receptive fields.

The accuracy and robustness of the proposed SR model was verified on a wide range of thermal and visible light databases, proving that the introduced architecture outperforms previous solutions in the image enhancement task. Conducted experiments included evaluation of different image downscaling factors, showing that image can be restored even from inputs as small as 20x15, producing satisfactory values of image quality metrics. This opens a lot of possibilities for remote medical diagnostic solutions that could provide non-disruptive way of monitoring health and emotional status of people during their daily activities. In-depth evaluation performed on thermal images with different resolutions and acquired with various thermal cameras prove the first part of the thesis II defined in the presented dissertation, as it has been shown that the introduced novel architecture of Deep Convolutional Neural Networks allows for increasing resolution of those sequences, outperforming previous super-resolution methods.

The second part of the formulated thesis II specifies the need for evaluating whether resolution of thermal images increased with the proposed DL model lead to improvement of facial areas detection accuracy. Our motivation for those studies is based on the fact that some examples of telemedicine solutions can potentially take advantage of more accurate localization of facial features, e.g. proposed and studied by us non-contact estimation of vital signs [169, 172], person identification [216] and emotion recognition from extracted video-based vital signs [158]. Next Chapter (Chapter 6) is devoted to introduction and explanation of those ideas in order to evaluate authenticity and genuineness of the second part of the thesis II.

Chapter 6

Improvement of Contactless Vital Signs Estimation

6.1 Introduction and Overview

The main goal of our work is to propose novel Deep Learning (DL) solutions for thermal image processing in order to enable new innovative applications in the area of remote medical diagnostics. The primary motivation for such studies is global aging and influence of latest inventions (i.e. wearable devices, smart home infrastructure) on our lives. Global aging and demographic shift in many societies (by 2030 there will be 3 times more super-aged nations around the world than today [41]) has revolutionized the current healthcare definition and led to growing expectations of healthcare providers to deliver solutions that would allow to perform some medical activities outside professional institutions, preferably without any supervision, e.g. medical consultations, such as melanoma detection [217]; virtual nursing [218]; remote person monitoring by collecting vital signs [44, 172] or evaluating physiological state [219]; therapy support, such as phobias treatment [220].

On the other hand, increased self-awareness of societies and easier access to health monitoring and tracking applications and devices steer the direction of a lot of studies towards Artificial Intelligence (AI) driven telemedicine use cases. DL has enabled many of such solutions by providing tools for generating predictions from various input sources (e.g. RGB cameras, microphones, low-cost thermal sensors, etc.) with human-like accuracy. For example, nowadays smart watches allow to track daily activities, measure basic vital patterns and has been proved to accurately detect emergency situations [221]. Similar solutions could be also embedded into other devices within the smart home infrastructure, e.g. smart home speakers reminding about regular exercises, or smart kitchen appliances equipped with algorithms for suggesting proper nutrition, controlling diet and improving eating habits [222] (e.g. to treat obesity or support diabetes monitoring).

Taking it into account, the next goal of our research was to evaluate proposed DL detection and thermal image enhancement models in possible non-contact vital signs monitoring applications. Specifically, we focus on estimation of Respiratory Rate (RR) from nostril areas by analysing pixel values changes associated with temperature differences during inhalation and exhalation. Our previous studies of this problem were utilizing original resolution data and were making use of manually marked facial regions [172, 151]. Here, we want to determine whether there is a relation between PSNR (achieved by the proposed super resolution DRESNet model and other analysed state-of-the-art networks) and accuracy of detecting facial regions that could be used for extraction of

breathing patterns, e.g. nose area. Utilization of accurate object detection algorithm may allow for implementing a fully automated remote diagnostic solution. Another application considered by us focuses on non-contact extraction of vital signs. As presented in our research, we are particularly interested whether thermal image resolution increased with Convolutional Neural Network (CNN) could lead to more accurate extraction of a breathing rate. Conducted experiments include analysis of different thermal datasets, scaling factors and respiratory rate estimators. Results achieved with the introduced novel Deep Neural Network (DNN) architecture are compared against Eulerian Video Magnification [223], color and motion magnification algorithm, applied in literature for amplifying heart rate and respiration patterns [224] Since we want to target applications of vital signs collection during daily activities, such as reading a book, working at the computer, driving a car, person identification may be essential for tracking changes in recorded patterns. Face recognition is one of the most common computer vision approaches applieed to solve this problem. Taking it into account, one of our studies [216] was aimed at evaluating whether DNN used for face recognition will produce more accurate predictions if thermal sequences are first enhanced with the proposed model. Finally, we also investigate possibility of extracting emotions from estimated vital signs, as such information may be valuable for monitoring psychological status of people suffering from neurological and psychological disorders or working under a stress [219].

6.2 Related Work in Thermal Domain

In this section we focus on providing a brief overview of existing image processing-based solutions for non-contact estimation of RR. We also present some studies conducted on tracking of facial areas used for respiratory signal extraction.

Non-contact estimation of respiratory patterns has a huge potential in various medically-oriented applications, e.g. stress analysis of autonomous vehicles passengers [225], remote diagnostics with smart home platforms [226] or drones in areas with a limited walkable/driveable access [227], vital signs analysis that could be applied to security applications, e.g. during border control [228], infants monitoring during hospital visits, especially important for premature babies [229], and many others.

Naturally, such solutions may strongly benefit from image processing techniques and as a result limit the use of additional sensors, as compact cameras are sufficient for most of them. Moreover, cameras could be potentially integrated into existing devices or home and other environments infrastructure, allowing for monitoring of people in a non-disturbing way. Processing of thermal image sequences for the needs of RR estimation has been already widely studied. Dilation-induced changes of pixel values around the nose tip and associated with them breathing cyclic information were also utilized for mental stress detection by Cho Y. et al. [230]. Sleep studies made use of thermal-based airflow changes extracted in a non-contact way, showing a high detection accuracy of apnea from recorded breathing activity [231]. Fei J. et al. [232] studied the influence of applying a narrow bandpass filter to the proposed respiratory rate imaging system and proved its suitability for an unobtrusive, desktop monitoring solution. Another photoplethysmography method was based on a statistical algorithm and quasi-periodicity phenomenon of breathing signal (lower and higher pixel values in data distribution corresponding to colder (inhalation) and warmer (exhalation) temperature of air in the nostril area during respiration [233]. Similar approach was introduced by Fei J. and Pavidis I. [234], where breathing signals were calculated from the nostril area, which was detected and tracked over time with the probabilistic models.

Later, Ruminski J. [235] proposed and evaluated a method for respiration rate and respiration pattern extraction from thermal sequences recorded with a portable thermal camera. The presented study utilized the same sensor as was used for collecting datasets presented in this work (Chapter 3, FLIR® Lepton camera) due to its relatively low cost and small spatial size which makes it possible to embed it in wearable platforms, such as developed by us eGlasses device [150]. Achieved results confirmed possibility to obtain reliable breathing signals from low resolution thermal sequences, however, as mentioned by the author, some motion artifacts may significantly affect results. Therefore, automatic detection of nostril region is important for eliminating such noise.

In fact, image processing-based vital signs estimation is typically a multi step procedure, where at first a Region of Interest (RoI), from which the signal is extracted, is either detected automatically of marked manually. Since breathing is associated with exchange of air, a commonly used RoI is defined at a nostril or a mouth area. Various techniques for tracking of RoI in thermal infrared imaging have been already proposed in literature. Zhou Y. et al. introduced an algorithm based on combination of a particle filter with probabilistic templates, producing a tracker which is insensitive to positional and physiological changes [236]. The tracking of nose area proposed in [237] was based on human face physiology and selection of salient features using temperature information. Other methods which adopt assumptions concerning facial geometry were also evaluated [26, 27]

After RoI selection, pixel values within the detected region are aggregated to produce single value, representing a given frame. Collections of those values over time form a signal which corresponds to color changes and thereby local temperature changes. Filtering and processing of the constructed signal gives us information about respiratory rate. There are different methods of determining the main frequency of such periodical signals, referred to as RR estimators. They can be based on estimating the dominating peak frequency, number of zero-crossings or number of signal peaks. More details about frequency estimators analysed in this work are provided in RR estimation methodology subsection (Sec. 6.3.2).

Studies introduced in [235] were continued by us in further research [172], where we compared different respiratory rate estimators and analysed machine learning methods for facial feature tracking. Haar-cascade and interest point detection-based algorithm, originally proposed by us in [58], were analysed in various scenarios (considering person's movement) in order to determine its applicability to non-contact RR estimation. Although the displacement of the detected area from its ground-truth position was satisfactory for Harris and SIFT detector, the produced error was dependant on different poses of the subject. Also, as mentioned by Marzec M. et al. [26] other disturbances (haircut, background objects) may also influence detection accuracy.

According to in-depth analysis of various detection algorithms presented in Chapter 2 this finding holds true for most of methods based on hand-crafted features. This is one of the reasons for recent advances in DL, which does not require pre-defined representations of objects, but instead is able to automatically extract them from complex background, at different angles and lighting conditions [39]. Thus, we are mainly interested in DL-based algorithms for detection of facial regions useful for RR estimation. It has been already proved that region from which vital signs are extracted is crucial for the resulting accuracy [172]. Also, according to previous studies, motion magnification techniques, e.g. Eulerian Video Magnification [223], lead to the increased performance of heart rate evaluation even at a very long distances from camera ($> 6m$) [169]. Yet, to the best of our knowledge, evaluation of super resolution algorithm in the context of improving accuracy of facial feature detection, especially in thermal imaging, hasn't been performed yet.

In addition, we go one step further and use deep enhancing models for improving results achieved

by deep object detection models. Specifically, our contribution lies in evaluation of existing and proposed by us Super Resolution (SR) Convolutional Neural Network (CNN) for improvement of state-of-the-art DNN aimed at detecting facial areas and computer vision algorithms used for extracting respiratory patterns.

6.3 Practical Applications of the Proposed Methods

In this section an in-depth evaluation of the DL model proposed for enhancing thermal data for practical remote medical diagnostic applications is performed. Achieved results are compared against accuracy produced with the use of previously presented in literature techniques aimed at performing similar tasks, i.e. super resolution and deblurring of thermal sequences, as well as magnification of vital signs patterns.

6.3.1 Facial Areas Detection

Objective

Temperature changes and resulting pixel values variations corresponding to vital signal are particularly visible in specific facial areas, e.g. breathing is associated with changes of inhaling/exhaling air temperature. Hence, as previously mentioned, RoI detection is usually a first step in non-contact vital signs extraction. This task may be challenging in thermal imaging because of features blurring and relatively low spatial resolution of available image acquisition sensors. According to our previous work [141, 167], even with the use of Deep Neural Networks, which have been proved to achieve human-like accuracy in computer vision tasks, the achieved accuracy of facial regions recognition was limited (Intersection over Union (IoU) of 0.32±0.38, 0.55±0.42 for eyes and nostril areas respectively). We believe that a possible reason for such results are similarity of different facial regions due to their blurring, as previously presented in Fig. 4.18. Therefore, the improvement of detection accuracy can be potentially achieved by improving resolution of collected sequences. In our studies [208], we evaluated the influence of thermal data enhancement (and corresponding image quality metrics: PSNR and SSIM) on the accuracy of facial feature detection with SSD - a Deep Neural Network, successfully used for similar tasks but in the visible light domain. The architecture of SSD object detection model was previously explained in Section 2.3.2.

Methodology

In previous chapter (Chapter 5) we explained methods proposed by us for image enhancement and evaluated them using various datasets in the audited thermal domain. At this step of our research we utilize generated by us super-resolved data to verify if increased resolution helps with producing better accuracy of facial areas detection. Experiments conducted on thermal datasets (SC3000-ADRA, Lepton-ADRA, IRIS), using the proposed SR network (DRESNet) and other state-of-the-art DNNs resulted in generation of 33 thermal data models. Specifically, following models were produced (nomenclature from Sec.5.3.4):

A) Datasets: SC3000-ADRA-8/16-test-S2-W1/W7/W30/W90; models: DRESNet, DRCN, DRRN

B) Datasets: Lepton-ADRA-8/16-test-S2/S4; models: DRESNet

C) Datasets: IRIS-test-S2; models: DRESNEt, SRCNN, DRCN, DRRN, p2p-deblur.

Detection model applied in this study requires object coordinates for training. Utilized thermal databases are missing such annotations, thus, there was a need for creating them. Since manual data annotation is a time-consuming process and we wanted to perform experiments to verify the assumed correlations between features resolution and accuracy of their detection, we selected two databases as representatives for initial tests: SC3000-ADRA collected by us and publicly available IRIS. Lepton-ADRA datasets was skipped as it was used only for the proposed model training to verify the repeatability of achieved results on different dataset than in the original research where DRESNet was proposed by us[208]. To prepare data for facial feature detection training, test subsets of both databases were at first enhanced with the trained SR and deblurring models, producing hallucinated images of a face (e.g. SC3000-ADRA-8-test-S2-W1-DRESNet is a set produced by enhancing data from SC3000 sensor with 8 bit-width, test subset of ADRA dataset, scale of 2, temporal averaging window of 1, using DRESNet model). Evaluation of produced results was conducted by using PSNR and SSIM image quality metrics in order to compare them later with values of IoU of detected facial regions.

Bounding boxes representing eye and nose regions were marked manually in original HR data and applied to LR and SR frames. It's important to preserve the same locations of annotations to perform a fair comparison, as in the case of annotating LR data separately by an expert, lack of facial features may be unwittingly taken into account and resolution degradation may be compensated by providing different coordinates of facial areas. The same may happen for super-resolved data, which may provide more accurate facial feature representations leading to more exact annotations. To avoid being biased by those issues and properly determine if resolution degradation has influence on feature representation, the same annotations were used for both enhanced and degraded frames. It was possible since loss of resolution was simulated by downscaling and upscaling of original HR data and produced LR images had the same size as hallucinated ones. Annotated sets were again divided into train, validation and test parts using a split of 70:15:15 to optimize object detection model. Each set from SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90}-{DRESNet/DRCN/DRRN} and IRIS-{test}-{S2}-{DRESNet/SRCNN/DRCN/DRRN/p2p}, was used to train a separate detection model in order to analyse relation between achieved image quality metrics and accuracy of facial feature localization.

Super-resolved frames were produced from test subsets of each database in order to measure a true performance of SR networks. Yet, as a result, produced enhanced sets had limited number of samples (test set consists of ˜194 images that were further divided for object detection optimization, leaving only ˜136 images for training). Due to this limitation, learning models from scratch is impossible and commonly used method is to apply transfer learning, as proposed in our previous work [141, 141], explained in Chapter 4. For this experiment, we decided to follow the same approach and tune model previously trained on bigger dataset to super-resolved sets generated with DL networks analysed by us. For transferring knowledge about features representation, a publicly available checkpoint [1] of Inception-based SSD300 detection model [138] trained on a huge amount of visible light images from COCO dataset [238] was used.

Selection of SSD network was motivated by its high accuracy comparing to other DL-based detectors, e.g. mAP of SSD300 fine-tuned on VOC2007 and VOC2012 datasets (07+12) is 79.6, while for Fast R-CNN it's 70 [239]. Although results are on pair with faster R-CNN (mAP of 78.8), we decided to use SSD due to its relatively simpler structure comparing to other detection

[1] https://github.com/tensorflow/models/blob/master/research/object$_detection/g3doc/detection_model_zoo.mdAccessed$: $2018-11-01$

networks. SSD does not need to generate object proposals at a run time, since bounding boxes of different aspect ratios and scales are produced at training step. During the inference, presence of each object within boxes is determined by assigning scores. After that, coordinates of boxes are adjusted using non-maxima suppression algorithm. This structure makes SSD suitable for real-time applications, what is very important for our target solutions where we want to track vital signs of persons during daily activities even at larger distance assuming the subject may be in motion [169]. Details about comparison of object detection architectures were provided in Section 2.2.2.

At first, a random search technique was applied using original HR data to find the best hyperparameters for SSD model training. Then, the same configuration was applied to all object detectors optimized on LR sets (SC3000-ADRA-{8/16}-{test}-{S2}-LR, IRIS-{test}-{S2}-LR) and hallucinated sets (SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90}-{DRESNet/DRCN/DRRN} and IRIS-{test}-{S2}-{DRESNet/SRCNN/DRCN/DRRN/p2p}) produced by enhancing corresponding LR data. Final set of hyperparameters consists of 40k training steps using batch size 32, initial value of learning rate 4.00e−3 and is decreased by a factor of 5.00e−2 every 5k steps.

Once all detection models were trained, images from test parts of sets were feed into the models, predicting locations of eye and nose regions. Achieved results were evaluated by calculating IoU values between all output bounding boxes and their ground-truth locations. Achieved outcomes are presented in the following subsection.

Results and Discussion

Figure 6.1 presents relation between IoU metric (average for all detected regions) and () image quality metric (average for all ground-truth areas) for all averaging window sizes (i.e. data produced with models: SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90}-{DRESNet/DRCN/DRRN}). Facial areas were detected using SSD models from sequences with both enhanced and decreased resolution. PSNR was calculated for all super-resolved frames and LR images (downscaled and then upscaled using bicubic interpolation). IoU for detected facial regions for all sets from SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90} are collected in Table 6.1.

Table 6.1. IoU for detected facial regions for all SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90} sets; (red - first best, blue - second best for each region, separately for each averaging window size)

	SC3000-ADRA-8-test-S2-W1					SC3000-ADRA-16-test-S2-W1				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.90 ±0.03	0.79 ±0.12	0.91 ±0.02	0.90 ±0.04	0.91 ±0.03	0.91 ±0.03	0.85 ±0.08	0.91 ±0.03	0.90 ±0.04	0.91 ±0.04
face	0.84 ±0.06	0.33 ±0.38	0.83 ±0.05	0.83 ±0.05	0.84 ±0.06	0.80 ±0.20	0.62 ±0.29	0.83 ±0.07	0.83 ±0.61	0.84 ±0.06
nose	0.83 ±0.06	0.31 ±0.38	0.84 ±0.06	0.83 ±0.07	0.85 ±0.08	0.85 ±0.08	0.59 ±0.35	0.84 ±0.08	0.85 ±0.07	0.86 ±0.07
avg.	0.86	0.48	0.86	0.85	0.87	0.85	0.69	0.86	0.86	0.87
	SC3000-ADRA-8-test-S2-W7					SC3000-ADRA-16-test-S2-W7				

region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.95 ±0.03	0.57 ±0.34	0.95 ±0.02	0.95 ±0.03	0.95 ±0.02	0.95 ±0.02	0.88 ±0.09	0.94 ±0.03	0.95 ±0.02	0.95 ±0.02
face	0.82 ±0.08	0.32 ±0.37	0.81 ±0.08	0.81 ±0.08	0.83 ±0.06	0.82 ±0.06	0.48 ±0.39	0.75 ±0.27	0.81 ±0.8	0.85 ±0.07
nose	0.86 ±0.05	0.45 ±0.39	0.87 ±0.06	0.86 ±0.06	0.88 ±0.06	0.86 ±0.05	0.65 ±0.25	0.86 ±0.05	0.86 ±0.6	0.87 ±0.05
avg.	0.88	0.45	0.88	0.87	0.88	0.88	0.67	0.86	0.85	0.89
	SC3000-ADRA-8-test-S2-W30					SC3000-ADRA-16-test-S2-W30				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.95 ±0.02	0.83 ±0.10	0.94 ±0.02	0.95 ±0.03	0.94 ±0.03	0.94 ±0.03	0.89 ±0.05	0.94 ±0.03	0.94 ±0.03	0.94 ±0.04
face	0.80 ±0.09	0.45 ±0.33	0.80 ±0.10	0.80 ±0.08	0.81 ±0.09	0.81 ±0.09	0.48 ±0.34	0.82 ±0.07	0.80 ±0.80	0.81 ±0.08
nose	0.80 ±0.08	0.70 ±0.21	0.81 ±0.07	0.81 ±0.10	0.80 ±0.09	0.82 ±0.07	0.62 ±0.32	0.82 ±0.09	0.80 ±0.10	0.81 ±0.08
avg.	0.86	0.66	0.85	0.85	0.85	0.86	0.66	0.86	0.84	0.85
	SC3000-ADRA-8-test-S2-W90					SC3000-ADRA-16-test-S2-W90				
region	orig.	bicub.	DRCN	DRRN	DRES	orig.	bicub.	DRCN	DRRN	DRES
eye	0.88 ±0.04	0.76 ±0.11	0.89 ±0.05	0.89 ±0.04	0.89 ±0.04	0.89 ±0.04	0.84 ±0.09	0.90 ±0.05	0.88 ±0.05	0.89 ±0.04
face	0.78 ±0.20	0.22 ±0.36	0.78 ±0.19	0.77 ±0.20	0.78 ±0.20	0.78 ±0.20	0.44 ±0.39	0.77 ±0.19	0.72 ±0.26	0.83 ±0.07
nose	0.84 ±0.10	0.34 ±0.39	0.81 ±0.07	0.84 ±0.07	0.83 ±0.09	0.82 ±0.09	0.79 ±0.06	0.83 ±0.07	0.83 ±0.08	0.81 ±0.08
avg.	0.83	0.44	0.84	0.83	0.84	0.83	0.69	0.83	0.81	0.84

IoU for eye and nostril classes detected with SSD model from LR data (generated with bicubic interpolation, scale 2) and then further enhanced with evaluated state-of-the-art networks and the proposed thermal SR model (DRESNet), as well as with deblurring p2p GAN are collected in Table 6.2. The presented results were produced for the reference IRIS thermal dataset IRIS-test-S2. Qualitative of applying SSD model on the same image from IRIS-test-S2 enhanced with different SR methods are presented in Fig. 6.2.

Although experiments described in the previous chapter confirmed that the introduced super-resolution model outperforms other DL networks in a task of thermal image enhancement, evaluation whether better image quality metrics have influence on accuracy of facial areas detection remains a very important question in the research on remote medical diagnostic solutions. This section focuses on making assessment of this relation.

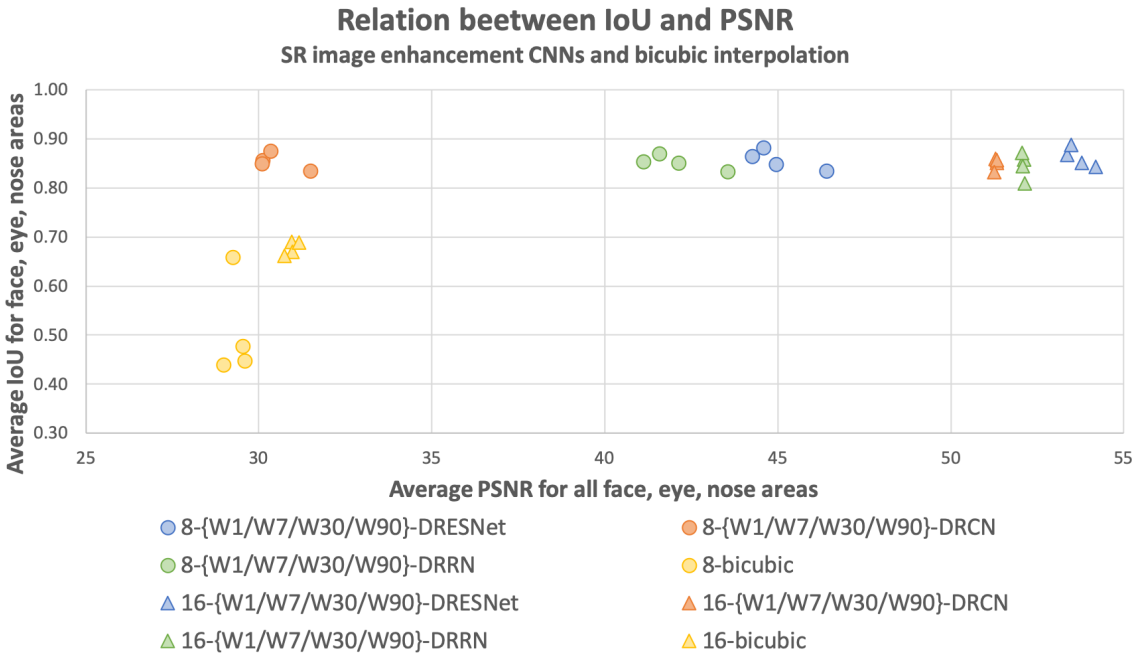


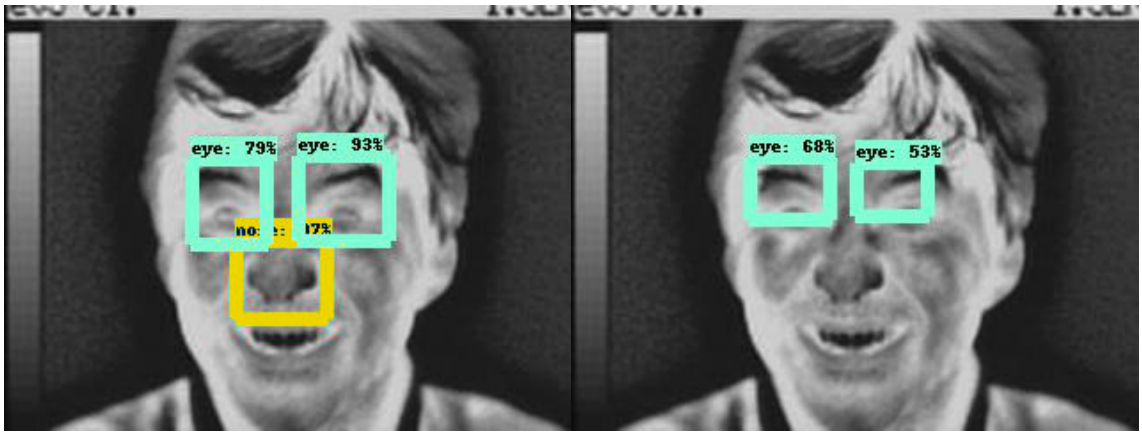
Figure 6.1. Relation between IoU metric (average for all detected regions) and PSNR image quality metric (average for all ground-truth areas) for all averaging window sizes on SC3000-ADRA-{8/16}-{test}-{S2}-{W1/W7/W30/W90}-{DRESNet/DRCN/DRRN} sets; for simplicity set names are shortened to changing parameters only, i.e. {8/16}-{W1/W7/W30/W90}-{DRESNet/DRCN/DRRN}

Table 6.2. IoU for eye and nostril classes detected with SSD model from LR and enhanced thermal data (red - first best, blue - second best for each region separately)

SSD trained on:	eye	nose
IRIS-test-S2-bicubic	0.67 ± 0.22	0.56 ± 0.32
IRIS-test-S2-DRESNet	0.72 ± 0.18	0.69 ± 0.21
IRIS-test-S2-DRCN	0.61 ± 0.20	0.60 ± 0.18
IRIS-test-S2-DRRN	0.68 ± 0.25	0.66 ± 0.33
IRIS-test-S2-SRCNN	0.59 ± 0.28	0.64 ± 0.22
IRIS-test-S2-p2p	0.68 ± 0.21	0.69 ± 0.16

Based on IoU values calculated for SC3000-ADRA (Table 6.1) and reference thermal set IRIS (Table 6.2) we can conclude that image enhancement is crucial for improving accuracy of facial areas detection. All applied super-resolution models allowed for increasing IoU by at least 15% comparing to LR bicubic images. In the best case (SC3000-ADRA-8-test-S2-W90,), the accuracy of face area detection after enhancing image with DRESNet in terms of IoU was improved by 0.56 comparing to corresponding low resolution image. The improvement was also very significant for other facial regions, e.g. nostril detection improved by 0.49 and eyes area improved by 0.13. This finding supports the second part of thesis II which states that increased resolution of thermal images lead to improvement of facial areas detection accuracy.

For the reference IRIS set (Table 6.2) we also observe improvement of detection accuracy, but



(a) Original HR data

(b) Enhanced with SRCNN

(c) Enhanced with p2p-deblur

(d) Enhanced with DRCN

(e) Enhanced with DRRN

(f) Enhanced with DRESNet

Figure 6.2. Facial regions detected with SSD model in the same image (IRIS) enhanced with different SR methods. Detected categories were marked as follows: eye - light green, nose - yellow.

the difference between LR and super-resolved images is smaller. The possible reason for this result is that IRIS set was collected in more dynamic conditions where motion content was higher. Thus, images might already have worse resolution leading to worse performance overall. IoU results are

also on pair with image quality metrics calculated for different image enhancement models in the previous chapter (Table 5.1 and Table 5.3). The proposed DRESNet model led to the best image quality metrics and data enhanced with this model resulted in the best detection accuracy results (for SC3000-ADRA-16-test-S2-W7 average IoU of 0.89; on IRIS set the best IoU was produced for eye area - 0.72).

However, other evaluated SR DL models achieved very close IoU results to those produced by the introduced by us thermal data oriented network. As can be seen in Fig. 6.1, after reaching a specific level of PSNR (˜30dB), detection accuracy remains constant regardless of further improvement of image quality. The same conclusion can be made for different sizes of averaging windows and data bit resolutions. Even though better quality was achieved by preserving original data bit resolution and reducing background noise using average of 90 subsequent frames (as presented in Chapter 5), no significant gain of IoU was achieved. This can be caused by the fact that universal learning models are capable of learning complex mapping functions equally well as single ones. For example, let's say the model is optimized for a detection task D, using super-resolved data S(x), where x is an input image, i.e. it learns the mapping defined as D(S(x)). Theoretically, the same model should be able to achieve similar accuracy if provided with enough samples of lower quality data D(x).

Yet, according to IoU results calculated for LR bicubic data, for images with lower quality and thereby lower values of PSNR (in our case below ˜30dB but this can vary between databases), image enhancement is crucial for increasing IoU values. A possible reason for this finding is very smooth representation of facial features in thermal images, which after downscaling are even more blurred. It's worth noting that CNN-based models utilize high frequency features for making predictions. Thus, it might be difficult for the network to correctly adjust region coordinates, if borders of specific areas in downscaled thermal images are completely distorted. This assumption was confirmed by IoU values calculated for low resolution data generated with bicubic interpolation for both SC3000-ADRA (Table 6.1) and IRIS-test-S2 (Table 6.2) sets. In the worst case (8 bit bicubic data from SC3000-ADRA, window size 1 or 7), IoU was below 0.5 (corresponding PSNR values in this case were below 28 dB - see Table 5.1). In this case, inputs had spatial size of 160x120 (320x240 downscaled with a factor of 2). We believe that very low values of PSNR indicate the complete lack of meaningful facial features representations, leading to poor detection accuracy. What's more, scenarios considered by us very often utilize even smaller inputs, e.g. Lepton 2 camera produces inputs of a size 80x60. Taking it into account, there is a need for enhancing such sequences in order to accurately extract regions important for vital signs estimation. On the other hand, for bigger inputs and higher PSNR values, the model is able to learn correct mapping regardless of the small differences in image quality metrics. Thus, after SSD reached its saturation level (Fig. 6.1, SSD maximum IoU of 0.85 [138]), no significant improvement of detection accuracy was observed even with increasing PSNR values.

Results produced by the GAN p2p network aimed at performing deblurring operation also proved its efficiency in improving the detection accuracy. For the nostril area, the IoU metric for images enhanced with the p2p was better than for the DRRN model and the same as for the proposed DRESNet, as presented in Table 6.2. Yet, it is important to note that deblurring mitigates the problem of reversing the convolution, while super-resolved images are the reversed version of down-sampled inputs, so the problems that they solve are exclusive. Potentially they could be applied together to further increase the performance. This experiment will be performed by us in next studies, as explained in Chapter 7. Also, other GAN-based SR models will be evaluated on the collected thermal data.

6.3.2 Respiratory Rate Estimation

Objective

Resolution enhancement is especially desired in healthcare industry, due to possibility of revealing details and components important for making diagnostic decisions that are usually not visible when using lower quality data. Also, it can limit the need of purchasing more expensive acquisition devices or help in cases where better devices are simply not available, e.g. images related to pathological anatomy [240]. Thus, producing the same quality of data as if a higher resolution device was utilized with software-enhanced resolution is one of the gals of this dissertation. Undoubtedly, higher resolution of images can ease their analysis in various medical imaging procedures [241, 183]. Yet, a very interesting research question is whether image enhancement can also help in telemedicine and remote medical diagnostic applications, where usually only standard webcams or low resolution thermal sensors are used.

Some algorithms for magnifying subtle intensity variations corresponding to signals representing vital signs have already been proposed in literature, e.g. Eulerian Video Magnification [223]. According to previous studies [224] it's possible to utilize such algorithms for vital signs enhancement. Yet, there are some problems with those techniques. First of all, they rely on strictly defined image priors or hand-designed features, so magnification may lead to excessive blurring especially when motion is very small. Secondly, they are constrained to a specific frequency spectrum, so unless the frequency is known, they are inapplicable.

Therefore, DL solutions are becoming more popular, as they allow for automatic knowledge extraction and learning of inputs representation. Oh T. et al. [242] proposed to apply Deep Convolutional Neural Network for motion magnification and evaluated its robustness on visible light images. In our study, we want to verify if similar approach would work equally well for thermal data. Specifically, our studies focus on evaluating whether resolution of thermal sequences, acquired in possible remote healthcare scenarios, affects the accuracy of respiratory rate estimation by comparing data enhanced with deep CNNs with original high-resolution and generated low-resolution frames.

Our contribution to the state-of-the-art is threefold: (a) we evaluate whether resolution of thermal sequences enhanced with Deep Neural Networks leads to increased accuracy of non-contact RR extraction in comparison to estimation from data of original resolution (analysis is performed on two thermal datasets to avoid being biased by a specific data distribution); (b) achieved results are compared against Eulerian Video Magnification algorithm previously proved in the literature to be successful for vital sign pattern magnification; (c) extensive benchmark evaluation covers various RR estimators, data aggregation operators, as well as different input data parameters, such as scaling factors and pixel values bitwidth.

Methodology

Since target applications analysed by us focus on non-contact vital signs estimation, only datasets with thermal sequences coollected by us are applicable to our study. Other databases, i.e. IRIS, containing images as single frames could not be adapted due to the lack of temporal information utilized for signal construction. Thus, for this experiment we utilized Lepton-ADRA and SC3000-ADRA databases. Another important aspect of data preparation and pre-processing is the fact that a continuous sequence of thermal frames collected for each person is needed to extract RR and evaluate its to relation PSNR. Since SR models are trained on images from both

datasets, it's important to make sure that samples from training subset are separate from samples used for inference and networks evaluation. This problem is already solved, as in our experiments with different scaling factors (see Section 5.3.4) data were divided into training and test sets based on volunteers' ids (first 15 volunteers used for models training, data of remaining kept aside for evaluation and further experiments). Thus, at this step models already trained for previous experiments were used (i.e. 8 DRESNet models: SC3000-ADRA-8/16-test-S2/S4-DRESNet and Lepton-ADRA-8/16-test-S2/S4-DRESNet).

Accuracy of all those models was evaluated by feeding all LR frames from thermal sequences recorded for remaining subjects, generating restored HR version of them and comparing produced results with original HR data using PSNR and Structural Similarity Index Metric (SSIM). Enhanced frames were then combined back into sequences with the same frame per second (FPS) value and processed to extract RR. Estimated values of the vital signs are compared against the same metrics obtained for original high resolution inputs and low resolution recordings generated with bicubic interpolation.

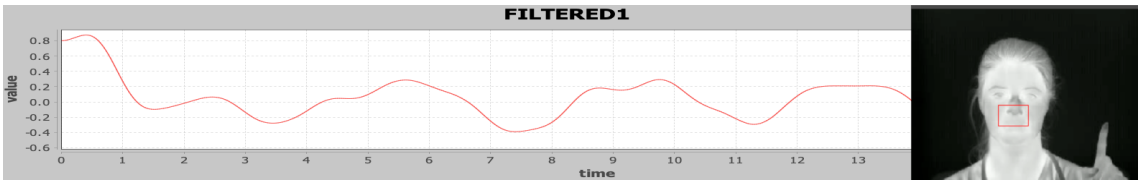
Non-contact estimation of basic vital signs has revolutionized conventional medical procedures, which involve the use of electrodes placed on a body. As already mentioned, previous studies showed that RR can be accurately obtained from very low resolution (80x60) sequences [235]. In our work we were mainly interested whether estimation accuracy can be further improved if vital signs are estimated from sequences super-resolved with the means of Deep Neural Networks. Two respiratory rate estimators: eRR_{sp} and eRR_{as}, previously verified by us in [172], are analysed in conducted experiments, also described by us in [213].

Estimator eRR_{sp} assumes that a signal representing respiratory activity dominates in the signal spectrum and thereby a RR value can be estimated by obtaining the frequency value of the dominating peak. Yet, this may result in unreliable estimation, as even for signals other than vital sings, such as noise, a maximum value in the frequency domain can be retrieved leading to false outcomes. In the second estimator, referred to as eRR_{as} due to the use of auto-correlation spectrum, a relation between a periodic signal and its auto-correlation sequence is utilized. This relation is based on the fact that the auto-correlation sequence has the same cyclic characteristic as its corresponding periodic signal. Thus, analysis of the auto-correlation sequence in different time spans can be used for calculation of RR values.

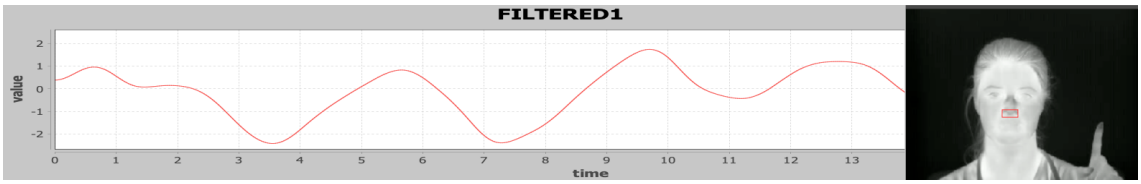
Our work focus on providing and improving solutions of remote medical diagnostics. Possible scenarios considered by us include non-contact estimation of basic vital signs in emergency rooms, principal care doctors' offices, as well as in smart homes for e.g. monitoring of infants, disabled or elderly people. To ensure fast responses and provide convenient solution, the acquisition time shouldn't be very long. Therefore, short data segments (300-400 frames) were utilized for signal extraction. Initial 50-100 samples of each sequence were skipped to reduce possible motion artifacts that usually are present at the beginning of data collection [172].

In order not to introduce additional factor that could influence results, we used RoI marked manually by an expert, instead of using object detection models. Combination of both algorithms in a single RR estimation pipeline will be explored by us in future work. The same regions were applied to all three types of sequences, i.e. original HR data, HR data downscaled and upscaled to simulate LR inputs, produced LR sequences enhanced with the proposed SR CNN. Then, the raw breathing signal was produced by aggregation of intensity values within the marked area. Two different aggregation operations were examined by us: skewness and averaging. Regions utilized for RR estimation were different depending on the applied aggregation operator. For the skewness, it

has been already proved [172] that extracted waveforms don't depend on the specific location and size of the area, as long as a whole nose is covered. On the other hand, averaging of pixel values leads to the smoothing and blurring what has a negative impact on visibility of changes related to respiratory patters. Thus, if averaging operator is used, the region has to be marked more carefully and cover a smaller region, e.g. nostrils only. Differences in region selection taking into account applied aggregation operation are presented in Fig. 6.3.



(a) Skewness aggregation



(b) Average aggregation

Figure 6.3. Selected RoI and extracted raw RR signal. Skewness aggregation is less prone to the specific location of nostril area used for RR extraction, thus selected region can be bigger. For average operator, it has to be more exact.

Values obtained by aggregating pixels values within the marked RoI over a sequence of samples resulted in the construction of a raw signal which was then filtered with a moving average and 4^{th} order Butterworth filters. For the high pass Butterworth filter the threshold frequency equals 0.125Hz. Applied methods were previously verified in our studies [151, 172].

After that step, multiple RR values for each volunteer were obtained, i.e. RR estimated from original HR input, from generated LR data and from super-resolved sequence. Thus, it was possible to compare the influence of resolution enhancement/degradation on the accuracy of RR estimation. Since the RoI remained the same for all sequences (per volunteer) we were able to evaluate whether vital signs estimated with the image processing techniques are sensitive to image blurring and enhancement of facial features. All calculated RR results were compared against the ground truth RR measurement using Root Mean Square Error (RMSE) metric. Reference RR values were obtained with the Vernier respiratory monitor belt (for Lepton data) and with the manual calculation of a number of finger movements performed by a subject corresponding to inhalation/exhalation events (for SC3000 data), as explained in Chapter 3.

In order to evaluate the robustness of the proposed SR-based RR estimation approach, achieved results were compared with values achieved for algorithms previously proposed in literature for magnification of vital signs patterns. Specifically, Eulerian Video Magnification (EVM) [223] was applied to test sequences from both datasets in order to enhance pixel color changes across time associated with respiratory signals. In case of Lepton database images had to be upscaled before feeding them to EVM algorithm, as the EVM tool [1] utilized by us for color changes magnification requires inputs with spatial resolution above 100x100.

[1] https://lambda.qrilab.com, Accessed: September 2018

EVM algorithm is based on spatial decomposition of input sequence into different frequency bands. After that, all bands are processed using temporal filtering and then magnified to reveal changes invisible to a naked eye, e.g. blood flow or vein movements. Magnified signals are added back to the input video to form the final output. Two parameters of EVM are adjustable: filtering frequency range and magnification factor. Based on the fact that a standard respiratory rate of an adult fluctuates between 10-20 breaths per minute (bmp), the filtering frequency was set to 0.16-0.33 Hz. Magnification factor was set to 20, since our previous studies [151, 169] proved high estimation accuracy using this value. Following the chosen filtering frequency (and the standard RR value for an adult), the color change should be observable in every ˜3-second time spans. Fig. 6.4 presents nostril areas extracted from each middle frame in such windows. It can be observed that intensities vary across them due to differences in temperature of inhaled and exhaled air.

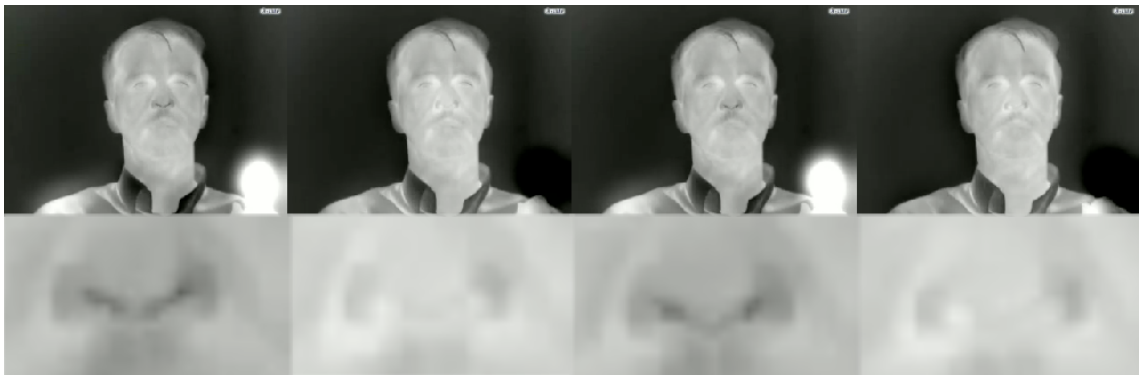


Figure 6.4. Every ˜45^{th} frame (inhalation/exhalation event can be observed every ˜3 seconds, FPS=15) extracted from SC3000-ADRA-8-test sequence magnified with EVM; inhaled air is colder, thus nostrils have darker color; exhaled air is warmer (heated by body), thus nostrils are lighter

Results and Discussion

Figure 6.5 illustrates the same frame from the Lepton-ADRA-8-test set after various modifications, i.e. original, after bicubic interpolation with different scales and then processed using the motion changes magnification algorithm EVM and the proposed thermal SR model DRESNet. Similar results but from SC3000-ADRA-8-test set are presented in Figure 6.6.

Error of respiratory rate estimations between reference value and values obtained with the described image processing algorithms and two respiration rate estimators: *eRR_sp* and *eRR_as* are presented in Table 6.3. Influence of different scaling factors on the estimation accuracy was compared. Also, the proposed thermal image enhancement model was compared with results achieved for original data and sequences with breathing patterns magnified using Eulerian Video Magnification (EVM) [223].

Analysis performed for various datasets in the remote RR estimation study revealed some limitations of utilized methods. Although for SC3000-ADRA-test-S2 data enhanced with the proposed DRESNet model RR was estimated with an error of 2.94 breaths per minute (bpm), for most of considered use cases the RR was much higher (close to 5 bpm), what is not acceptable in professional medical applications. On the other hand, the aim of this study was to investigate whether a use of super-resolved sequences allow for improving accuracy of contactless vital signs estimation, not to outperform existing RR estimators.

(a) Original HR data

(b) Eulerian Video Magnification

(c) Bicubic interpolation scale 2

(d) super resolved using DRESNet scale 2

(e) Bicubic interpolation scale 4

(f) super resolved using DRESNet scale 4

Figure 6.5. The same frame from Lepton-ADRA-8-test set processed with techniques evaluated in the study of respiratory rate evaluation; please note that although bicubic images were completely blurred, the proposed SR model was able to restore facial features from those samples (especially for scale 4, where restoration was done from 20x15 inputs); restored facial features are distorted but allow for determining locations of different facial regions

(a) Original HR data

(b) Eulerian Video Magnification

(c) Bicubic interpolation scale 2

(d) super resolved using DRESNet scale 2

(e) Bicubic interpolation scale 4

(f) super resolved using DRESNet scale 4

Figure 6.6. The same frame from SC3000-ADRA-8-test set processed with techniques evaluated in the study of respiratory rate evaluation; please note that restored features are characterized by more clear boundaries between facial regions

Table 6.3. Root Mean Square Error between reference RR values and RR values estimated from original, LR and enhanced thermal sequences from SC3000ADRA and Lepton-ADRA sets for different scaling factors using two respiration rate estimators: *eRR_sp* and *eRR_as*. Raw pixel values inside manually selected regions of interests aggregated using average (Avg.) or skewness (Skew.) operators (red - first best, blue - second best for Lepton and SC3000 images separately)

Dataset	Method	Bits	RR Estimator Aggregation Operation			
			eRR_sp Avg.	eRR_sp Skew.	eRR_as Avg.	eRR_as Skew.
Lepton-ADRA-test-S0	**orig.**	8 bits	4.97	6.28	15.61	12.80
		16 bits	5.15	6.35	5.68	7.39
	EVM*	8 bits	5.58	7.04	9.40	11.94
		16 bits	5.58	6.81	7.98	11.28
Lepton-ADRA-test-S2	**bicubic**	8 bits	5.66	7.21	9.14	7.96
		16 bits	4.93	7.20	8.08	7.77
	DRESNet	8 bits	4.89	5.64	4.95	6.21
		16 bits	4.93	6.72	6.29	7.41
Lepton-ADRA-test-S4	**bicubic**	8 bits	5.61	7.64	8.57	7.90
		16 bits	6.40	7.32	8.37	7.34
	DRESNet	8 bits	4.96	5.93	7.41	12.25
		16 bits	4.89	6.10	10.31	8.00
SC3000-ADRA-test-S0	**orig.**	8 bits	3.48	3.59	17.19	11.06
		16 bits	3.61	5.61	12.11	14.72
	EVM	8 bits	5.00	6.15	5.98	7.82
		16 bits	4.56	6.09	5.84	7.65
SC3000-ADRA-test-S2	**bicubic**	8 bits	6.35	6.04	17.05	11.26
		16 bits	5.91	8.46	34.43	14.52
	DRESNet	8 bits	2.94	2.46	5.56	4.27
		16 bits	4.09	3.59	8.39	8.90
SC3000-ADRA-test-S4	**bicubic**	8 bits	5.73	8.23	17.05	11.65
		16 bits	5.73	6.32	14.31	11.35
	DRESNet	8 bits	3.48	5.11	13.36	12.12
		16 bits	3.48	5.38	14.48	14.61

* Lepton dataset upscaled to 100 x 100 due to requirements of the EVM tool

According to results presented in Table 4.6 the goal of the presented work was achieved, as for both datasets the smallest RMSE values were produced if thermal sequences were at first super-resolved using DRESNet model. Furthermore, the proposed approach allowed for outperforming not only LR bicubic sequences, but also original high resolution data, e.g. RMSE for inputs as small as 80x15 enhanced with DRESNet (Lepton-ADRA-test-S4) was smaller than for original 80x60 images (Lepton-ADRA-test-S0) - RMSE equals 4.89 vs 5.15 for eRR_{sp} avg. aggregation. Similarly for SC3000 sequences, RMSE was reduced by 0.13 bpm for scaling factor of 4 and 16-bit data and by 0.54 for 8-bit and scale 2. Also, for scaling factor of 2, DRESNet allowed for achieving the lowest values of RMSE regardless of applied RR estimators and aggregation operators. For scaling factor of 4, this conclusion is valid in case of eRR_{sp} estimator, as it has been shown that eRR_{as} estimator is very sensitive to areas selected for signals extraction. As a result, it was very difficult to obtain accurate RR values, what led to inconclusive results. More experiments with this estimator should be performed in future work.

Also, it's worth noting that the number of samples used for RR estimation has a direct influence on the accuracy. Since applications considered by us involve vital signs calculation during physical exam or remote monitoring of subjects during daily activities, the number of utilized samples was limited in order to verify whether short data acquisition time allow for accurate measurement of RR. Utilization of shorter data collection process is also beneficial for limiting possible motion of volunteers. In case of Lepton-ADRA set (sampling frequency f_s=12) we utilized 300 samples ($N_{samples}$) from acquired sequences, while for SC3000-ADRA database (f_s=30) 400 samples were used. Therefore, the frequency resolution defined as:

$$\delta f = \frac{fs}{N_{samples}} * 60 \tag{6.1}$$

equals 2.4bpm and 4.5bpm for Lepton-ADRA and SC3000-ADRA, respectively. Increasing number of samples at the same sampling frequency will lead to decreasing of δf and thus more accurate measurements. On the other hand, we want to utilize small measurement windows (small N) to be able to provide information about vital signs without delays caused by long data acquisition process. Moreover, the presented study was based on manual selection of RoI used for signal extraction. However, as shown in our previous studies on extraction of average pixel values from areas detected with DL networks, more accurate signals can be obtained for dynamic areas locations in case if their spatial size is relatively small, e.g. for nostril area (see Section 4.2.2). Taking it into account, in future work we would like to combine the proposed RR estimation method with facial areas detection using Deep Neural Network, described in the previous section (Sec. 6.3.1). Additionally, we would like to propose and evaluate networks that would allow for detecting areas, where vital signs are the most accurate instead of using pre-defined facial regions, such as nostrils. Some preliminary work in this area has been already proposed by us [**?**], however we would like to further investigate this idea.

Another limitation of conducted experiments are strictly defined measurement conditions that might be difficult to achieve in real-life scenarios. The best results are achieved if nostrils are clearly visible in a frame [172] showing differences of temperature between inhaled and exhaled air used for RR estimation. Even though we assumed simple data collection process (i.e. person looking towards camera no necessarily tilting head backward and no additional data pre-processing algorithms, such as motion/lighting compensation), it would be useful to perform similar experiments for sequences that contain possible higher motion content and better simulate possible scenarios of remote medical diagnostics.

Similarly to research conducted for facial areas detection from thermal image sequences (Section 6.3.1), utilization of higher bit resolution data turned out to be beneficial for RR estimation accuracy for both LR bicubic images and sequences enhanced with EVM. This finding further supports the need for preserving original raw formats of acquired data instead of using lossy conversion to 8-bit image formats. For analysed SR models the positive influence of 16-bit format on RR estimation accuracy is not that clear. We believe that it may be caused by the fact that both 8-bit and 16-bit models were trained using the same hyperparameters and training termination procedure, while 16-bit images were more detailed. As a result, 16-bit Deep Neural Networks might have overfitted to the training part of utilized databases, resulting in worse generalization capabilities. This problem will be also investigated by us in future work.

Comparison with state-of-the-art vital signs patterns magnification algorithms showed the robustness of the proposed SR model over them. Specifically, 8 and 16-bit sequences downscaled 2-times and then enhanced with DRESNet resulted in smaller RMSE values for both eRR_{sp} and eRR_{as} estimators. Furthermore, for most of super-resolved inputs 4 times smaller than sequences magnified with EVM, RR estimation accuracy was also better. Taking into account the size of those downscaled images (20x15), we believe that the proposed method has a huge potential in remote person monitoring solutions allowing for initial estimation of health status at e.g. long distances, where interesting RoI occupy only a small part of a frame [169] or using very small thermal sensors embedded e.g. in wearable devices, such as developed by us smart glasses [150]. Also, improvement of thermal image resolution using image processing techniques can enable various innovative remote medical solutions, which were previously difficult to achieve due to e.g. higher cost, lower availability and bigger sizes of suitable thermal cameras, e.g. driver's drowsiness detection [243].

6.4 Other Relevant Applications

Here, we focus on evaluating other relevant applications of the proposed model which could be integrated into potential remote diagnostics solutions. At first, we perform experiments with DL-based face recognition models from thermal data previously enhanced with the introduced SR network. In addition, other application that could possibly benefit from the proposed resolution enhancement model and transferring the knowledge from other image domains is described. Specifically, we describe how to recognize emotions from extracted in a non-contact way vital signs. Preliminary work in this area has been already proposed by us and described in [158]. In this Section, we will specify methodology applied by us to perform emotion recognition from vital signs and provide ideas for improvement of the proposed method with the means of Deep Neural Networks.

6.4.1 Face Recognition

Objective

In order to make remote medical diagnostic solutions more intelligent, context information about subjects should be retrieved. Possible ways of obtaining this information include the use of graphical markers, as proposed by us in our previous study [153], as well as direct processing of recorded video sequences to recognize objects, persons or performed actions. However, a few important concerns are associated with the use of visible light data. First of all, algorithms are usually sensitive to changing lighting conditions , especially when traditional machine learning

algorithms are used, which require definition of specific feature sets characterized by different representation depending on the lighting [68]. Some solutions have been proposed to address this problem. Kalaiselvi P. and Nithya S. [244] suggested to define features that are insensitive to different illumination conditions, Huang F. and Bian H. [245] introduced illuminance-invariant face recognition system by using contrast equalization and Gamma correction. It has been also shown that utilization of different color spaces, e.g. HSV may reduce the influence of poor lighting condition on authentication accuracy [246].

This, on the other hand, lead to an increased computational complexity of the whole pipeline, what may impact performance and responsiveness to incoming events , since remote diagnostics solutions are usually resource constraint due to the target deployment platforms, e.g. Internet of Things (IoT) devices, wearable solutions, or Systems on a Chip (SoC). Thus, more studies are focused on DL models optimization. We also showed how to quantize state-of-the-art classification models in order to reduce processing time and model size without impacting accuracy [247].

Yet, some problems with the use of visible light data, such as privacy and data security concerns remain unsolved. Those issues are especially important in medical applications due to a high sensitivity of collected data and possibility of revealing private information if diagnostic solutions are an integrated part of a smart home infrastructure, e.g. used for fall detection in bathroom. In such cases, thermography is often preferred due to the way of how images are constructed, i.e. high level features are more blurred and it's difficult to identify a person even with a naked eye as objects are represented by temperature distributions.

On the other hand, recent advances in Deep Neural Networks (DNNs) allowed for achieving human-like accuracy in various computer vision tasks [39]. Taking it into account, it's important to verify if a person can be recognized from thermal images if DL-based models are used. Moreover, we also evaluate if resolution enhancement/degradation has influence on the accuracy of person recognition by utilization of thermal image sequences (from both Lepton and SC3000 datasets) super-resolved with the proposed DRESNet model. As far as we are concerned, our work is a first attempt to evaluate effect of increased thermal image resolution on robustness and accuracy of person recognition.

Facial embeddings used for subject identification are extracted with Convolutional Neural Network from all three types of inputs, i.e. downscaled with bicubic interpolation, original HR frames and LR data enhanced with our SR model. In this way, we are able to evaluate influence of image resolution on biometrics representation generated with DL techniques. Results of this study were presented by us in [216].

Methodology

The first step in person recognition pipeline requires cropped facial areas that could be used for building representation of each person. To detect faces, we followed similar approach as in our previous experiments (Sec. 6.3.1) and utilized SSD DL model. Transfer learning technique proposed in Chapter 4 for improvement of accuracy was used to re-purpose model previously trained on visible light data to thermal datasets. For this experiment only 8-bit width data from SC3000 was used, preserving every 180^{th} frame to ensure high variability of facial representations. As a result, subset of SC3000-ADRA contained of 766 images grouped into 40 categories corresponding to volunteers' for whom data was obtained. This set will be thereafter referred as SC3000-FR (Face Recognition). As a second, reference database we decided to utilize IRIS set (4190 images divided into 30 subjects) in order to make sure that results are not biased towards one set and to provide

comparable metrics, since IRIS is publicly available. This set is referred to as IRIS-FR. All images were manually annotated with bounding boxes corresponding to location of facial regions.

A separate SSD model was trained for each dataset using hyperparameters configuration verified by us in our previous work [167, 208] using 70% of images from each set for training, and 15% for validation and testing. Trained networks were evaluated on test subsets using IoU metric, representing ratio of similarity between manually marked ground-truth area and detected face coordinates. Results showed that model trained on SC3000 set is more accurate (84.1 $\pm$ 6% vs 79.4 $\pm$ 14% on IRIS). Thus, it was used to crop faces from both sets. From all extracted regions we randomly selected 20% to generate profiles of each volunteer using face recognition model, the remaining part was preserved for evaluating accuracy of the proposed pipeline.

The convolutional-based FaceNet model [248] was applied to build personal profiles. The output of this model is represented as a vector containing 512 features that uniquely characterize each person from the dataset. A larger embedding vector was selected, because it has been showed that it allows for capturing more subtle facial differences and thus leads to higher recognition accuracy. A checkpoint of the FaceNet model, previously trained on data from visible light spectrum was directly applied to thermal data without additional re-training. In this way, we wanted to determine whether model learnt to extract high frequency features present in RGB images will produce representations of thermal images sufficient for person recognition task.

The final representation of each person was build by calculating average of vectors produced for all profile images of this subject (20% of all frames) and saved in a database for future reference. Every time a new frame is collected by a system, it is fed to the same model for extraction of current embedding that is then compared against all stored users' profiles in order to perform person identification. Comparison of feature vectors (current one vs. all stored profile) was done with two methods: Support Vector Machine (SVM) with linear kernel and Euclidean Distance between two vectors. The selection of linear kernel was motivated by the fact that linear SVM is faster and performs very well if a number of samples is relatively small, while a number of feature within each sample is larger. In this case, mapping to a higher dimensional space is not necessary, as performance would remain constant.

To evaluate influence of resolution degradation on the proposed method, additional sets were created from the constructed SC3000-FR and IRIS-FR sets. At first images were downscaled and upscaled with a scale 2 to produce LR versions of original data. Then, DRESNet models with configuration and hyperparameters verified by us in previous experiments were optimized on those sets to learn kernels which would allow for restoration of facial features. Trained models were used to generate super-resolved facial images. All created sets, i.e. LR images: SC3000-FR-S2, IRIS-FR-S2 and corresponding restored HR outputs: SC3000-FR-DRESNet, IRIS-FR-DRESNet were processed to build users' profiles and evaluate face recognition pipeline in the same way as described above for original HR data. Results produced for all subsets were compared in order to determine if different resolution of thermal images affects possibility of identifying subjects.

Results and Discussion

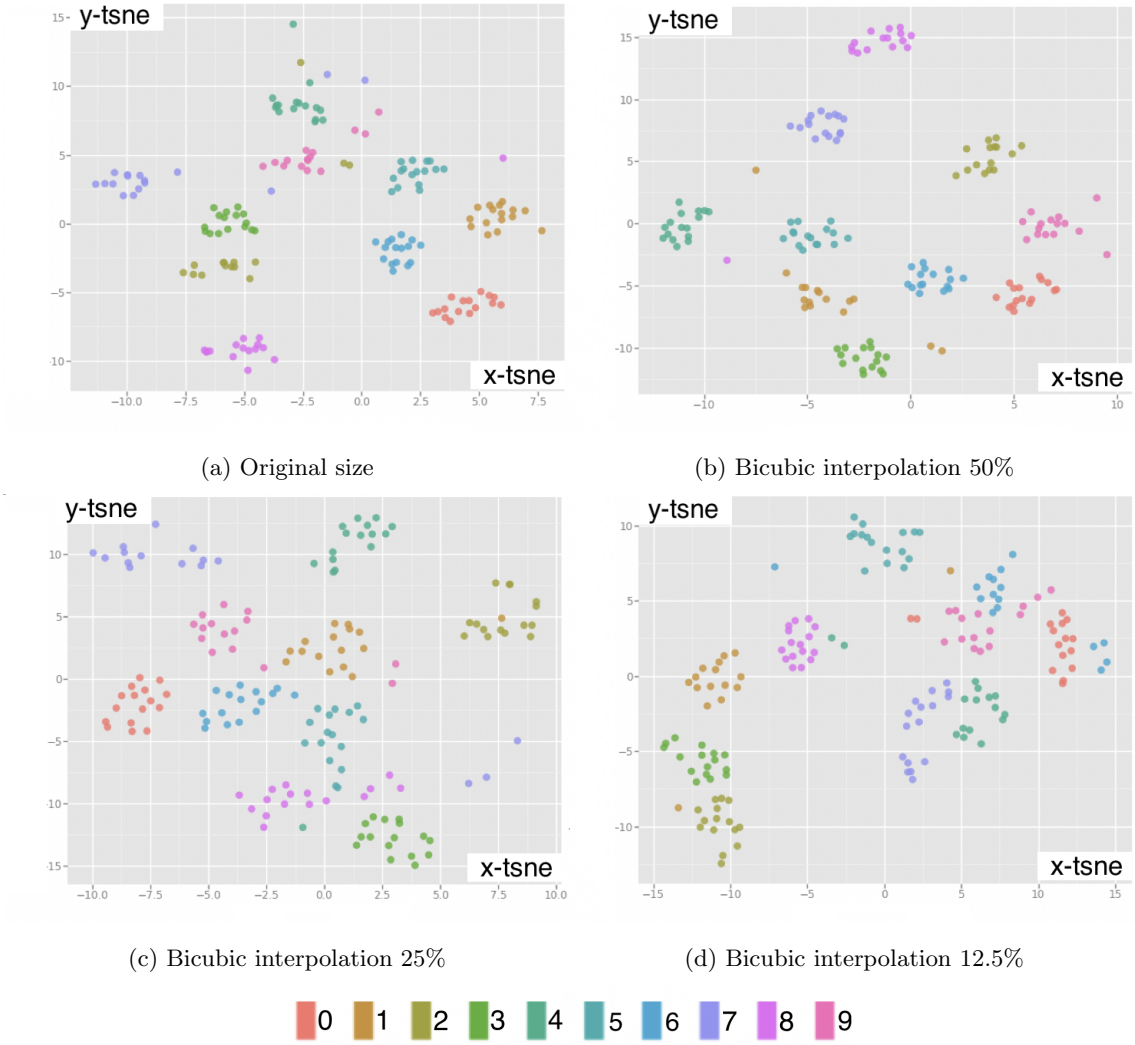
Accuracy of person recognition from generated set of images is presented in Table 6.4. For testing 80% of images from each subset was used. The rest 20% was utilized to create users' profiles that new images are compared against during verification process. We also collected accuracy for images downscaled 4 times to evaluate influence of resolution degradation on person identification.

Figure 6.7 presents embedding vectors produced with Face Recognition Deep Neural Network for images from SC3000-FR (first 10 volunteers). In addition to images used for accuracy calculation (LR with scale 2), we also included representations of lower resolution frames (scale 4 and 8) to visualize effect of resolution degradation on classes separability. Reduction of high dimensionality was done with t-Distributed Stochastic Neighbouring Entities (t-SNE) technique. Graphs in Fig. 6.8 and 6.9 show relation between resolution degradation/enhancement and accuracy of face recognition using Euclidean distance between extracted embeddings on IRIS-FR and SC3000-FR datasets, respectively.

Table 6.4. Accuracy of person recognition from test subsets of SC3000-FR and IRIS-FR datasets (80% of all images used for testing, the remaining 20% was used to generate embeddings representing users' profiles in a database)(red - first best, blue - second best for each dataset separately)

SVM with linear kernel								
Dataset	**SC3000-FR**				**IRIS-FR**			
profiles / **test**	orig.	S2	S4	DRESNet	orig.	S2	S4	DRESNet
orig.	99.5	-	-	-	82.14	-	-	-
S2	-	99.17	-	-	-	81.33	-	-
S4	-	-	96.36	-	-	-	74.01	-
DRESNet	-	-	-	99.33	-	-	-	81.87
Euclidean distance								
Dataset	**SC3000-FR**				**IRIS-FR**			
profiles / **test**	orig.	S2	S4	DRESNet	orig.	S2	S4	DRESNet
orig.	99.66	-	-	-	63.48	-	-	-
S2	-	98.67	-	-	-	58.01	-	-
S4	-	-	90.42	-	-	-	57.68	-
DRESNet	-	-	-	98.84	-	-	-	60.85

The aim of the study presented in this section was twofold. First of all, we wanted to determine if representation of facial features in thermal images allows for accurate person recognition. Produced results showed that it is possible to identify subjects from a group of 40 (SC3000-FR) and 30 (IRIS-FR) volunteers with accuracy above 74% using SVM classifier and subjects' representations created with Deep Neural Network FaceNet. Moreover, in case of our dataset those results were much better. Even for the smallest input size (80x60) the recognition accuracy was very high (Table 6.4 96.36%). It turned out that the use of Euclidean distance for person recognition from thermal data depends on the dataset characteristic. As shown in Table 6.4, our dataset produced similar results regardless of the used classifier, while accuracy for IRIS database differs significantly for Euclidean distance comparing to SVM, resulting in ˜20% worse performance. It's important to note that SC3000 sequences were collected by us in strictly defined conditions, i.e. volunteers asked to



(a) Original size

(b) Bicubic interpolation 50%

(c) Bicubic interpolation 25%

(d) Bicubic interpolation 12.5%

Figure 6.7. 2D visualization of embedding vectors produced by Face Recognition NN from images of a) original size, b) S2, c) S4 and d) S8 for volunteers 0-9 from SC3000-FR test set

remain still with a face placed towards the camera. In this way, possibility of a noise presence was reduced. IRIS set, on the other hand, was collected for different facial expressions and poses, thus contained more variability what can lead to worse recognition accuracy. In future studies, we will examine influence of motion and different body poses on person recognition task.

Furthermore, it has been proved that high recognition accuracy can be achieved with a very limited number of samples used for generating users' profiles. A proportion of 2:8 (users' profiles: testing set) was selected in the performed experiments, resulting in only 4 images from the SC3000 and 42 images from IRIS used for extracting facial embedding. Yet, the use of DL model for generation of users' profiles allowed for selecting proper representation that could uniquely describe each person. In our study we decided to use linear kernel. This choice was justified by different factors. First, linear kernels are often preferable with small number of feature vectors and relatively big amount of features within each vector. Secondly, as can be seen in Fig. 6.7, which presents clusters of extracted embeddings, multiple users' representations can be linearly separated with one-against-one approach. This characteristic of utilized data allowed for using linear kernel and thus preserving simplicity of the solution.

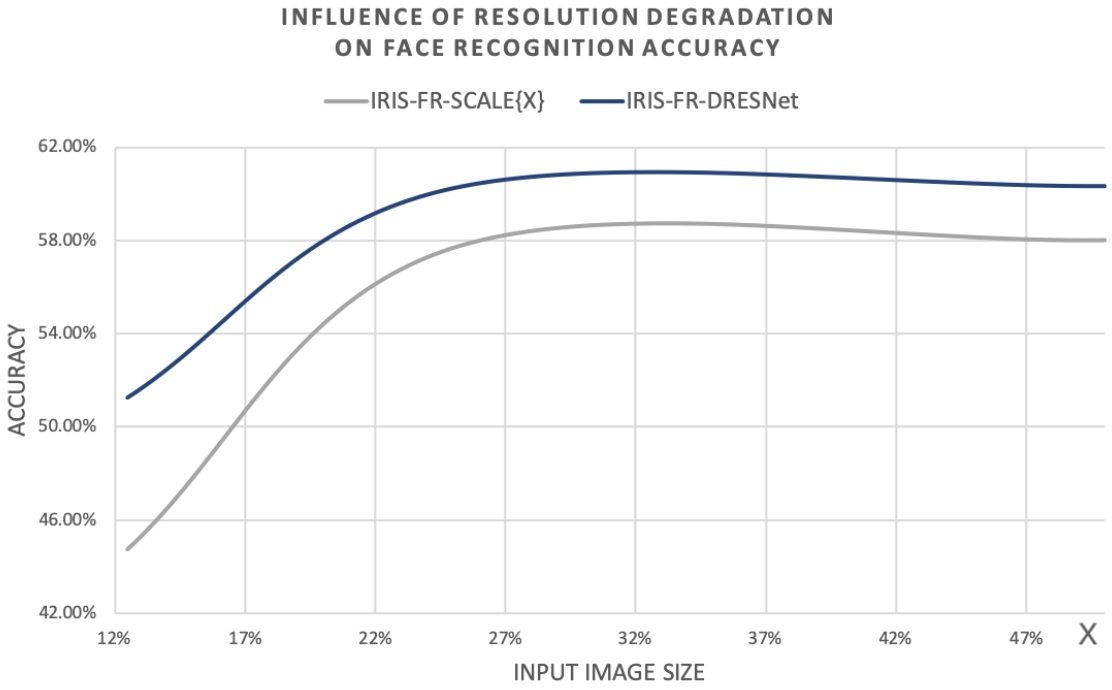


Figure 6.8. Relation between resolution degradation/enhancement and accuracy of face recognition using Euclidean distance between extracted embeddings on IRIS-FR

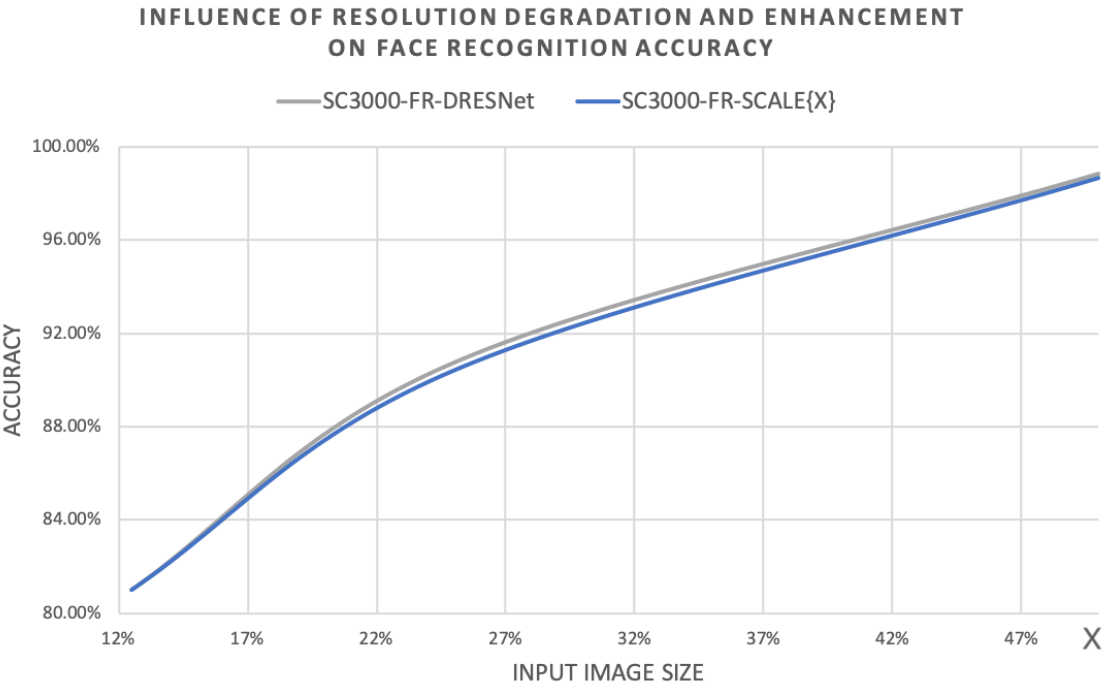


Figure 6.9. Relation between resolution degradation/enhancement and accuracy of face recognition using Euclidean distance between extracted embeddings on SC3000-FR

Calculated accuracy values proved the robustness of the selected kernel, however, we would also like to explore how other kernels affect recognition results from acquired thermal sequences in future work. Another important finding of the presented work is related to the idea of transferring

knowledge between image domain. As previously proved in our studies on facial feature detection (Chapter 4) it is possible to re-purpose deep models trained on visible light data to a novel task in thermal spectrum by utilizing learnt feature representation and optimizing only final classification layer of the network. In experiments presented here, we made use of the same approach and generated thermal embeddings of faces using FaceNet optimized on visible light images. Even though data from both domains have different characteristic, it was still possible to build facial representations that uniquely present each volunteer and achieve high recognition accuracy, as presented in Table 6.4.

The second goal of conducted face recognition experiments is even more important in the context of the presented doctoral dissertation as it allows for verifying the applicability of the introduced super resolution model in other practical applications than the one proposed in thesis II. In all evaluated cases it has been proved that super-resolution leads to increase of person recognition accuracy. However, this gain is higher for scenarios where overall recognition accuracy was lower. For example, for SVM classifier, the performance was improved by less than 1%, what is almost negligible and could have been caused by some random errors.

On the other hand, for IRIS-FR set and Euclidean distance, the accuracy increased by ˜3%. This finding can be also observed by analysis of Fig. 6.8 and 6.9. The increase of recognition accuracy is clearly observable for IRIS-FR database, while for SC3000-FR the increase is minimal. A possible reason for this result is different characteristic of both sets. Our sequences produced very good results for both LR and HR images, so it was very difficult to further improve them with enhanced data. Yet, for datasets like IRIS where motion content is higher, it turned out that improved image quality is beneficial for person identification task. Considering real-life scenarios of remote medical diagnostics, we should assume that presence of motion is inevitable, especially for solutions that will be used without super-vision of a third person, e.g. remote monitoring of vital signs in smart home environment [235, 44] or detection of dangerous health-states in drivers [19]. In such cases, as shown by results on IRIS-FR set, utilization of data enhancement algorithms may be crucial.

It has been also confirmed that decrease of resolution results in decreased recognition accuracy (e.g. 90.42% for SC3000-FR-S4 vs 99.66 for original SC3000-FR inputs). Fig. 6.7 shows influence of image resolution on users' representations separability. Clusters can be easily obtained in case of original high resolution data, but they start to overlap for downscaled images, making classification much more difficult. Thus, utilization of super-resolution algorithms might be required in order to achieve satisfactory performance for very low resolution images in scenarios where target devices impose requirements on spatial size of used sensors (e.g. thermal cameras embedded in wearable devices). Our future work will focus on performing more experiments with different datasets which contain very low resolution sequences (e.g. 80x60 data from Lepton camera). In addition, we will also examine scenarios where subjects are performing different head and body movements in order to verify our assumption that SR helps to improve person recognition accuracy from data that contain more dynamic content.

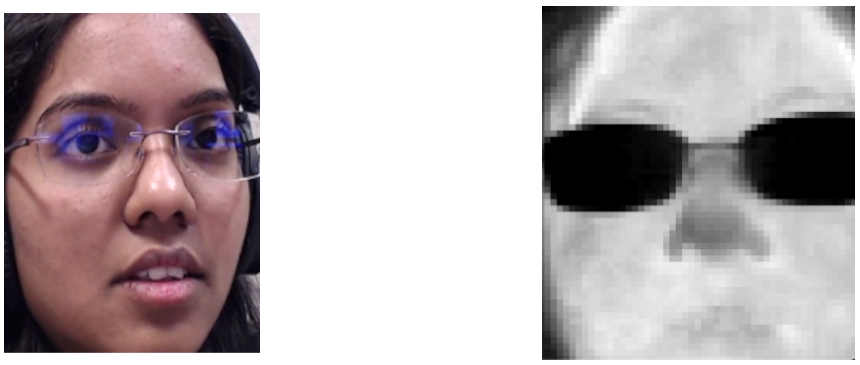
Although achieved results are satisfactory, we are aware of some limitations of the presented study. Data utilized for method evaluation was gathered during a single data collection procedure. As a result, images use for generating users' profiles and for testing the proposed solution were uniform. A very interesting research question is whether similar accuracy could be achieved for sequences collected over longer period of time. This problem should be further explored and analysed in future studies.

6.4.2 Emotion Recognition

Objective

Remote medical diagnostic solutions usually utilize facial areas for extracting information about well-being and state of health [249] The main reason for using face is that it is a highly sensitive part of a body , which can reveal many important details about medical conditions of subjects. Face can be analysed over time e.g. to obtain basic vital signs, e.g. heart rate [44] and respiratory rate [235] or evaluate motor skills of facial muscles in paralysed patients [20], as well as using static images. Applications of the latter are usually associated with non-verbal cues that utilize emotion recognition for e.g. pain level estimation [18] or sentiment analysis [250].

Emotion recognition is usually done by analysis of facial expressions. In visible light domain this problem has been widely studied, using various techniques including machine learning [251], and DL models [252] proving high detection accuracy. One of the reasons of the majority of studies being conducted in visible light is that representations of various facial expressions is clearly distinguishable when high frequency features are present (see Fig. 6.10a).



(a) Visible light image

(b) Thermal image

Figure 6.10. Images acquired simultaneously for invoking fear emotion using visual stimulus

Thermal data, on the other hand (see Fig. 6.10b), record temperature distribution of a facial region instead of its geometric. Solutions which utilize features designed for visible light spectrum, i.e. shapes, edges, points and other appearance information may fail to correctly recognize emotions from thermal images. That's why emotion recognition studies in the thermal image domain received less attention that similar work in the visible light spectrum. Wang S. et al. [253] proposed to use Boltzmann machine for emotion recognition using features learnt from forehead, eyes and mouth regions. Liu Z. and Wang S. [254] utilized histograms of temperature differences between subsequent frames in the recorded thermal sequences . Various techniques based on machine learning algorithms have been also studied, e.g. using linear discriminant analyses [255]. Nevertheless, analysis of facial expressions from thermal images is more difficult due to two main factors. First of all, geometric features are more blurred when temperature distribution of a face is used. Secondly, various diseases and external environment conditions may affect representation of facial expression by changing distribution of temperatures, e.g. paralysed muscles may have different temperature than fully functional ones [20]. Another motivation for our work is the fact that some correlations between stressful situations and temperature distributions in a body have been already discovered [256]. Thus, potentially, it is possible to determine emotional response using signals extracted from thermal image sequences by analysis of color changes within specific facial areas over time.

Taking it into account, we propose to analyze emotional responses from detected vital signs. Some attempts to this problem have been already done by Yin Z. et al. [257], who proposed to make use of various physiological signals (e.g. electroencephalogram EEG, electrocardiogram ECG or electromyogram EMG signals) to evaluate and identify emotions. Guo H. et al. [258] also made use of ECG data. In another research, heart rate and skin conductance were measured to distinguish neutral, positive and negative feelings [259]. By using physiological states, predicted results may be more reliable as masking or suppressing of biosignals is much harder than changing facial expressions. Moreover, it can be used when subjects are not able to communicate emotions in other ways, e.g. infants or paralysed people. We follow similar approach and propose to use respiratory information for emotion analysis. Yet, contrary to [257, 259], biosignals utilized by us are extracted in a contactless way from video sequences (as previously described in Section 6.3.2) instead of using external sensors. In addition to respiratory rate, we also determine whether heart rate changes with emotions and how it correlates with extracted respiratory rate. In this way, our solution can be applied in various remote diagnostic scenarios in a non disruptive way, without forcing any action from users, such as placement of electrodes or attachment of external devices.

Methodology

In the view of foregoing, our study presented in [158] focused on determining whether vital signs (heart and respiratory rate) extracted from recorded sequences can be used for analysis of emotional responses. Specifically, we evaluated how those parameters change for imitated and video-invoked emotions and what conditions should be met to design provide remote diagnostic solution which makes use of emotional responses of monitored subjects.

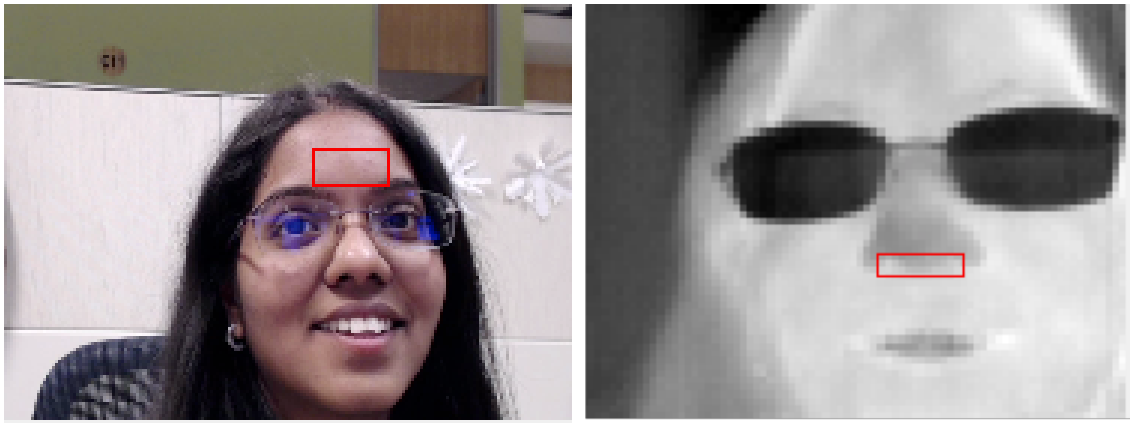
For this study, we utilized datasets collected simultaneously with thermal and RGB cameras in order to make use of multimodal responses associated with emotions. Vital signs were estimated in a contactless way by processing recorded sequences (respiratory rate from thermal data, heart rate from visible light data). Furthermore, facial expression were detected from RGB frames using Microsoft Emotion Cognitive Service (MECS) [1]. The output from MECS is represented as a vector containing confidence levels corresponding to various emotions: anger, contempt, disgust, fear, joy, neutral, sadness, surprise. In addition, perceptible emotional response was collected using an online questionnaire (described in details in Chapter 3), as a type of an experienced emotion is an individual matter. As a result, we were able to evaluate correlation between facial expression, estimated physiological signals and the real emotion experienced by volunteers and noted by them in the survey. Details about data collection were presented in Chapter 3. According to the nomenclature specified in this chapter, databases are referred thereafter as Logitech900-ER-simulated, Lepton-ER-simulated, Logitech900-ER-invoked and Lepton-ER-invoked, where first part of the name corresponds to the used sensor, second means Emotion Recognition (ER) study and the last indicates if emotion was invoked with visual stimuli or simulated by a participant.

Vital signs were estimated from short fragments of recorded sequences to ensure a fast response provided to a user in target applications (˜400 samples from thermal data and ˜500 subsequent RGB frames). In order to accommodate possible inertia of expressed emotion, vital signs were estimated twice: from the beginning and the end of each recording. Respiratory rate (RR) of each volunteer was extracted using the same method as in the study on resolution influence on the accuracy of RR estimation (Sec. 6.3.2). For this experiment, we were dynamically selecting aggregation operator leading to best estimation result (average, variance or skewness). Also, only

[1] http://azure.microsoft.com/en-us/services/cognitive-services, Accessed: September 2018

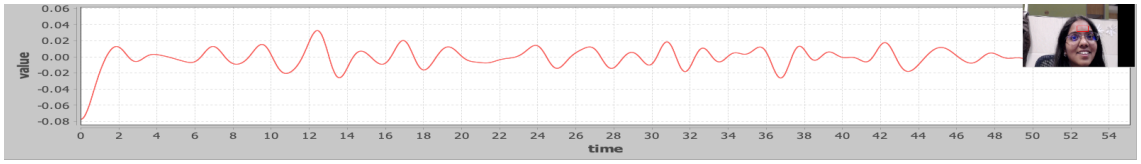
the best performing RR estimator according to our previous work [172] was applied, i.e. estimator based on frequency of the dominated peak in the frequency spectrum eRR_{sp}.

Heart rate (HR) was obtained from recorded visible light sequences in a similar non-contact way by analysis of color changes within manually marked RoI. As presented in literature [44] and proved by our previous studies [169], the best results are achieved for signals extracted from a forehead area using YUV color space. Thus, sequences were at first converted using H264 codec and transformed to YUV color space. Having target platforms with limited compute resources in mind, images were downscaled to 800x600 and a frame rate was reduced to 15 frames per second (FPS). Then, RoIs were selected to cover forehead regions for all participants, instead of nostrils as in case of RR estimation. Examples of areas marked for both vital signs are presented in Fig. 6.11. After that, raw signals were constructed by averaging pixel values within marked forehead areas and filtered with a band pass Butterworth filter (frequency bandwidth between 0.67Hz and 4Hz), as verified in [21]. Finally, the same estimator as for RR was used for calculation of a pulse rate. Examples of raw heart rate and respiratory rate signals are shown in Fig. 6.12. Due to poor lighting conditions in some recorded RGB sequences, frames from only 6 volunteers contained information usable for HR estimation. Thus, HR was calculated only for those participants. In future work, we are planning to acquire sequences for more volunteers, making sure facial region is sufficiently lit.

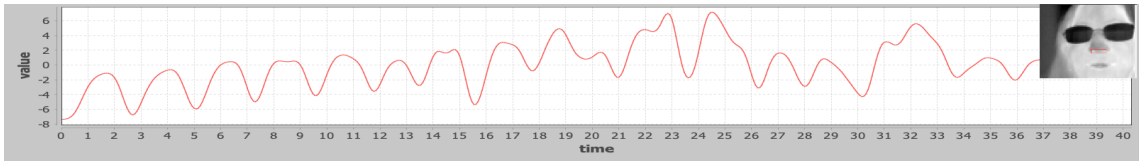


(a) RoI for heart rate estimation (b) RoI for respiratory rate estimation

Figure 6.11. RoIs used for vital signs estimation during invoked joy emotion



(a) Heart rate



(b) Respiratory rate

Figure 6.12. Examples of raw signals extracted from selected RoIs (presented in Fig. 6.11)

Results and Discussion

Values of respiratory rate estimated from collected thermal sequences for emotions simulated (S) by volunteers (Lepton-ER-simulated) and for responses invoked (I) by video stimulus (Lepton-ER-invoked) are presented in Table 6.5. Corresponding results for pulse rate were collected in Table 6.6. To accommodate for eventual inertia of emotional response, vital signs were estimated from selected beginning (B) and last (L) samples of collected sequence.

Table 6.5. Respiratory rate estimated from beginning (B) and last (L) samples of thermal sequences; S - simulated, I-invoked, tp - technical problem during data collection, fb - face turned away

	S:neutral		**S:joy**		**S:fear**		**S:disgust**		**I:neutral**		**I:joy**	
Subject	B	L	B	L	B	L	B	L	B	L	B	L
1	21.6	20.3	18.9	21.6	21.6	21.6	16.2	17.6	21.6	18.9	tp	tp
2	20.3	20.3	18	19.8	18.9	18.9	24.3	25.2	18.9	18.9	28.4	20.3
3	18.7	22.9	21.6	21.6	18.9	18.9	19.8	18.0	18.9	21.6	17.6	17.6
4	11.7	10.8	14.7	16.7	16.2	16.2	16.2	14.4	16.2	14.4	17.2	14.9
5	13.5	13.5	16.2	16.2	12.2	12.2	14.9	13.5	12.2	15.9	tp	15.1
6	21.6	22.7	21.6	23.1	22.2	22.2	17.3	20.9	22.2	14.3	21.6	20.3
7	14.4	13.1	14.4	14.4	tp	tp	12.6	14.4	tp	tp	22.9	16.2
8	18.5	18.5	14.0	15.8	27.0	27.0	24.0	27.7	27.0	18.9	21.6	22.7
9	14.4	12.6	21.6	21.6	11.2	11.2	14.4	21.6	11.2	11.2	17.6	18.9
10	18.0	18.0	18.0	19.8	17.6	17.6	19.8	21.6	17.6	17.6	21.6	18.9
11	21.6	21.6	21.6	19.8	17.6	17.6	21.6	20.3	17.6	17.6	18.9	18.9
	I:neutral		**I:disgust**		**I:neutral**		**I:fear**		**I:neutral**		**I:sad**	
Subject	B	L	B	L	B	L	B	L	B	L	B	L
1	18.9	21.6	16.2	17.6	18.9	21.6	17.6	21.6	18.9	20.3	21.6	21.6
2	18.9	17.6	18.9	17.6	18.9	14.9	23.4	fb	21.6	23.0	20.3	fb
3	18.9	19.8	20.3	18.9	16.2	18.9	20.3	20.3	18.9	17.6	16.2	20.3
4	15.1	14.7	14.0	10.8	11.9	14.4	15.1	15.1	12.1	10.8	13.5	14.9
5	12.1	13.5	13.5	16.2	12.1	13.5	15.8	fb	12.1	13.5	12.2	13.5
6	16.2	15.4	14.4	30.6	12.9	12.9	18.0	21.6	18.0	14.4	17.1	17.8
7	12.6	12.6	13.5	16.2	12.6	12.6	14.9	13.5	10.8	10.8	14.4	14.4
8	14.5	14.5	20.1	19.8	21.6	18.0	27.0	23.6	14.4	12.6	25.2	25.2
9	16.2	16.2	14.9	14.9	14.9	14.9	14.9	16.2	13.5	14.9	16.2	16.2
10	18.0	18.0	18.0	18.0	15.0	19.8	18.0	16.2	18.0	18.0	18.0	18.0
11	12.1	12.1	16.8	15.6	17.6	17.6	19.8	18.0	16.2	18.0	18.0	19.8

Table 6.6. Heart rate estimated from beginning (B) and last (L) samples of visible light sequences; S - simulated, I-invoked, tp - technical problem during data collection, fb - face turned away

	S:neutral		**S:joy**		**S:fear**		**S:disgust**		**I:neutral**		**I:joy**	
Subject	B	L	B	L	B	L	B	L	B	L	B	L
6	59.8	59.8	57.1	59.5	61.7	59.4	59.6	61.7	59.9	57.8	66.7	62.0
7	57.2	56.6	69.4	72.0	66.9	66.9	66.9	70.2	61.7	59.1	72.7	88.2
8	84.6	85.0	90.0	91.6	82.8	88.2	88.2	90.0	82.8	81.0	87.2	79.3
9	59.4	57.6	82.8	84.6	63.0	64.8	64.8	59.9	57.7	57.7	61.2	66.6
10	75.3	76.9	74.9	75.3	76.8	79.9	80.4	81.0	76.5	79.2	80.5	82.9
11	73.4	70.6	76.8	70.7	75.8	80.1	78.6	71.4	75.5	72.2	75.0	75.9
	I:neutral		**I:disgust**		**I:neutral**		**I:fear**		**I:neutral**		**I:sad**	
Subject	B	L	B	L	B	L	B	L	B	L	B	L
6	61.7	59.1	64.1	59.5	57.9	60.0	61.2	59.5	59.8	59.8	59.9	62.7
7	57.7	59.5	69.5	63.1	55.9	55.9	57.7	57.7	55.8	59.8	57.6	59.0
8	90.0	77.4	86.5	81.1	82.9	89.2	75.7	72.1	82.8	81.7	70.2	70.2
9	61.2	77.4	63.0	63.0	59.4	61.2	63.0	61.2	59.4	61.7	59.5	58.7
10	78.0	78.9	84.1	81.7	85.4	81.4	80.4	83.7	80.4	80.6	82.9	81.3
11	73.1	79.0	78.2	77.2	82.3	76.1	75.2	83.2	72.1	80.2	73.7	75.3

Table 6.7. Emotions self estimated by volunteers and collected using the questionnaire filled out during data collection

	Video number: pre-defined emotion							
Subject	1:neutral	2:joy	3:neutral	4:disgust	5:neutral	6:fear	7:neutral	8:sad
1	neutral	joy	neutral	disgust	neutral	fear	neutral	sad
2	neutral	joy	neutral	disgust	neutral	fear	neutral	sad
3	neutral	neutral	neutral	disgust	neutral	fear	neutral	neutral
4	neutral	joy	neutral	disgust	neutral	joy	neutral	sad
5	joy	joy	neutral	disgust	neutral	fear	neutral	sad
6	neutral	joy	neutral	disgust	neutral	fear	neutral	sad
7	neutral	joy	neutral	disgust	neutral	neutral	neutral	neutral
8	neutral	joy	joy	disgust	neutral	fear	joy	sad
9	neutral	joy	neutral	disgust	neutral	joy	neutral	sad
10	neutral	joy	neutral	neutral	neutral	neutral	neutral	sad
11	neutral	neutral	neutral	disgust	neutral	fear	neutral	sad

Emotions specified by participants in the online questionnaire as dominant during watching each video stimulus are collected in Table 6.7. In addition, we also wanted to evaluate whether facial expression corresponds to actual emotion felt by volunteers and with relations indicated by changes in vital signs. Thus, Microsoft Emotion Cognitive Service was utilized in this study to extract emotions from visible light images by classifying facial expressions, as already mentioned in Methodology Section. The most dominant emotion for each volunteer while watching different videos and percentage of frames in which it was detected is presented in Table 6.8.

Table 6.8. Facial expression detected from RGB frames using Microsoft Emotion Cognitive Service (MECS). For each subject and video the most frequent emotion was noted, followed by a percentage of frames in a sequence, where this emotion was dominant

	Video number: pre-defined emotion							
Subject	1:neutral	2:joy	3:neutral	4:disgust	5:neutral	6:fear	7:neutral	8:sad
1	neutral	joy	neutral	disgust	neutral	neutral	neutral	neutral
	100	74	100	71	100	99	100	100
2	?	joy	?	neutral	neutral	?	neutral	?
	78	72	51	77	72	46	89	78
3	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
	100	93	100	88	100	92	100	99
4	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
	96	82	89	71	95	91	85	93
5	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
	93	64	99	93	97	49	89	82
6	neutral	joy	neutral	neutral	neutral	neutral	neutral	neutral
	100	92	100	78	100	98	100	100
7	neutral	joy	neutral	joy	neutral	neutral	neutral	neutral
	100	89	100	59	100	98	100	100
8	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
	96	65	100	61	100	71	100	96
9	neutral	joy	neutral	joy	neutral	neutral	neutral	neutral
	100	58	100	64	56	80	100	98
10	?	?	neutral	neutral	?	?	neutral	?
	68	53	100	68	51	57	50	55
11	?	?	neutral	?	?	neutral	neutral	neutral
	62	85	55	74	53	70	95	87

To compare values of vital signs and estimate which one is potentially more dependant on emotional responses, a relation between calculated pulse and respiratory rates was plotted in Fig. 6.13 for joy and neutral responses for volunteers 6-11. Since, we believe emotional response is

an individual matter, we also presented relationship between vital signs for two chosen subjects during watching videos aimed at simulating joy and neutral responses (see Fig. 6.14). Changes in estimated vital signs during transitions from various emotional stages invoked with video stimulus are shown in Fig. 6.15 and 6.16 for pulse and respiratory rates, respectively.

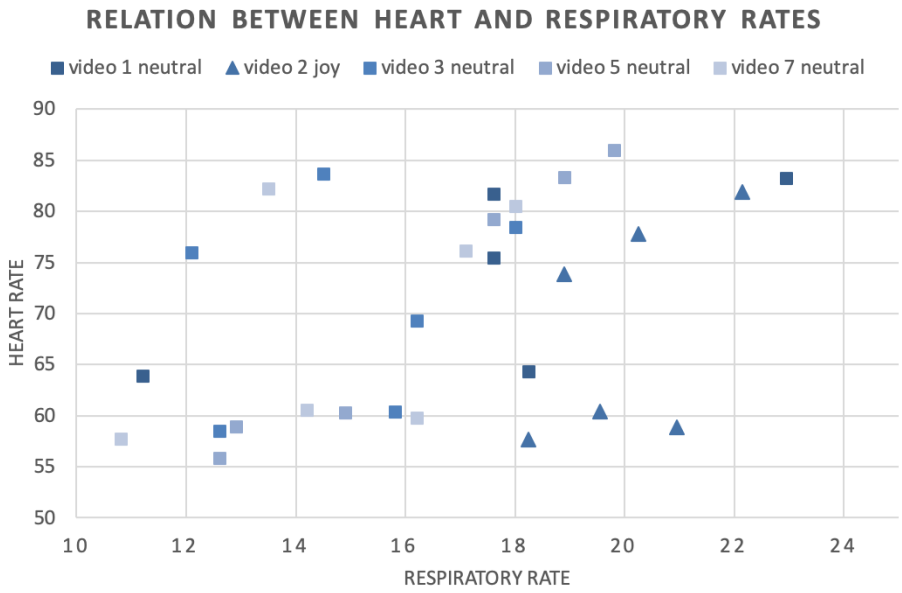


Figure 6.13. Relation between vital signs for stimulation video 2: joy (triangle) and neutral videos (squares) for volunteers 6 to 11

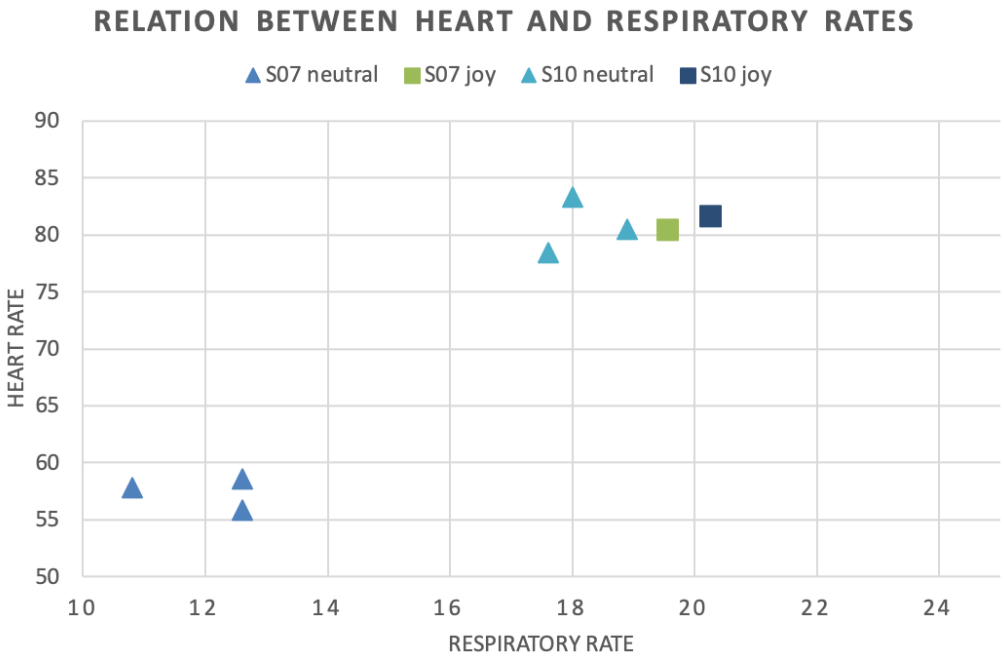


Figure 6.14. Relation between vital signs for stimulation video 2: joy (square) and neutral videos (triangles) for 2 chosen subjects (subject 7 and subject 10)

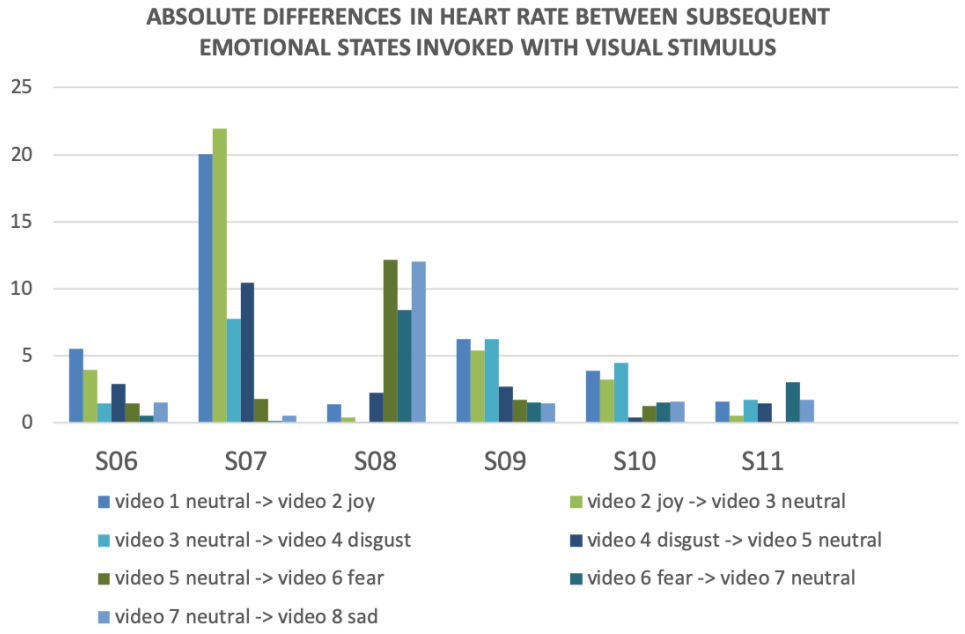


Figure 6.15. Changes in estimated pulse rate values during transition from emotions invoked by video stimulus for subjects S06-S11

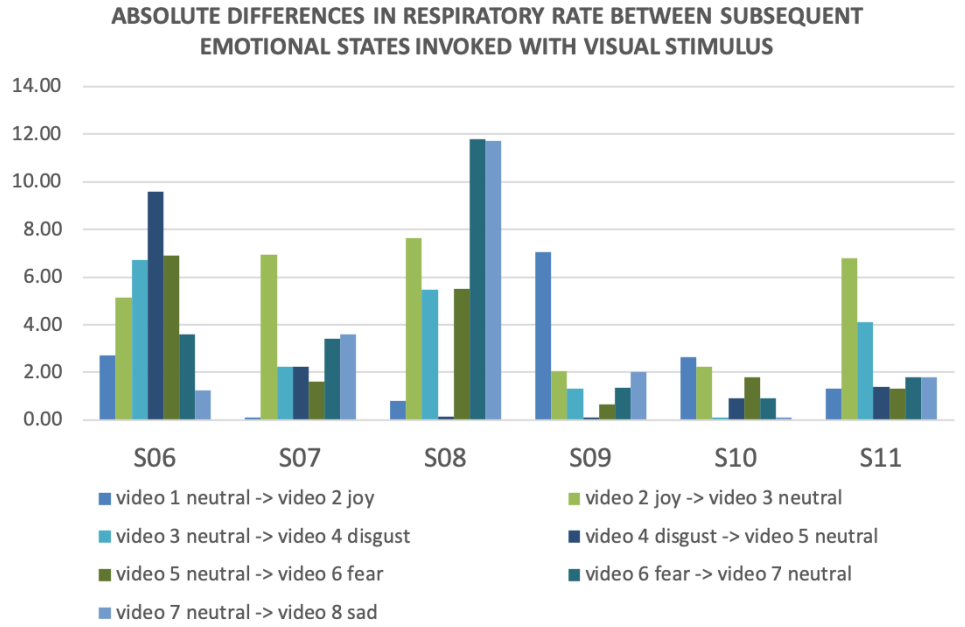


Figure 6.16. Changes in estimated respiratory rate values during transition from emotions invoked by video stimulus for subjects S06-S11

This study focused on evaluating influence of emotional states on vital signs. Performed experiments included scenarios for both invoked and simulated emotions. Vital signs were obtained in a contactless way from facial regions, as presented in previous Section (Sec. 6.3.2). According to users' responses collected in Table 6.7 we can observe that volunteers claimed to feel emotions quite consistent with assumptions made by us while selecting video stimulus. However, as can be seen from Table 6.8, facial expressions detected from visible light sequences don't correspond to those

responses, resulting in neutral output in most of the cases. This result clearly indicates the need for obtaining information about emotions from inputs other than facial expressions. Therefore, we believe that utilization of vital signs is very beneficial, allowing for getting reliable response, as masking vital patterns is very difficult.

The threshold for indicating the observable change of vital signs between different emotional states (δf) was set to 1.8, following the finite frequency resolution of obtained signals, defined in the same way as in our previous experiments, i.e. with Eq. 6.1. Since visible light sequences were sampled with frequency 15Hz and 500 frames were used for estimation, $\delta f = \frac{15}{500} * 60 = 1.8$ beats per minute. Similarly for thermal data collected with frequency of 9Hz and 300 samples used for estimation, $\delta f = 1.8$ breaths per minute. Thus, we believed this value is a safe threshold for making valuable conclusions about influence of emotions on vital signs.

Analysis of vital signs data collected in Table 6.6 and 6.5 prove the existence of changes in heart rate and respiratory rate values during transitions between different emotional states. This finding is also visible in visualization of vital signs differences presented in Figures 6.15 and 6.16. Analysis of plotted graphs confirm that change of emotions is an individual issue and different changes were observed for different subjects. On the other hand, we can also see that respiratory rate show more clear distinction between different states than the heart rate and some results are consistent among volunteers. For example, in 4 cases the biggest RR change was observed for transition either between neutral and joy or joy and neutral states. Differences in heart rate are more aleatory. 2 subjects reacted with the biggest HR change for joy, 2 for disgust and 2 for fear emotion. For invoked joy emotion, 67% of cases resulted in respiratory rate difference higher than 1.8 bpm comparing to its previous value during neutral stimulus. Taking into account the average RR value for all neutral states, this difference was higher than 1.8bpm in almost all cases (89%). Other emotions depend heavily on the individual. Figure 6.14 shows that for volunteer 7 there is a clear distinction in pulse and respiratory rate values between joy and neutral states, while for volunteer 10 the differences are much smaller and estimated values are closely aggregated. Therefore, we believe that the proposed solution is more suitable for detecting general categories of emotional states, e.g. in a binary classification of positive vs negative response rather than predicting specific outcomes, such as fear, disgust, joy, etc.

According to Fig. 6.13, it can be noted that respiratory signals lead to better separation of evaluated emotions than the pulse rate. Yet, it is still an individual matter and it's difficult to define a general rule for emotion recognition. Also, some subjects can perform dynamic movements in different emotional stages that would lead to some problems with signal acquisition, e.g. invisibility of facial regions. In order to increase system robustness, additional motion compensation and/or gaze and face detection algorithms should be introduced to detect the moment when interesting facial region disappeared from the field of view. Due to high influence of personal features and behaviour on achieved results, we believe that the proposed system should be tuned separately for each volunteer and learn his/her patterns in order to make valuable decisions in the future.

It's important to note that the proposed emotion recognition study is only preliminary and neither image enhancement nor emotion classification was performed with the use of AI. Yet, in the future work we would like to focus on integrating the proposed method with DL and evaluate possibility of improving recognition accuracy, e.g. by classifying emotions from vital signs patterns using Recurrent Neural Network or obtaining predictions from super-resolved sequences. Although results are promising, we still observe dependence of results on the subject and our initial conclusions should be further confirmed in future experiments.

6.5 Summary

In this chapter we presented examples of applications which benefited from the proposed DL techniques and models verified and designed specifically for thermal images. Experiments performed for facial areas detection showed that image enhancement is crucial for improving accuracy of the system. Evaluated super-resolution models allowed for increasing IoU by at least 15% comparing to LR bicubic images. In the case of using the network proposed by us, the accuracy gain was even higher. Face area detection after enhancing image with DRESNet (window size 90) was improved by 56% comparing to corresponding low resolution image. The improvement was also very significant for other facial regions, e.g. nostril detection improved by 49% and eyes area improved by 13%. This finding supports the second part of thesis II which states that increased resolution of thermal images lead to improvement of facial areas detection accuracy.

Furthermore, we also proposed and discusses possibilities of applying the introduced network in other applications potentially useful for remote medical diagnostic systems. Experiments performed for contactless respiratory rate estimation and face recognition using enhanced thermal sequences proved the robustness of the introduced SR model and its positive influence on resulting accuracy in analysed scenarios. Since the proposed SR network allows for improving methods used at different stages of remote person monitoring, i.e. person recognition, facial areas detection and vital signs estimation, we could potentially build a very accurate medical diagnostic system allowing for obtaining user-specific vital signs patterns during daily activities.

Additionally, we described details of preliminary study conducted by us for the task of emotion recognition from physiological signals estimated using image processing techniques. Potentially, the introduced work could also take advantage of image enhancement techniques, as more accurate estimations of vital signs can improve detection of emotional responses. This problem will be studied by us in future work.

Chapter 7

Future work

7.1 Introduction and Overview

Extensive benchmark evaluation performed for proposed novel thermal image processing Deep Learning (DL) approaches and practical remote diagnostic solutions, which utilize designed neural network architecture exposed some limitations of introduced techniques that we would like to address in future work. In this chapter, we will focus on providing ideas for improvement of the proposed Super Resolution Model.

Furthermore, we will suggest an innovative approach to neural network design using evolutionary algorithms that could ease the search for the optimal architecture. Other methods for enhancing thermal imagery in order to increase accuracy of contactless medical diagnostics will be also described. According to remote diagnostics solutions verified by us (see Chapter 6), the designed Super Resolution (SR) model helps with improving performance of applications which utilize image processing algorithms. Yet, we only examined a limited set of such scenarios. Thus, at the end of the chapter, we will provide an overview of other remote computer vision-based medical solutions that could potentially benefit from higher image resolution increased with the means of the proposed image enhancement method.

7.2 Improvements of the proposed Super Resolution Model

The idea of improving DL techniques and methods proposed by us for thermal image analysis can be treated as a double entendre. First of all, due to a huge interest and thereby very fast progress in Artificial Intelligence (AI) research, new ideas for network design and training are being continually introduced to the state-of-the-art knowledge. Although, results proved high accuracy of the designed super resolution model, we would like to examine other techniques which have been developed in DL area to examine if performance can be further improved.

Secondly, presented studies have been performed only in an experimental setup. To make sure that proposed algorithms are suitable for production-ready solutions a more in-depth analysis should be conducted bearing in mind possible factors that could influence results. Both aspects of network improvements are discussed in this section.

7.2.1 Architectural and Optimization Changes

Removal of Batch Normalization

One of architecture changes that we want to explore in further studies is removal of batch norm layer. Currently, the best performing DRESNet architecture normalizes outputs of each convolutional layer in the feature extraction subnetwork. This design decision was inspired by studies conducted by Ioffe S. and Szegedy C. [210] and authors of the ResNet model [113], which proved that training time can be reduced, while improving network accuracy if normalization is included in the architecture itself.

Batch normalization can be better understood by revisiting the process of network training. Before introduction of this operation, the only normalization performed during network training was introduced to network inputs, e.g. z normalization, also known as data standardization, defined as:

$$x_{out} = \frac{x_{in} - \mu}{\sigma} \tag{7.1}$$

where x_{out} is a new value of each input x_{in}, μ is a mean value of a batch and σ is a standard deviation of a batch. Yet, as can be noted during NN training we also observe changes in inputs distribution at a layer level due to adjustments of weights. As a result, lower learning rates lead to better performance, since steps used for network optimization are less prone to those changes. This problem is referred to as covariate shift and as proposed in [210] can be solved by performing normalization of each layer inputs. Batch normalization extends the Eq. 7.1 by introducing two parameters (β and γ) learnt separately for each level:

$$x_{out} = \gamma(\frac{x_{in} - \mu}{\sigma}) + \beta \tag{7.2}$$

It's worth mentioning that in case of images, every pixel is treated as an example. Thus, the mean value (μ) and the standard deviation value (σ) are calculated over $N * W * H$ samples, where H and W are image height and width and N is a number of images in a batch. As shown by [210], batch normalization approach reduces possibility of exploding and vanishing gradients and eliminates a need of applying some other techniques, e.g. dropout, as it already acts as a regularization method.

On the other hand, as shown by Lim B. et al. [209] in their Enhanced Deep Residual (EDSR) Super Resolution network, removal of batch normalization layers is advantageous in case of regression tasks, such as Single Image Super Resolution (SISR). The goal of SISR is to restore image features by performing image-to-image mapping instead of building abstract representations used for classification, as it is in a case of ResNet. Therefore, if input and output of the network is highly similar and correlated, batch normalization can lead to decrease of performance. Taking it into account, we would like to evaluate accuracy achieved by DRESNet if batch normalization is not applied.

Width and Size of Convolutional Kernels

Second limitation of the proposed SR thermal data-oriented model is a number and a size of filters. In order to ease the training procedure by keeping a number of parameters at minimum, we only tested configurations where all convolutions were using 96 filters of a size 3x3. Yet, as shown in studies performed with SRCNN [190] (pioneer Convolutional Neural Network (CNN) SR network), we can observe performance gain for networks with bigger convolution width (the term width corresponds to the number of filters used within each convolution). Other state-of-the-art

SR models also took advantage of this finding, introducing 128 (DRRN [193]) or even 256 filters (DRCN [192]). On the other hand, utilization of more filters comes at a cost of processing time, what can be especially important for our target applications, often limited in terms of available computational resources. In such cases, it may turn out that smaller network width is preferable, since achieved accuracy is still satisfactory while keeping higher restoration speed (as proved by DRESNet which outperformed other networks in performed experiments using only 96 filters).

Another aspect of convolutional filters design is associated with their size. It has been also shown by SRCNN that larger kernels allow for increasing model accuracy [190]. The intuition behind applying bigger masks for feature extraction lies in utilization of richer structural information. At the non-linear mapping step it's also beneficial as bigger filters allow for utilizing more distant neighborhood relations. Yet, similar results may be achieved by using residuals and recursions with shared weights, as proposed in DRESNet. Convolution filters larger than 1×1 applied iteratively lead to widening of the receptive field and thus achieving comparable effect as in a case of bigger kernels. Contrary to bigger filters though, residuals and recursions with shared weights do not cause expansion of parameters number, what allows for keeping small network size. Yet, the interesting research question is whether combination of bigger filter size and convolutions used recursively will lead to even better results than with the proposed DRESNet configuration. The choice of network width and filters size should always involve analysis of a trade-off between inference time and restoration accuracy. It would be useful to perform such comparison in future studies.

Gradient Clipping

Potential improvements could also target the training procedure itself. According to previous studies on Convolutional Neural Networks [39], deeper architectures lead to better performance. However simple stacking of more layers is not efficient, as it may result in vanishing/exploding gradient problem. Various techniques to address this problem have been already proposed and also introduced by us in the proposed SR model, e.g. use of specific activation functions, such as ReLU [260] which allow for selecting features better for image recognition; applying supervision to recursion to make backpropagation of gradients easier [192]; utilization of residual blocks [113]; previously described batch normalization [210]; specific weight initializer, e.g. He algorithm [212].

There is also one more very effective approach, which haven't been tried by us so far. This technique, known as gradient clipping, was proposed a few years ago in the study on recursive neural networks [261]. The basic rule of gradient clipping is based on clipping the gradient value whenever it exceeds a fixed threshold. This simple, yet very effective strategy, turned out to be crucial for network convergence using Stochastic Gradient Descent Optimizer. Later, gradient clipping has been also shown to improve training procedure of SR convolutional networks by possibility of using higher learning rates (VDSR [191] applied 4 times smaller learning rate than SRCNN, while producing even better results). Although the usage of gradient clipping is still limited in CNNs, as it was originally designed for recurrent networks, we would like to examine how our proposed network saturates when clipping strategy is utilized.

7.2.2 Production-ready Solution

This aim of this work was to propose novel Deep Neural Network (DNN) architectures that would allow for accurate processing of thermal sequences for remote medical diagnostic solutions. Conducted experiments were established on strictly defined conditions of data collection process,

e.g. person looking towards the camera, controlled laboratory environment, etc. Some initial studies for evaluation of proposed techniques in practice have been also proposed and described. Specifically, one of databases (see Section 3.2.1) was collected with thermal camera mounted on a wearable device (eGlasses platform [150]) to simulate eventual influence of physician's movement on possibility to detect facial areas. Experiments performed with this set were described in Section 4.2.1.

Yet, a more detailed experiments in challenging environments and real life setting are necessary in order to ensure the software is reliable and production-ready. One of the factors that should be considered is deployment of models on target platforms. All of presented results were produced by performing an *offline* testing, meaning that datasets were collected in one step and then processed on a separate device, i.e. desktop PC, laptop or NVIDIA® DGX-1™ Station equipped with four GPU Tesla V100 cards, dedicated to processing of DL workloads. In order to accurately evaluate utilization of computational resources, verify if size of models is optimal and make sure that responses from system are returned to users with acceptable latency, end to end solution should be run directly on smart glasses or other resource-constraint device that is usually used in smart home environments or telemedicine systems. Inference time should be measured as a time required by the device to process a sequence of thermal images collected using camera connected to the platform and return results of estimated vital signs. It should be also verified if model and a sequence of frames needed for performing medical diagnostics are within the memory capacity offered by the device. Examples of latest embedded AI computing platforms and AI accelerators include NVIDIA© Jetson [1], Intel© Nervana Neural Network Processor [2], or Intel© Movidius [3].

Another important aspect that should be verified involves definition of measurement conditions. In real life scenarios distance between subjects and camera, body position and other settings vary. Performed tests were conducted on different databases collected using numerous acquisition devices in order to provide reliable conclusions. However, to confirm results, test sets should be acquired in different conditions and environments to determine if trained networks can generalize well and aren't biased towards previously utilized datasets.

It would be also beneficial to provide algorithmic redundancy to increase reliability of response. Although we identified the configuration of DRESNet that leads to the best restoration accuracy in Chapter 5, other structures may turn out to perform better in different scenarios. Numerous configurations of the proposed SR model could be used for resolution enhancement and then respiratory rates could be obtained with different estimators (e.g. all estimators utilized in experiments described in Section 6.3.2) simultaneously. After that, results may be combined to build algorithmic-redundant software and assess the overall reliability of the system. It is also important to specify clear terms of use and indicate that results produced by the proposed solution are not professional medical diagnosis and should be consulted with specialists.

In order to perform analysis of practical applications of introduced DL thermal image processing techniques for the needs of remote medical diagnostics, users' ratings should be collected after testing proposed neural networks, e.g. using a questionnaire. Collected responses should contain information about a number of trials needed to obtain the measurement, time spent on data collection and processing, a number of failed attempts, and general rating indicating out of the box experience, ease of use, clarity of instructions and interfaces. To perform such evaluation, specific usage scenarios should be prepared and conducted on a bigger group of volunteers.

[1] https://www.nvidia.com/en-us/autonomous-machines/embedded-systems, Accessed: February 2020
[2] https://www.intel.ai/nervana-nnp/nnpi, Accessed: February 2020
[3] https://software.intel.com/en-us/movidius-ncs, Accessed: February 2020

7.3 Optimal Architecture Design with Neuroevolution

One of the presented dissertation theses aimed at verifying whether introduced novel Deep Neural Network architecture allow for thermal image enhancement and thereby improvement of facial areas detection accuracy in order to provide more reliable non-contact vital signs diagnostic solutions. Experiments conducted on various datasets proved this statement and resulted in proposal of the best configuration of convolutional-based Super Resolution (SR) model which outperformed other state of the art networks (see Chapter 5 and 6).

In our experiments conducted for thermal data enhancement the best performing neural network architecture was selected by randomly applying various number of residual and recursion blocks to all proposed subnetworks of the SR model (i.e. feature extraction, non-linear mapping and reconstruction), training them and comparing image quality metrics (Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM)). In total more than 60 configurations were evaluated. Although this approach allowed for in-depth analysis of placement and number of utilized blocks leading to best image reconstruction results, outperforming other state-of-the-art solutions. Yet, the process of finding the best architecture was very time-consuming and doesn't guarantee that the best configuration has been found, since parameters were chosen from limited sets of recursive and residual blocks (as presented in Section 5.3.4). Taking it into account, it would be beneficial to take advantage of other approaches used for generation of optimal neural network structures.

Neuroevolution is one of such solutions. It is based on stochastic search methods and operate on a population of genotypes mapped to neural network structures evolved to find the best model fitness. Neuroevolution has been first proposed for generation of Artificial Neural Networks in 2009 [262] and has recently gained a lot of popularity in DL studies [263] . Evolutionary algorithms possess some advantages over other commonly used optimization algorithms, e.g. back-propagation. First of all, parameters of neural network can be encoded in genomes and evolved during training based on applied evaluation metrics. It has been shown that such approach is more effective than optimization using cost functions and allows for finding better performing network in fewer computational cycles [264].

Moreover, it is also possible to combine neuroevolution algorithms with learning methods in order to provide solutions that can dynamically adapt to various environmental changes [265]. Evolutionary algorithms can be also utilized for hyperparameters search. While using conventional learning algorithms, initialization of hyperparameters may have a significant influence on final model accuracy. Thus, various techniques have been proposed to improve state-of-the-art results, e.g. He [212] initializer. Yet, as proved by research conducted on evolution-driven supervised learning [266], a very promising results can be achieved if initial parameters are find with evolution algorithms and then network is optimized in a conventional way, e.g. using back-propagation.

Inspired by those findings, we are interested whether neuroevolution could help with evolving super resolution thermal data-oriented architecture that will result in better resolution enhancement than proposed, found with random search, but still fixed topology. Detailed description of various approaches for evolving neural networks has been provided by Floreaon D. et al. [262]. We are planning to evaluate them in future studies in order to propose other configurations of the introduced DRESNet architecture.

7.4 Other Algorithms used for Image Enhancement

In the presented dissertation, we focused on improving quality of thermal data using Super Resolution Deep Neural Networks. Experiments performed for target remote medical diagnostic solutions, described in Chapter 6, showed that image enhancement has a positive effect on facial areas detection and extraction of vital signs by analysis of pixel intensities changes (within detected regions). Yet, it is important to note that Super Resolution algorithms solve a specific problem of image quality degradation, defined by Eq. 5.2, i.e. reversing the effect of data down-scaling.

Hence, there are also other operations that lead to degradation of image quality and as a result may negatively affect computer vision-based telemedicine solutions. Some examples of such problems include influence of noise or blurring of features present in collected samples. All of those degradation effects can be mitigated by solving ill-posed inverse problems, similarly to super-resolution approach. Yet, each of them is slightly different, e.g. denoising focuses on restoration of a clean image from noisy inputs, deblurring aims at alleviating effect of convolution between sharp data and blurry kernel, etc. Therefore, we can treat all of those tasks as exclusive problems.

According to our studies, described in Section 5.3.4, conducted on collected low resolution (80x60) thermal dataset, Generative Adversarial Network (p2p-deblur) designed as image-to-image translation solution performing deblurring operation produced image quality metrics close to the proposed SR convolutional network. Potentially, restored high resolution outputs could be further improved by combining solutions for reversing different degradation algorithms in a single pipeline.

Various techniques for blur removal have been already proposed, including solutions based on image priors [267], conventional machine learning algorithms [268], and recently also Deep Neural Networks, e.g. Sun j. et al. [269] proposed to use a simple CNN structure to predict motion kernels. After that, Markov random field model was applied to estimate motion blur field and remove it using deblurring based on patch-level image priors. More recent studies showed that deblurring can be also solved with GAN [270], similarly to the p2p deblurring network evaluated in our work (see Chapter 5).

An interesting research question is what is the purpose of residual blocks in embedding sub-network learnt by the SR model. Those blocks may act as either image pre-processing, e.g. unsharp/deblur or as feature extraction operations. In the first case, unsharp masking may be realized by introduction of additional convolutional blocks, as presented in Fig. 7.1.

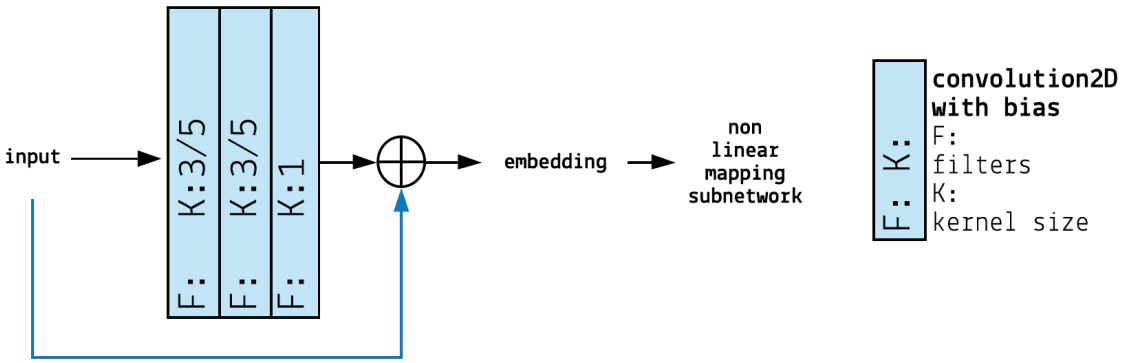


Figure 7.1. Unsharp operation can be realized by applying a sequence of convolution operations, where 1x1 convolution models unsharp masking kernel with weights learnt during model training instead of their manual selection as in a standard approach

It's worth noting that such structure is equivalent to a simple unsharp model (see Fig. 7.2) for which the output is restored as:

$$x_{out} = x_{in} + (x_{in} - x_{blurred}) \circledast K \tag{7.3}$$

where K is the kernel for unsharp masking, which can be modeled by convolution operation.

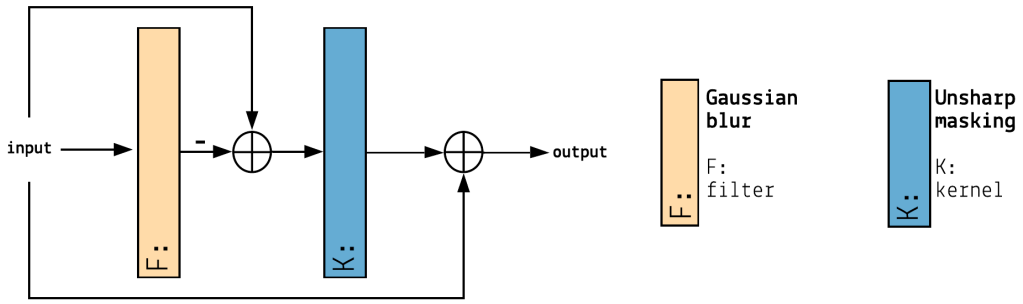


Figure 7.2. Steps of unsharp operation

Visualization of final weights of convolutional operations in the unsharp block may be useful for determining whether they in fact act as image pre-processing or as extraction of features. Such experiments may be also performed by analysing feature maps instead of image data, as shown in Fig. 7.3

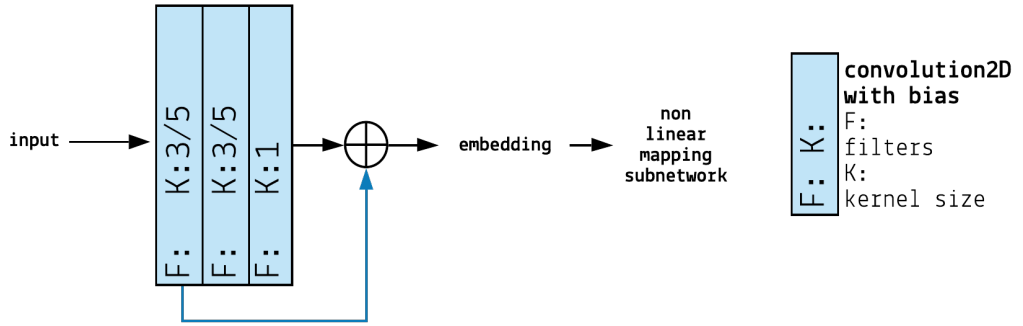


Figure 7.3. Unsharp operation applied to feature maps extracted after first convolution

On the other hand, if more experiments are conducted, it may turn out that the proposed DRESNet model is sufficient for reversing all discussed degradation operation at once. The reason for this assumption is that theoretically a universal DL model can learn a single function $SR(x)$ equally well as combination of functions $SR(DN(DB(x)))$, where SR denotes super resolution, DN denoising and DB deblurring applied to an image x. Yet, more tests should be performed to confirm this hypothesis. Especially, we would like to further reduce image quality by introducing additional noise and/or convolving images with blurry kernels. Data augmentation introduced in this way may allow for training universal thermal image enhancement solution, significantly improving image quality metrics.

As another aspect of future work on SR model architecture, we would also like to evaluate other SR networks on thermal data, e.g. Generative Adversarial Networks, which have recently gained a lot of attention due to achieving very promising results in various computer vision tasks, including super resolution of visible light images [196].

7.5 Other Potential Applications

In Chapter 6 a detailed analysis of exemplary remote medical diagnostic solutions based on the proposed DL ideas (i.e. transferring knowledge from visible light domain and enhancement of thermal images) was presented. Conducted experiments showed that accuracy of such solutions can be improved with AI. In this way, we proved that introduced neural network techniques and novel architectures are suitable for practical remote healthcare monitoring applications, what confirmed theses proposed in the dissertation.

We also identified a wide range of other potential use cases which could benefit from DL approaches described in this study. Although those applications are not the subject of this work, we would like to examine them in the future work. First of all, research on producing hallucinated visible light images is much more advanced than similar studies in thermal domain. Many artificial neural networks have been proved to restore very accurate versions of original high resolution data due to ability to learn high frequency representations of such images, as discussed in Section 5.2.3. Thus, generating super-resolved visible light sequences might also have a huge potential for remote diagnostics, e.g. non-contact heart rate extraction [44]. Initial evaluation of such solution has been discussed by McDuff D. [271]. We would like to verify it and compare with the network proposed by us for thermal image enhancement.

Secondly, evaluated applications may be also deployed in other markets, apart from telemedicine. A very interesting potential solution would be to utilize discussed algorithms for security purposes. Recent epidemic treats has caused a great commotion all over the world. Researchers are looking for novel health state evaluation techniques that would allow for making fast decisions about necessary treatment. By using image enhancement and vital signs magnification techniques, it is possible to estimate vital signs at longer distances, as showed in our previous work [169]. Taking it into account, proposed SR network could be deployed at border control and security checkpoints in order to improve accuracy of vital signs estimation or body temperature pattern analysis and reduce the risk of infection spreading by taking immediate medical actions.

Moreover, acquired and enhanced sequences could be also processed in order to recognize emergency situations or detect violent behaviours by e.g. performing action recognition [272]. In this case, examples of scenarios include fall detection applications deployed in smart homes [273], driver sleep alerts using different temperatures around the nose and mouth areas [19], crowd density classification in order to identify potential [274] security treats at airports and other public places, etc. The number of potential applications is countless and we believe that as long as image processing algorithms are used, techniques evaluated in this study and proposed neural networks could lead to improvement of the accuracy of such solutions. We would like to expand our research in this area and verify robustness of designed models for other potential applications.

7.6 Summary

This Chapter overviewed ideas for improving proposed AI algorithms in two ways: a) by modifying the introduced architecture using latest advances in DL; b) by utilization of other learning methods that enable more efficient search for optimal network architecture. Moreover, we also suggested other potential applications that could benefit from the use of the proposed AI model. Factors that should be considered in order to make the analysed solutions production ready were also presented and discussed.

Chapter 8

Conclusion

8.1 Summary

In the face of demographic transformations happening at a global scale, vision of healthy lifestyle and medical services is going to diverge from its current definition. This progress can be already observed in development of more advanced data processing systems and more intelligent devices equipped with algorithms capable of providing information about our state of health, such as wearable devices [150, 275] or kitchen appliances suggesting proper nutrition [222]. It has been also shown that contactless estimation of fundamental vital signs important for indicating various health problems is possible by analysis of pixel intensities changes in specific body regions [44, 235] even at long distances [169] if vital signs patterns are properly magnified and enhanced. Yet, based on conducted analysis of state-of-the-art solutions utilized for improvement of quality of collected data and detection of areas useful for contactless extraction of vital signs, we realized that the majority of studies focus on visible light data only. It is important to note though that images obtained in various domains have different characteristic and frequently solution developed for one imaging domain may not be directly applicable and transferable to other representations. Since thermal data is of interest in computer vision-based remote monitoring systems due to ability of providing additional medical information (temperature patterns can be used for pain analysis [18], sleep detection [19], evaluation of facial muscle paralysis [20] or respiratory rate estimation [21]), while being insensitive to different illumination condition and ensuring better data privacy, we believe there is a need of expanding research on such solutions to the thermal domain.

The presented doctoral dissertation focused on proposal of novel Deep Learning (DL) based solutions and Neural Network architectures that would improve state-of-the-art knowledge in the area of thermal image processing. Our main goal was to design algorithms that would lead to better accuracy of possible remote medial diagnostic applications. In-depth analysis of state-of-the-art architectures allowed for determining drawbacks of those networks for processing of thermal images which have different characteristic than visible light ones. In order to mitigate identified limitations, we proposed innovative modifications of existing solutions and novel Deep Neural Network (DNN) architectures more suitable for thermal images. To evaluate recommended approaches, datasets of thermal sequences were collected using different imaging sensors from more than 70 volunteers in total taking into account examples of target remote medical diagnostic applications, e.g. estimation of respiratory rate, emotion analysis studies, facial areas detection and deployment of proposed algorithms on wearable devices.

In addition, tests were also performed on publicly available datasets to avoid results being biased by our sets. Experiments conducted for all databases proved that introduced solutions outperformed algorithms used so far for solving the same tasks. Specifically, we modified the original flow of the Inception [111] model in a way that classification task was turned into detection during the inference time. Using the proposed solution, we observed a significant improvement of facial areas detection from low resolution thermal images in comparison to state-of-the-art network, i.e. precision improved by 8% and recall improved by 63%. Modification introduced during the prediction time allowed for eliminating the need of providing bounding box annotations and led to reduction of the processing time - ˜16FPS for a single image stream utilizing only ˜5% of the resources on NVIDIA® DGX-1™ Station vs. ˜2FPS for state-of-the-art SSD model.

It has also been showed that the proposed thermal image enhancement method lead to improvement of accuracy of existing algorithms applied for contactless vital signs estimation. According to achieved results, accuracy of respiratory rate estimation using super resolution thermal model is comparable or better than for results achieved with 4 times bigger inputs (Root Mean Square Error for inputs of a size 20x15 pixels, enhanced with proposed DRESNet equals 4.89 vs. 5.15 for original data and 5.58 for other magnification algorithms - Section 6.3.2). The same finding holds true for person recognition studies, as it has been proved that person identification accuracy can be improved by more than 8% for images as small as 60x80 if their quality is enhanced with the proposed neural network.

This opens a lot of new possibilities for modern healthcare systems which could benefit from DL-based image processing techniques without the need of providing better quality, more expensive sensors. Conducted experiments allowed for determining the configuration of DL models leading to the best image quality metrics by evaluating different number and placement of residual and recursive blocks within feature extraction, non-linear mapping and reconstruction subnetworks. Although presented results are very promising, we are also aware of some limitations of the introduced methods and discussed them in details in Chapter 7, providing ideas for future studies and improvements.

Based on performed analysis, we believe that the goal of the presented dissertation was achieved. Conclusions of conducted studies were summarized in each Section, underlying innovative results which proved theses formulated in the presented doctoral dissertation. Specifically, Sections 4.2.1, 4.2.2, and 4.3 allowed for justifying first thesis statement:

I) Architecture of Deep Neural Network designed for classification of visible light images can be modified in such a way that distribution of extracted features will be recreated enabling detection of facial areas from low resolution thermal data.

Sections 5.3, 5.3.2, 5.3.4, 5.3.4, 5.3.4, and 6.3.1 demonstrated correctness of the second thesis:

II) Proposed architecture of deep Convolutional Neural Networks allows for increasing resolution of thermal images leading to improvement of facial areas detection accuracy.

Finally, in sections 6.3.2 and 6.4.1 we also showed that other algorithms used in remote medical diagnostic solutions, i.e. contactless estimation of vital signs and face recognition, can also benefit from applying thermal image enhancement model introduced in experiments performed to prove the second thesis.

8.2 Novel Outcomes

Extensive benchmark evaluation conducted in the presented doctoral dissertation in order to prove, explain and elaborate on formulated theses resulted in following original, high-impact and cutting edge outcomes:

1. Limitations of existing image processing methods for facial areas detection in thermography were identified in the critical analysis of state of the art techniques (Chapter 2.

2. New databases that could be used for training of DNN for facial area detection were collected using different thermal acquisition devices (Section 4.2.1).

3. Convolutional Neural Networks and other novel DL architectures insensitive to body rotations were proved to accurately detect facial areas from thermal images (Sections 4.2.1 and 4.3).

4. Possibility of transferring knowledge from visible light data to thermal images in order to improve detection accuracy was verified using Deep Learning models (Section 4.2.1 and 6.3.1).

5. Innovative structure of Convolutional Neural Network leading to restoration of features distribution to determine coordinates of facial areas used for contactless vital sign estimation was proposed, implemented and evaluated. (Section 4.2.2); comparison of the proposed solution with state-of-the-art detection model showed improvement of performance and reduction of training time.

6. Novel DNN architecture of Super Resolution model for enhancing thermal images was proposed, designed and implemented. The introduced innovative architecture is a first attempt (to the best of our knowledge) to address thermal data characteristic by widening of the receptive field, to take into account more distant relations between facial components due to heat flow in objects. Comparison of the proposed neural network with other existing image enhancement DL models proved its superiority on different thermal datasets (Chapter 5).

7. Experimental analysis of proposed DL techniques and neural network structures showed their robustness in potential practical applications of remote medical diagnostics and other relevant use cases (Chapter 6).

Summing up, in our opinion the aim of the presented dissertation was achieved. Innovative methods of thermal image processing using Deep Neural Networks in order to enhance their quality were proposed. Evaluation performed with the introduced techniques proved the increase of facial areas detection accuracy, what is beneficial for the needs of remote medical diagnostic solutions. Theses formulated in the dissertation were confirmed, showing their authenticity and genuineness. Novel outcomes were achieved and published in a wide range of publications, which are summarized in Appendix A.

Bibliography

[1] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[2] James S Duncan, Michael F Insana, and Nicholas Ayache. Biomedical imaging and analysis in the age of big data and deep learning. *Proceedings of the IEEE*, 108(1):3–10, 2019.

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[4] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 2019.

[5] EY-K Ng, SC Fok, YC Peh, FC Ng, and LSJ Sim. Computerized detection of breast cancer with artificial intelligence and thermograms. *Journal of medical engineering & technology*, 26(4):152–157, 2002.

[6] Zhi-Hua Zhou, Yuan Jiang, Yu-Bin Yang, and Shi-Fu Chen. Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, 24(1):25–36, 2002.

[7] Dalia Wajeeh Abu Kashf, Ayah Nedal Okasha, Noor Ashraf Sahyoun, Roaa Emal El-Rabi, and Samy S Abu-Naser. Predicting dna lung cancer using artificial neural network. *International Journal of Academic Pedagogical Research (IJAPR)*, 2(10):6–13, 2018.

[8] Kyung-Joong Kim and Sung-Bae Cho. Prediction of colon cancer using an evolutionary neural network. *Neurocomputing*, 61:361–379, 2004.

[9] Junji Shiraishi, Qiang Li, Daniel Appelbaum, and Kunio Doi. Computer-aided diagnosis and artificial intelligence in clinical imaging. In *Seminars in nuclear medicine*, volume 41, pages 449–462. Elsevier, 2011.

[10] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[11] Lucas von Chamier, Romain F Laine, and Ricardo Henriques. Artificial intelligence for microscopy: what you should know. *Biochemical Society Transactions*, 47(4):1029–1040, 2019.

[12] Yuhua Chen, Yibin Xie, Zhengwei Zhou, Feng Shi, Anthony G Christodoulou, and Debiao Li. Brain mri super resolution using 3d deep densely connected neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 739–742. IEEE, 2018.

[13] Kensuke Umehara, Junko Ota, and Takayuki Ishida. Application of super-resolution convolutional neural network for enhancing image resolution in chest ct. *Journal of digital imaging*, 31(4):441–450, 2018.

[14] Abhimanyu S Ahuja. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.

[15] Regina E Herzlinger. Why innovation in health care is so hard. *Harvard business review*, 84(5):58, 2006.

[16] Matthew Collier, Richard Fu, Lucy Yin, and P Christiansen. Artificial intelligence: healthcare's new nervous system. *Viewable at https://. accenture. com/t20170418T023006Z__w__/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence. pdf*, 2017.

[17] Telemedicine market. `https://www.gminsights.com/industry-analysis/telemedicine-market`. Accessed: 2019-11-14.

[18] Marco Bellantonio, Mohammad A Haque, Pau Rodriguez, Kamal Nasrollahi, Taisi Telve, Sergio Escalera, Jordi Gonzalez, Thomas B Moeslund, Pejman Rasti, and Gholamreza Anbarjafari. Spatio-temporal pain recognition in cnn-based super-resolved facial images. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, pages 151–162. Springer, 2016.

[19] James Russell Clarke Sr and Phyllis Maurer Clarke. Sleep detection and driver alert apparatus, November 18 1997. US Patent 5,689,241.

[20] Shu He, John J Soraghan, and Brian F O'Reilly. Objective grading of facial paralysis using local binary patterns in video processing. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4805–4808. IEEE, 2008.

[21] Jacek Rumiński. Reliability of pulse measurements in videoplethysmography. *Metrology and Measurement Systems*, 23(3):359–371, 2016.

[22] Mritunjay Rai, Tanmoy Maity, and RK Yadav. Thermal imaging system and its real time applications: a survey. *Journal of Engineering Technology*, 6(2):290–303, 2017.

[23] Mahnaz EtehadTavakol, Vinod Chandran, EYK Ng, and Raheleh Kafieh. Breast cancer detection from thermal images using bispectral invariant features. *International Journal of Thermal Sciences*, 69:21–36, 2013.

[24] Leonardo Trujillo, Gustavo Olague, Riad Hammoud, and Benjamin Hernandez. Automatic feature localization in thermal images for facial expression recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 14–14. IEEE, 2005.

[25] Marcin Kopaczka, Jan Nestler, and Dorit Merhof. Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 518–529. Springer, 2017.

[26] Mariusz Marzec, Robert Koprowski, and Zygmunt Wróbel. Detection of selected face areas on thermograms with elimination of typical problems. *Journal of medical informatics & technologies*, 16, 2010.

[27] Mariusz Marzec, Robert Koprowski, Zygmunt Wróbel, Agnieszka Kleszcz, and Sławomir Wilczyński. Automatic method for detection of characteristic areas in thermal face images. *Multimedia Tools and Applications*, 74(12):4351–4368, 2015.

[28] Mariusz Marzec, Robert Koprowski, and Zygmunt Wróbel. Methods of face localization in thermograms. *Biocybernetics and Biomedical Engineering*, 35(2):138–146, 2015.

[29] Massimo Bertozzi, Alberto Broggi, Paolo Grisleri, Thorsten Graf, and Michael Meinecke. Pedestrian detection in infrared images. In *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No. 03TH8683)*, pages 662–667. IEEE, 2003.

[30] Weihong Wang, Jian Zhang, and Chunhua Shen. Improved human detection and classification in thermal images. In *2010 IEEE International Conference on Image Processing*, pages 2313–2316. IEEE, 2010.

[31] Kai Jungling and Michael Arens. Feature based person detection beyond the visible spectrum. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–37. IEEE, 2009.

[32] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[33] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[34] Robert Koprowski, Katarzyna Wojaczynska-Stanek, and Zygmunt Wrobel. Automatic segmentation of characteristic areas of the human head on thermographic images. *Machine Graphics & Vision International Journal*, 16(3):251–274, 2007.

[35] Mariusz Marzec, Aleksander Lamża, Zygmunt Wróbel, and Andrzej Dziech. Fast eye localization from thermal images using neural networks. *Multimedia Tools and Applications*, pages 1–14, 2016.

[36] Shan He, Shangfei Wang, Wuwei Lan, Huan Fu, and Qiang Ji. Facial expression recognition using deep boltzmann machine from thermal infrared images. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 239–244. IEEE, 2013.

[37] Vijay John, Seiichi Mita, Zheng Liu, and Bin Qi. Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 246–249. IEEE, 2015.

[38] Aly Metwaly, Jorge Peña Queralta, Victor Kathan Sarker, Tuan Nguyen Gia, Omar Nasir, and Tomi Westerlund. Edge computing with embedded ai: Thermal image analysis for occupancy estimation in intelligent buildings. *INTelligent Embedded Systems Architectures and Applications, INTESA@ ESWEEK*, 2019.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[40] R Abbott, JM Del Rincon, B Connor, and N Robertson. Deep object classification in low resolution lwir imagery via transfer learning. In *Proceedings of 5th IMA Conference on Mathematics in Defence*, volume 2, 2017.

[41] Alicja Kwaśniewska, Anna Giczewska, and Jacek Rumiński. Big data significance in remote medical diagnostics based on deep learning techniques. *Task Quarterly*, 21:309–319, 2017.

[42] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, pages 174–181. IEEE, 2018.

[43] James W Davis and Vinay Sharma. Robust background-subtraction for person detection in thermal imagery. In *CVPR Workshops*, page 128, 2004.

[44] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011.

[45] Deng-Yuan Huang, Ta-Wei Lin, Chun-Ying Ho, and Wu-Chih Hu. Face detection based on feature analysis and edge detection against skin color-like backgrounds. In *2010 Fourth International Conference on Genetic and Evolutionary Computing*, pages 687–690. IEEE, 2010.

[46] Anamika Singh, Manminder Singh, and Birmohan Singh. Face detection and eyes extraction using sobel edge detection and morphological operations. In *2016 Conference on Advances in Signal Processing (CASP)*, pages 295–300. IEEE, 2016.

[47] Kayvan Najarian and Robert Splinter. *Biomedical signal and image processing*. CRC press, 2005.

[48] Farah Q Al-Khalidi, Reza Saatchi, Derek Burke, and Heather Elphick. Facial tracking method for noncontact respiration rate monitoring. In *2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*, pages 751–754. IEEE, 2010.

[49] Abdulbasit AlAZZAWI, Osman Nuri Uçan, and Oğus Bayat. Face recognition based on multi features extractors. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, 2017.

[50] Peng Zhao-Yi, Zhu Yan-Hui, and Zhou Yu. Real-time facial expression recognition based on adaptive canny operator edge detection. In *2010 Second International Conference on MultiMedia and Information Technology*, volume 2, pages 154–157. IEEE, 2010.

[51] Kar-Kin Lee, Wai-Kuen Cham, and Qin-Ran Chen. Chin contour estimation using modified canny edge detector. In *7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002.*, volume 2, pages 770–775. IEEE, 2002.

[52] Thomas C Chang, Thomas S Huang, and Carol Novak. Facial feature extraction from color images. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 39–43. IEEE, 1994.

[53] Rainer Herpers, Markus Michaelis, K-H Lichtenauer, and Gerald Sommer. Edge and keypoint detection in facial regions. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 212–217. IEEE, 1996.

[54] Jian Huang Lai, Pong C Yuen, Wen Sheng Chen, Shihong Lao, and Masato Kawade. Robust facial feature point detection under nonlinear illuminations. In *Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 168–174. IEEE, 2001.

[55] Taro Yokoyama, Haiyuan Wu, and Masahiko Yachida. Automatic detection of facial feature points and contours. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*, pages 335–340. IEEE, 1996.

[56] Tumpa Dey and Tamojay Deb. Facial landmark detection using fast corner detector of ugc-ddmc face database of tripura tribes. In *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, pages 1–4. IEEE, 2015.

[57] IEEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under grant R01-1344-18; FAA/NSSA grant R01-1344-48/49; Office of Naval Research under grant N000143010022.

[58] Alicja Kwaśniewska and Jacek Rumiński. Real-time facial feature tracking in poor quality thermal imagery. In *2016 9th International Conference on Human System Interactions (HSI)*, pages 504–510. IEEE, 2016.

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

[60] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[61] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer, 2011.

[62] Mark Asbach, Peter Hosten, and Michael Unger. An evaluation of local features for face detection and localization. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 32–35. IEEE, 2008.

[63] Mehmet Sezgin and Bülent Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–166, 2004.

[64] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[65] Tanuj N Palghamol and Shilpa P Metkar. Flexible luminance thresholding for detecting eyes in color images. In *Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC)*, pages 568–572. IEEE.

[66] Imran Naseem and Mohamed Deriche. Robust human face detection in complex color images. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–338. IEEE, 2005.

[67] Madhusudhana Gargesha and Sethuraman Panchanathan. Face detection from color images by iterative thresholding on skin probability maps. In *2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No. 02CH37353)*, volume 5, pages V–V. IEEE, 2002.

[68] Alicja Kwaśniewska and Jacek Rumiński. Face detection in image sequences using a portable thermal camera. In *Proceedings of the 13th Quantitative Infrared Thermography Conference*, 2016.

[69] Jianping Fan, David KY Yau, Ahmed K Elmagarmid, and Walid G Aref. Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE transactions on image processing*, 10(10):1454–1466, 2001.

[70] Yufeng Zheng. Face detection and eyeglasses detection for thermal face recognition. In *Image Processing: Machine Vision Applications V*, volume 8300, page 83000C. International Society for Optics and Photonics, 2012.

[71] Yuen Kiat Cheong, Vooi Voon Yap, and Humaira Nisar. A novel face detection algorithm using thermal imaging. In *2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pages 208–213. IEEE, 2014.

[72] Hossein Ebrahimpour-Komleh, Vinod Chandran, and Sridha Sridharan. Face recognition using fractal codes. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pages 58–61. IEEE, 2001.

[73] Julian Supardi, Abdiansah Ns, and Nys Anditha. Watershed segmentation for face detection using artificial intelligence neural network. *International Conference on Computer Science and Engineering*, 10 2014.

[74] James W Davis and Vinay Sharma. Background-subtraction in thermal imagery using contour saliency. *International Journal of Computer Vision*, 71(2):161–181, 2007.

[75] Yongsheng Gao and Maylor KH Leung. Face recognition using line edge map. *IEEE transactions on pattern analysis and machine intelligence*, 24(6):764–779, 2002.

[76] Yasufumi Suzuki and Tadashi Shibata. Multiple-clue face detection algorithm using edge-based feature vectors. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–737. IEEE, 2004.

[77] Haiyuan Wu, Qian Chen, and Masahiko Yachida. Face detection from color images using a fuzzy pattern matching method. *IEEE transactions on pattern analysis and machine intelligence*, 21(6):557–563, 1999.

[78] Sanun Srisuk and Werasak Kurutach. A new robust face detection in color images. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 306–311. IEEE, 2002.

[79] Debotosh Bhattacharjee, Ayan Seal, Suranjan Ganguly, Mita Nasipuri, and Dipak Kumar Basu. Comparative study of human thermal face recognition based on haar wavelet transform and local binary pattern. *Computational intelligence and neuroscience*, 2012:6, 2012.

[80] Xiaohua Qian, Jiahui Wang, Shuxu Guo, and Qiang Li. An active contour model for medical image segmentation with application to brain ct image. *Medical physics*, 40(2):021911, 2013.

[81] Haiyuan Wu, Taro Yokoyama, Dadet Pramadihanto, and Masahiko Yachida. Face and facial feature extraction from color image. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pages 345–350. IEEE, 1996.

[82] Maurício Pamplona Segundo, Luciano Silva, Olga Regina Pereira Bellon, and Chauã C Queirolo. Automatic face segmentation and facial landmark detection in range images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5):1319–1330, 2010.

[83] David Crespo, Carlos M Travieso, and Jesús B Alonso. Thermal face verification based on scale-invariant feature transform and vocabulary tree. 2012.

[84] Ranjana Sikarwar, Arun Agrawal, and Rajendra Singh Kushwah. An edge based efficient method of face detection and feature extraction. In *2015 Fifth International Conference on Communication Systems and Network Technologies*, pages 1147–1151. IEEE, 2015.

[85] Minh-Tri Pham, Yang Gao, Viet-Dung D Hoang, and Tat-Jen Cham. Fast polygonal integration and its application in extending haar-like features to improve object detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 942–949. IEEE, 2010.

[86] Arwa M Basbrain, John Q Gan, and Adrian Clark. Accuracy enhancement of the viola-jones algorithm for thermal face detection. In *International Conference on Intelligent Computing*, pages 71–82. Springer, 2017.

[87] Md Omar Faruqe and Md Al Mehedi Hasan. Face recognition using pca and svm. In *2009 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication*, pages 97–101. IEEE, 2009.

[88] Harihara Santosh Dadi and Gopala Krishna Mohan Pillutla. Improved face recognition rate using hog features and svm classifier. *IOSR Journal of Electronics and Communication Engineering*, 11(4):34–44, 2016.

[89] Ergun Gumus, Niyazi Kilic, Ahmet Sertbas, and Osman N Ucan. Evaluation of face recognition techniques using pca, wavelets and svm. *Expert Systems with Applications*, 37(9):6404–6408, 2010.

[90] 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.

[91] Ramiro Donoso Floody, César San Martín, and Heydi Méndez-Vázquez. Face recognition using tof, lbp and svm in thermal infrared images. In *Iberoamerican Congress on Pattern Recognition*, pages 683–691. Springer, 2011.

[92] Shangfei Wang, Zhilei Liu, Peijia Shen, and Qiang Ji. Eye localization from thermal infrared images. *Pattern Recognition*, 46(10):2613–2621, 2013.

[93] Brais Martinez, Xavier Binefa, and Maja Pantic. Facial component detection in thermal imagery. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 48–54. IEEE, 2010.

[94] Wang Xiaoyu, Chen Jihong, Wang Pingjiang, and Huang Zhihong. Infrared human face auto locating based on svm and a smart thermal biometrics system. In *Sixth International Conference on Intelligent Systems Design and Applications*, volume 2, pages 1066–1072. IEEE, 2006.

[95] Mrinal Kanti Bhowmik, Barin Kumar De, Debotosh Bhattacharjee, Dipak Kumar Basu, and Mita Nasipuri. Multisensor fusion of visual and thermal images for human face identification using different svm kernels. In *2012 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pages 1–7. IEEE, 2012.

[96] Pallabi Parveen and Bhavani Thuraisingham. Face recognition using multiple classifiers. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 179–186. IEEE, 2006.

[97] J Prabin Jose, P Poornima, and Kukkapalli Manoj Kumar. A novel method for color face recognition using knn classifier. In *2012 International Conference on Computing, Communication and Applications*, pages 1–3. IEEE, 2012.

[98] Yufeng Zheng, Adel S Elmaghraby, and Kristopher Reese. Performance improvement of face recognition using multispectral images and stereo images. In *2012 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 000280–000285. IEEE, 2012.

[99] Shangfei Wang, Peijia Shen, and Zhilei Liu. Facial expression recognition from infrared thermal images using temperature difference by voting. In *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, volume 1, pages 94–98. IEEE, 2012.

[100] Shangfei Wang and Zhilei Liu. Infrared face recognition based on histogram and k-nearest neighbor classification. In *International Symposium on Neural Networks*, pages 104–111. Springer, 2010.

[101] George A Papakostas, Vassilis G Kaburlasos, and Th Pachidis. Thermal infrared face recognition based on lattice computing (lc) techniques. In *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2013.

[102] Chan Su Park and Hi Seok Kim. Color enhancement algorithm using the integrated k-means clustering and inverted otsu method for thermal object characterization. *International Journal of Intelligent Engineering & Systems*, 11(4).

[103] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color and edge information. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 1, pages 525–528. IEEE, 2003.

[104] Duy Nguyen, David Halupka, Parham Aarabi, and Ali Sheikholeslami. Real-time face detection and lip feature extraction using field-programmable gate arrays. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):902–912, 2006.

[105] Pradeep Buddharaju, Ioannis T Pavlidis, Panagiotis Tsiamyrtzis, and Mike Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):613–626, 2007.

[106] Christopher Gordon, Mark Acosta, Nathan Short, Shuowen Hu, and Alex L Chan. Toward automated face detection in thermal and polarimetric thermal imagery. In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXV*, volume 9842, page 984212. International Society for Optics and Photonics, 2016.

[107] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Linear regression for face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):2106–2112, 2010.

[108] Ayan Seal, Suranjan Ganguly, Debotosh Bhattacharjee, Mita Nasipuri, and Dipak Kumar Basu. A comparative study of human thermal face recognition based on haar wavelet transform (hwt) and local binary pattern (lbp). *arXiv preprint arXiv:1309.1009*, 2013.

[109] 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.

[110] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[111] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[112] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[113] 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.

[114] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[115] Yuda Song, Yunfang Zhu, and Xin Du. Dynamic residual dense network for image denoising. *Sensors*, 19(17):3809, 2019.

[116] Margarita Osadchy, Yann Le Cun, and Matthew L Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, 8(May):1197–1215, 2007.

[117] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[118] Yuxin Jiang, Songbin Li, Peng Liu, and Qiongxing Dai. Multi-feature deep learning for face gender recognition. In *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, pages 507–511. IEEE, 2014.

[119] Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. 2010.

[120] Brian Cheung. Convolutional neural networks applied to human face classification. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 580–583. IEEE, 2012.

[121] Ricky Anderson, Aryo Pradipta Gema, Sani M Isa, et al. Facial attractiveness classification using deep learning. In *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, pages 34–38. IEEE, 2018.

[122] Marc Oliu Simón, Ciprian Corneanu, Kamal Nasrollahi, Olegs Nikisins, Sergio Escalera, Yunlian Sun, Haiqing Li, Zhenan Sun, Thomas B Moeslund, and Modris Greitans. Improved rgb-dt based face recognition. *Iet Biometrics*, 5(4):297–303, 2016.

[123] Min Peng, Chongyang Wang, Tong Chen, and Guangyuan Liu. Nirfacenet: A convolutional neural network for near-infrared face identification. *Information*, 7(4):61, 2016.

[124] Jongwoo Seo and In-Jeong Chung. Face liveness detection using thermal face-cnn with external knowledge. *Symmetry*, 11(3):360, 2019.

[125] Seyed Mehdi Iranmanesh, Ali Dabouei, Hadi Kazemi, and Nasser M Nasrabadi. Deep cross polarimetric thermal-to-visible face recognition. In *2018 International Conference on Biometrics (ICB)*, pages 166–173. IEEE, 2018.

[126] Alperen Kantarcı and Hazım Kemal Ekenel. Thermal to visible face recognition using deep autoencoders. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2019.

[127] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for thermal to visible face recognition. *arXiv preprint arXiv:1507.02879*, 2015.

[128] Artur Grudzien, Marcin Kowalski, and Norbert Palka. Face re-identification in thermal infrared spectrum based on thermalfacenet neural network. In *2018 22nd International Microwave and Radar Conference (MIKON)*, pages 179–180. IEEE, 2018.

[129] Zhan Wu, Min Peng, and Tong Chen. Thermal face recognition using convolutional neural network. In *2016 International Conference on Optoelectronics and Image Processing (ICOIP)*, pages 6–9. IEEE, 2016.

[130] O Obi-Alago, SN Yanushkevich, and HM Wetherley. Detecting thermal face signature abnormalities. In *2019 Eighth International Conference on Emerging Security Technologies (EST)*, pages 1–6. IEEE, 2019.

[131] Sooraj Menon, J Swathi, SK Anit, Anu P Nair, and S Sarath. Driver face recognition and sober drunk classification using thermal images. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0400–0404. IEEE, 2019.

[132] Hemang M Shah, Aadhithya Dinesh, and T Sree Sharmila. Analysis of facial landmark features to determine the best subset for finding face orientation. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–4. IEEE, 2019.

[133] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[134] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38, 1998.

[135] Ross Girshick. Fast r-cnn object detection with caffe. *Microsoft Research*, 2015.

[136] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[137] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[138] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[139] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[140] Domenick Poster, Shuowen Hu, Nasser Nasrabadi, and Benjamin Riggan. An examination of deep-learning based landmark detection methods on thermal face imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[141] Alicja Kwaśniewska, Jacek Rumiński, Krzysztof Czuszyński, and Maciej Szankin. Real-time facial features detection from low resolution thermal images with deep classification models. *Journal of Medical Imaging and Health Informatics*, 8(5):979–987, 2018.

[142] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[143] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[144] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[145] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[146] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[147] Charles Kornreich and Pierre Philippot. Dysfunctions of facial emotion recognition in adult neuropsychiatric disorders: Influence on interpersonal difficulties. *Psychologica belgica*, 46(1-2), 2006.

[148] Jacek Ruminski, Adam Bujnowski, Krzysztof Czuszynski, and Tomasz Kocejko. Estimation of respiration rate using an accelerometer and thermal camera in eglasses. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1431–1434. IEEE, 2016.

[149] Mariusz Kaczmarek, Adam Bujnowski, Jerzy Wtorek, and A Polinski. Multimodal platform for continuous monitoring of the elderly and disabled. *Journal of Medical Imaging and Health Informatics*, 2(1):56–63, 2012.

[150] Roderick McCall, Nicolas Louveton, and Jacek Rumiński. D2. 1 the specification and overall requirements of the eglasses platform. Technical report, University of Luxembourg, 2014.

[151] Alicja Kwaśniewska, Jacek Rumiński, and Jerzy Wtorek. The motion influence on respiration rate estimation from low-resolution thermal sequences during attention focusing tasks. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1421–1424. IEEE, 2017.

[152] Jacek Ruminski, Adam Bujnowski, Jerzy Wtorek, Aliaksei Andrushevich, Martin Biallas, and Rolf Kistler. Interactions with recognized objects. In *2014 7th International Conference on Human System Interactions (HSI)*, pages 101–105. IEEE, 2014.

[153] Alicja Kwaśniewska, Joanna Klimiuk-Myszk, Jacek Ruminski, Jérôme Forrier, Benoît Martin, and Isabelle Pecci. Quality of graphical markers for the needs of eyewear devices. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 388–395. IEEE, 2015.

[154] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[155] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[156] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[157] Jacek Ruminski. The accuracy of pulse rate estimation from the sequence of face images. In *2016 9th International Conference on Human System Interactions (HSI)*, pages 518–524. IEEE, 2016.

[158] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Krzysztof Czuszynski. Remote estimation of video-based vital signs in emotion invocation studies. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4872–4876. IEEE, 2018.

[159] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.

[160] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646, 1996.

[161] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[162] Georg Wimmer, Andreas Vécsei, and Andreas Uhl. Cnn transfer learning for the automated diagnosis of celiac disease. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[163] Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE journal of biomedical and health informatics*, 21(1):76–84, 2016.

[164] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[165] Maciej Szankin, Alicja Kwaśniewska, Jacek Ruminski, and Rey Nicolas. Road condition evaluation using fusion of multiple deep models on always-on vision processor. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pages 3273–3279. IEEE, 2018.

[166] Jerzy Wtorek, Adam Bujnowski, Jacek Rumiński, Artur Poliński, Mariusz Kaczmarek, and Antoni Nowakowski. Assessment of cardiovascular risk in assisted living. *Metrology and measurement systems*, 19(2):231–244, 2012.

[167] Alicja Kwaśniewska, Jacek Rumiński, and Paul Rad. Deep features class activation map for thermal face detection and tracking. In *2017 10th International Conference on Human System Interactions (HSI)*, pages 41–47. IEEE, 2017.

[168] Bao Lei, Rene Klein Gunnewiek, and Peter HN De With. Reuse of motion processing for camera stabilization and video coding. In *2006 IEEE International Conference on Multimedia and Expo*, pages 597–600. IEEE, 2006.

[169] Maciej Szankin, Alicja Kwasniewska, Tejaswini Sirlapu, Mingshan Wang, Jacek Ruminski, Rey Nicolas, and Marko Bartscherer. Long distance vital signs monitoring with person identification for smart home solutions. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1558–1561. IEEE, 2018.

[170] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[171] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14:8, 2012.

[172] Jacek Ruminski and Alicja Kwasniewska. Evaluation of respiration rate using thermal imaging in mobile conditions. In *Application of Infrared to Biomedical Sciences*, pages 311–346. Springer, 2017.

[173] Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650. ACM, 2015.

[174] Michael D Breitenstein, Daniel Kuettel, Thibaut Weise, Luc Van Gool, and Hanspeter Pfister. Real-time face pose estimation from single range images. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[175] Reza Shoja Ghiass, Ognjen Arandjelović, and Denis Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor. In *Proceedings of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, pages 25–34. ACM, 2015.

[176] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[177] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

[178] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE transactions on Image Processing*, 15(8):2226–2238, 2006.

[179] Yaniv Romano, Matan Protter, and Michael Elad. Single image interpolation via adaptive nonlocal sparsity-based modeling. *IEEE Transactions on Image Processing*, 23(7):3085–3098, 2014.

[180] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

[181] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[182] Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, Osama A Omer, and Domenec Puig. Breast tumor classification in ultrasound images using texture analysis and super-resolution methods. *Engineering Applications of Artificial Intelligence*, 59:84–92, 2017.

[183] Yunxing Gao, Hengjian Li, Jiwen Dong, and Guang Feng. A deep convolutional network for medical image super-resolution. In *2017 Chinese Automation Congress (CAC)*, pages 5310–5315. IEEE, 2017.

[184] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In *Advances in neural information processing systems*, pages 769–776, 2009.

[185] David Eigen, Dilip Krishnan, and Rob Fergus. Restoring an image taken through a window covered with dirt or rain. In *Proceedings of the IEEE international conference on computer vision*, pages 633–640, 2013.

[186] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *2012 IEEE conference on computer vision and pattern recognition*, pages 2392–2399. IEEE, 2012.

[187] Shuangteng Zhang and Ezzatollah Salari. Image denoising using a neural network based non-linear filter in wavelet domain. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–989. IEEE, 2005.

[188] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019.

[189] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen. Deep network cascade for image super-resolution. In *European Conference on Computer Vision*, pages 49–64. Springer, 2014.

[190] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[191] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[192] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[193] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[194] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[195] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Tao Zhou, Yong Xia, and Dinggang Shen. Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's

disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 455–463. Springer, 2018.

[196] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[197] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[198] Yutian Zhang, Xiaohua Li, and Jiliu Zhou. Sftgan: a generative adversarial network for pan-sharpening equipped with spatial feature transform layers. *Journal of Applied Remote Sensing*, 13(2):026507, 2019.

[199] Tom Brosch and Roger Tam. Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2d and 3d images. *Neural computation*, 27(1):211–227, 2015.

[200] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5439–5448, 2017.

[201] Antigoni Panagiotopoulou and Vassilis Anastassopoulos. Super-resolution reconstruction of thermal infrared images. In *Proceedings of the 4th WSEAS International Conference on REMOTE SENSING*, 2008.

[202] Y Kiran, V Shrinidhi, W Jino Hans, and N Venkateswaran. A single-image super-resolution algorithm for infrared thermal images. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY*, 17(10):256–261, 2017.

[203] Emanuele Mandanici, Luca Tavasci, Francesco Corsini, and Stefano Gandolfi. A multi-image super-resolution algorithm applied to thermal imagery. *Applied Geomatics*, 11(3):215–228, 2019.

[204] Feras Almasri and Olivier Debeir. Multimodal sensor fusion in single thermal image super-resolution. In *Asian Conference on Computer Vision*, pages 418–433. Springer, 2018.

[205] Xiaohui Chen, Guangtao Zhai, Jia Wang, Chunjia Hu, and Yuanchun Chen. Color guided thermal image super resolution. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2016.

[206] Xudong Zhang, Chunlai Li, Qingpeng Meng, Shijie Liu, Yue Zhang, and Jianyu Wang. Infrared image super resolution by combining compressive sensing and deep learning. *Sensors*, 18(8):2587, 2018.

[207] Xiaodong Kuang, Xiubao Sui, Yuan Liu, Qian Chen, and Guohua Gu. Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing*, 332:119–128, 2019.

[208] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Mariusz Kaczmarek. Super-resolved thermal imagery for high-accuracy facial areas detection and analysis. *Engineering Applications of Artificial Intelligence*, 87:103263, 2020.

[209] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[210] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[211] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[212] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[213] Alicja Kwasniewska, Jacek Ruminski, and Maciej Szankin. Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. *Applied Sciences*, 9(20):4405, 2019.

[214] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[215] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[216] Maciej Szankin, Alicja Kwasniewska, and Jacek Ruminski. Influence of thermal imagery resolution on accuracy of deep learning based face recognition. In *2019 12th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2019.

[217] Cesare Massone, Rainer Hofmann-Wellenhof, Verena Ahlgrimm-Siess, Gerald Gabler, Christoph Ebner, and H Peter Soyer. Melanoma screening with cellular phones. *PloS one*, 2(5), 2007.

[218] Carol A Kilmon, Leonard Brown, Sumit Ghosh, and Artur Mikitiuk. Immersive virtual reality simulations in nursing education. *Nursing education perspectives*, 31(5):314–317, 2010.

[219] Christos D Katsis, George Ganiatsas, and Dimitrios I Fotiadis. An integrated telemedicine platform for the assessment of affective physiological states. *Diagnostic pathology*, 1(1):16, 2006.

[220] M. Christofi and D. Michael-Grigoriou. Virtual environments design assessment for the treatment of claustrophobia. In *2016 22nd International Conference on Virtual System Multimedia (VSMM)*, pages 1–8, 2016.

[221] Wanhong Wu, Jiannong Cao, Yuan Zheng, and Yong-Ping Zheng. Waiter: A wearable personal healthcare and emergency aid system. In *2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 680–685. IEEE, 2008.

[222] Suhuai Luo, Jesse S Jin, Jiaming Li, et al. A smart fridge with an ability to enhance health and enable better nutrition. *International Journal of Multimedia and Ubiquitous Engineering*, 4(2):69–80, 2009.

[223] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012.

[224] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, 2013.

[225] Taishi Sawabe, Tomoya Okajima, Masayuki Kanbara, and Norihiro Hagita. Evaluating passenger characteristics for ride comfort in autonomous wheelchairs. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 102–107. IEEE, 2017.

[226] Marjorie Skubic, Gregory Alexander, Mihail Popescu, Marilyn Rantz, and James Keller. A smart home application to eldercare: Current status and lessons learned. *Technology and Health Care*, 17(3):183–201, 2009.

[227] Chuantao Li, Fuming Chen, Jingxi Jin, Hao Lv, Sheng Li, Guohua Lu, and Jianqi Wang. A method for remotely sensing vital signs of human subjects outdoors. *Sensors*, 15(7):14830–14844, 2015.

[228] Natalia V Rivera, Swaroop Venkatesh, Chris Anderson, and R Michael Buehrer. Multi-target estimation of heart and respiration rates using ultra wideband sensors. In *2006 14th European Signal Processing Conference*, pages 1–6. IEEE, 2006.

[229] 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.

[230] Youngjun Cho, Simon J Julier, and Nadia Bianchi-Berthouze. Instant stress: Detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR mental health*, 6(4):e10140, 2019.

[231] Jayasimha N Murthy, Johan Van Jaarsveld, Jin Fei, Ioannis Pavlidis, Rajesh I Harrykissoon, Joseph F Lucke, Saadia Faiz, and Richard J Castriotta. Thermal infrared imaging: a novel method to monitor airflow during polysomnography. *Sleep*, 32(11):1521–1527, 2009.

[232] Jin Fei, Zhen Zhu, and Ioannis Pavlidis. Imaging breathing rate in the co 2 absorption band. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 700–705. IEEE, 2006.

[233] Ramya Murthy and Ioannis Pavlidis. Non-contact monitoring of breathing function using infrared imaging. Technical report, Technical Report Number UH-CS-05-09, 9 April. Department of Computer Science, 2005.

[234] Jin Fei and Ioannis Pavlidis. Thermistor at a distance: unobtrusive measurement of breathing. *IEEE Transactions on Biomedical Engineering*, 57(4):988–998, 2009.

[235] Jacek Rumiński. Evaluation of the respiration rate and pattern using a portable thermal camera. In *Proc. Of the 13th Quantitative Infrared Thermography Conference*, 2016.

[236] Yan Zhou, Panagiotis Tsiamyrtzis, Peggy Lindner, Ilya Timofeyev, and Ioannis Pavlidis. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5):1280–1289, 2012.

[237] Farah Q Al-Khalidi, Reza Saatchi, Derek Burke, and Heather Elphick. Tracking human face features in thermal images for respiration monitoring. In *ACS/IEEE International Conference on Computer Systems and Applications-AICCSA 2010*, pages 1–6. IEEE, 2010.

[238] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[239] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[240] Lara G Villanueva, Gustavo M Callicó, Félix Tobajas, Sebastián López, Valentín De Armas, José F López, and Roberto Sarmiento. Medical diagnosis improvement through image quality enhancement based on super-resolution. In *2010 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools*, pages 259–262. IEEE, 2010.

[241] José V Manjón, Pierrick Coupé, Antonio Buades, D Louis Collins, and Montserrat Robles. Mri superresolution using self-similarity and image priors. *International journal of biomedical imaging*, 2010, 2010.

[242] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018.

[243] Serajeddin Ebrahimian Hadi Kiashari, Ali Nahvi, Amirhossein Homayounfard, and Hamidreza Bakhoda. Monitoring the variation in driver respiration rate from wakefulness to drowsiness: a non-intrusive method for drowsiness detection using thermal imaging. *Journal of Sleep Sciences*, 3(1-2):1–9, 2018.

[244] Haitao Wang, Stan Z Li, and Yangsheng Wang. Face recognition under varying lighting conditions using self quotient image. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 819–824. IEEE, 2004.

[245] Fuzhen Huang and Houqin Bian. Identity authentication system using face recognition techniques in human-computer interaction. In *Proceedings of the 32nd Chinese Control Conference*, pages 3823–3827. IEEE, 2013.

[246] J Birgitta Martinkauppi, Maricor N Soriano, and Mika V Laaksonen. Behavior of skin color under varying illumination seen by different cameras at different color spaces. In *Machine Vision Applications in Industrial Inspection IX*, volume 4301, pages 102–112. International Society for Optics and Photonics, 2001.

[247] Alicja Kwasniewska, Maciej Szankin, Mateusz Ozga, Jason Wolfe, Arun Das, Adam Zajac, Jacek Ruminski, and Paul Rad. Deep learning optimization for edge devices: Analysis of training quantization parameters. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, volume 1, pages 96–101. IEEE, 2019.

[248] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[249] Fang Zhao, Meng Li, Yi Qian, and Joe Z Tsien. Remote measurements of heart and respiration rates for telemedicine. *PloS one*, 8(10), 2013.

[250] Alex D Torres, Hao Yan, Armin Haj Aboutalebi, Arun Das, Lide Duan, and Paul Rad. Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. In *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*, pages 61–89. Elsevier, 2018.

[251] Curtis Padgett and Garrison W Cottrell. Representing face images for emotion classification. In *Advances in neural information processing systems*, pages 894–900, 1997.

[252] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013.

[253] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. Emotion recognition from thermal infrared images using deep boltzmann machine. *Frontiers of Computer Science*, 8(4):609–618, 2014.

[254] Zhilei Liu and Shangfei Wang. Emotion recognition using hidden markov models from facial temperature sequence. In *International Conference on Affective Computing and Intelligent Interaction*, pages 240–247. Springer, 2011.

[255] Abhiram Kolli, Alireza Fasih, Fadi Al Machot, and Kyandoghere Kyamakya. Non-intrusive car driver's emotion recognition using thermal camera. In *Proceedings of the Joint INDS'11 & ISTET'11*, pages 1–5. IEEE, 2011.

[256] Barbara Manini, Daniela Cardone, Sjoerd Ebisch, Daniela Bafunno, Tiziana Aureli, and Arcangelo Merla. Mom feels what her child feels: thermal signatures of vicarious autonomic response while watching children in a stressful situation. *Frontiers in human neuroscience*, 7:299, 2013.

[257] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140:93–110, 2017.

[258] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 274–277. IEEE, 2016.

[259] Kazuhiko Takahashi. Remarks on computational emotion recognition from vital information. In *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*, pages 299–304. IEEE, 2009.

[260] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[261] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[262] Dario Floreano, Peter Dürr, and Claudio Mattiussi. Neuroevolution: from architectures to learning. *Evolutionary intelligence*, 1(1):47–62, 2008.

[263] Phillip Verbancsics and Josh Harguess. Image classification using generative neuro evolution for deep learning. In *2015 IEEE winter conference on applications of computer vision*, pages 488–493. IEEE, 2015.

[264] David J Montana and Lawrence Davis. Training feedforward neural networks using genetic algorithms. In *IJCAI*, volume 89, pages 762–767, 1989.

[265] Joseba Urzelai and Dario Floreano. Evolution of adaptive synapses: Robots with fast adaptive behavior in new environments. *Evolutionary computation*, 9(4):495–524, 2001.

[266] Richard K Belew, John McInerney, and Nicol N Schraudolph. Evolving networks: Using the genetic algorithm with connectionist learning. In *In*. Citeseer, 1990.

[267] Meiguang Jin, Stefan Roth, and Paolo Favaro. Normalized blind deconvolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–684, 2018.

[268] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794. 2006.

[269] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015.

[270] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE international conference on computer vision*, pages 251–260, 2017.

[271] Daniel McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018.

[272] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[273] Georgios Mastorakis and Dimitrios Makris. Fall detection system using kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9(4):635–646, 2014.

[274] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2017.

[275] Subramaniam Venkatraman and Shelten Gee Jao Yuen. Wearable heart rate monitor, April 7 2015. US Patent 8,998,815.

List of Figures

List of Tables

Appendix A

Scientific Work

A.1 Publications in the Dissertation Area

[1] Alicja Kwaśniewska, Anna Giczewska, and Jacek Rumiński. Big data significance in remote medical diagnostics based on deep learning techniques. *Task Quarterly*, 21:309–319, 2017.

[2] Alicja Kwaśniewska, Anna Giczewska, and Jacek Rumiński. Duże zbiory danych w zdalnej diagnostyce medycznej z wykorzystaniem technik głębokiego uczenia. In *VIII Konferencja Naukowa Infobazy, Gdansk Poland*, 2017.

[3] Alicja Kwaśniewska, Joanna Klimiuk-Myszk, Jacek Ruminski, Jérôme Forrier, Benoît Martin, and Isabelle Pecci. Quality of graphical markers for the needs of eyewear devices. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 388–395. IEEE, 2015.

[4] Alicja Kwaśniewska and Jacek Rumiński. Face detection in image sequences using a portable thermal camera. In *Proceedings of the 13th Quantitative Infrared Thermography Conference*, 2016.

[5] Alicja Kwaśniewska and Jacek Rumiński. Real-time facial feature tracking in poor quality thermal imagery. In *2016 9th International Conference on Human System Interactions (HSI)*, pages 504–510. IEEE, 2016.

[6] Alicja Kwaśniewska, Jacek Rumiński, Krzysztof Czuszyński, and Maciej Szankin. Real-time facial features detection from low resolution thermal images with deep classification models. *Journal of Medical Imaging and Health Informatics*, 8(5):979–987, 2018.

[7] Alicja Kwaśniewska, Jacek Rumiński, and Paul Rad. Deep features class activation map for thermal face detection and tracking. In *2017 10th International Conference on Human System Interactions (HSI)*, pages 41–47. IEEE, 2017.

[8] Alicja Kwasniewska, Jacek Ruminski, and Maciej Szankin. Improving accuracy of contactless respiratory rate estimation by enhancing thermal sequences with deep neural networks. *Applied Sciences*, 9(20):4405, 2019.

[9] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Krzysztof Czuszynski. Pose-invariant face detection by replacing deep neurons with capsules for thermal imagery in telemedicine. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 778–781. IEEE, 2018.

[10] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Krzysztof Czuszynski. Remote estimation of video-based vital signs in emotion invocation studies. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4872–4876. IEEE, 2018.

[11] Alicja Kwasniewska, Jacek Ruminski, Maciej Szankin, and Mariusz Kaczmarek. Super-resolved thermal imagery for high-accuracy facial areas detection and analysis. *Engineering Applications of Artificial Intelligence*, 87:103263, 2020.

[12] Alicja Kwaśniewska, Jacek Rumiński, and Jerzy Wtorek. The motion influence on respiration rate estimation from low-resolution thermal sequences during attention focusing tasks. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1421–1424. IEEE, 2017.

[13] Alicja Kwasniewska, Maciej Szankin, Jacek Ruminski, and Mariusz Kaczmarek. Evaluating accuracy of respiratory rate estimation from super resolved thermal imagery. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2744–2747. IEEE, 2019.

[14] Jacek Ruminski and Alicja Kwasniewska. Evaluation of respiration rate using thermal imaging in mobile conditions. In *Application of Infrared to Biomedical Sciences*, pages 311–346. Springer, Singapore, 2017.

[15] Jacek Rumiński, Alicja Kwaśniewska, Maciej Szankin, Tomasz Kocejko, and Magdalena Mazur-Milecka. Evaluation of facial pulse signals using deep neural net models. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3399–3403. IEEE, 2019.

[16] Maciej Szankin, Alicja Kwasniewska, and Jacek Ruminski. Influence of thermal imagery resolution on accuracy of deep learning based face recognition. In *2019 12th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2019.

[17] Maciej Szankin, Alicja Kwasniewska, Tejaswini Sirlapu, Mingshan Wang, Jacek Ruminski, Rey Nicolas, and Marko Bartscherer. Long distance vital signs monitoring with person identification for smart home solutions. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1558–1561. IEEE, 2018.

A.2 Publications in Other Fields

[1] Krzysztof Czuszynski, Alicja Kwasniewska, Maciej Szankin, and Jacek Ruminski. Optical sensor based gestures inference using recurrent neural network in mobile conditions. In *2018 11th International Conference on Human System Interaction (HSI)*, pages 101–106. IEEE, 2018.

[2] Krzysztof Czuszyński, Jacek Rumiński, and Alicja Kwaśniewska. Gesture recognition with the linear optical sensor and recurrent neural networks. *IEEE Sensors Journal*, 18(13):5429–5438, 2018.

[3] Alicja Kwasniewska, Maciej Szankin, Mateusz Ozga, Jason Wolfe, Arun Das, Adam Zajac, Jacek Ruminski, and Paul Rad. Deep learning optimization for edge devices: Analysis of training

quantization parameters. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, volume 1, pages 96–101. IEEE, 2019.

[4] Maciej Szankin, Alicja Kwaśniewska, Jacek Ruminski, and Rey Nicolas. Road condition evaluation using fusion of multiple deep models on always-on vision processor. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pages 3273–3279. IEEE, 2018.

[5] Libin Tang, Harish Subramony, Weian Chen, Jimin Ha, Hassnaa Moustafa, Tejaswini Sirlapu, Gauri Deshpande, and Alicja Kwasniewska. Edge assisted efficient data annotation for realtime video big data. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, pages 6197–6201. IEEE, 2018.

[6] Mingshan Wang, Tejaswini Sirlapu, Alicja Kwasniewska, Maciej Szankin, Marko Bartscherer, and Rey Nicolas. Speaker recognition using convolutional neural network with minimal training data for smart home solutions. In *2018 11th International Conference on Human System Interaction (HSI)*, pages 139–145. IEEE, 2018.

A.3 Projects

1. Power 3.5 - European Funding for Research and Development of DL-based project with the focus on remote healthcare using thermal image processing techniques

2. ERA-NET, CHIST-ERA program Project eGlasses - Interactive eye glasses for mobile, perceptual computing. Funded by European Coordinated Research program. Contractor: Design and development of mobile application and algorithms for contactless estimation of heart and breathing rates.

A.4 Artificial Intelligence Initiatives

1. Speaker and panelist: IEEE HSI 2015 (Poland), 2016 (UK), 2017 (S. Korea), 2018 (Poland); IEEE IECON 2018 (USA), 2019 (Portugal); QIRT 2016 (Poland); IEEE EMBC 2017 (S. Korea), 2018 (USA); Women In Tech Summit 2018 (Poland), Grace Hooper 2018 (USA); International Summer School on Deep Learning 2018 (Poland), 2019 (Poland);

2. Co-organizer: International Summer School on Deep Learning 2018 (Poland), 2019 (Poland), 2020 (Poland; Special Session "Connected-and-Automated Vehicle Integration, Safety, and Environment Design" IEEE IECON 2018 (USA); Special Session "Deep Neural Networks in Multimedia Data Analysis for HumanSystem Interaction" IEEE HSI 2018 (Poland); IEEE HSI Conference 2015

3. AI Committee Member: Grace Hooper 2019 (USA), 2020 (USA)

4. Reviewer: MDPI Journal; Transactions on Industrial Electronics Journal; Manning Publications; IEEE IECON; IEEE HSI

5. Cooperation with companies and universities: Non-disclosure Agreement between Gdansk University of Technology and Intel Corporation, USA in the area of artificial intelligence for smart home and autonomous driving; cooperation with University of Texas San Antonio - one year scholarship

A.5 Awards

1. Best paper award for: Maciej Szankin, Alicja Kwasniewska, and Jacek Ruminski. "Influence of thermal imagery resolution on accuracy of deep learning based face recognition." In2019 12th International Conference on Human System Interaction (HSI), pages 1–6. IEEE, 2019

2. Best paper award for: M. Wang, T. Sirlapu, A. Kwasniewska, M. Szankin, M. Bartscherer and R. Nicolas, "Speaker Recognition Using Convolutional Neural Network with Minimal Training Data for Smart Home Solutions." 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, 2018, pp. 139-145. doi: 10.1109/HSI.2018.8431363

3. Best paper award for: K. Czuszynski, A. Kwasniewska, M. Szankin and J. Ruminski, "Optical Sensor Based Gestures Inference Using Recurrent Neural Network in Mobile Conditions." 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, 2018, pp. 101-106. doi: 10.1109/HSI.2018.8430823

4. IEEE IES Young Professionals and Students Best Paper Recognition for: M. Wang, T. Sirlapu, A. Kwasniewska, M. Szankin, M. Bartscherer and R. Nicolas, "Speaker Recognition Using Convolutional Neural Network with Minimal Training Data for Smart Home Solutions," 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, 2018, pp. 139-145. doi: 10.1109/HSI.2018.8431363

A.6 Citations

The academic achievements of the author of the presented doctoral dissertation include (as of May 2020) 4 published articles from the JCR list, 1 paper published as a chapter in monography and 18 conference papers (15 indexed by Web of Science, Core Collection). Author's publications have been cited 117 times according to Google Scholar and 49 times according to Web of Science (Core Collection). Author's Hirsh index equals 7 according to Google Scholar and 5 according to Web of Science (Core Collection).

Appendix B

Super Resolution architecture

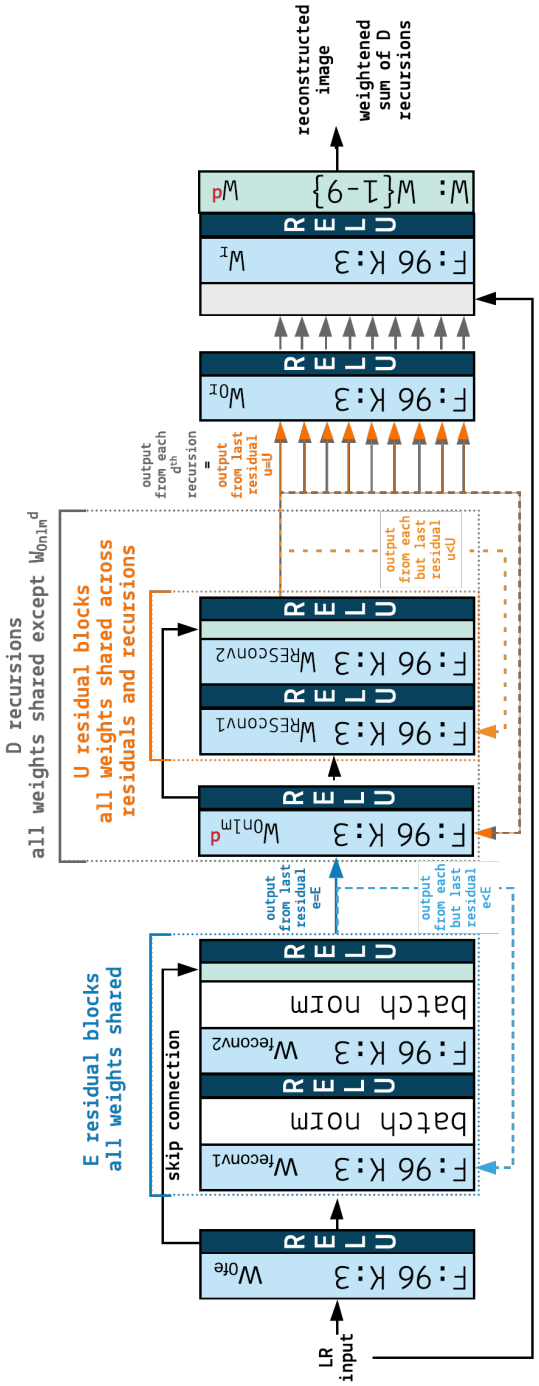


Figure B.1. Proposed CNN for thermal image enhancement