VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Paulius TUMAS

IMPROVEMENT OF INTELLIGENT METHODS FOR PEDESTRIAN DETECTION IN FAR-INFRARED RADIATION IMAGES

DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
ELECTRICAL AND ELECTRONIC ENGINEERING (T 001)

Vilnius, 2021

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2016–2021.

Scientific supervisor

Prof. Dr Artūras SERACKIS (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

The Dissertation Defense Council of Scientific Field of Electrical and Electronic Engineering of Vilnius Gediminas Technical University:

Chairman

Assoc. Prof. Dr Vitalij NOVICKIJ (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

Members:

Dr Habil. Pawel FORCZMANSKI (West Pomeranian University of Technology in Szczecin, Poland, Electrical and Electronic Engineering – T 001),

Prof. Dr Rytis MASKELIŪNAS (Kaunas University of Technology, Informatics Engineering – T 007),

Assoc. Prof. Dr Raimondas POMARNACKI (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001),

Prof. Dr Nerija ŽURAUSKIENĖ (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

The dissertation will be defended at the public meeting of the Dissertation Defense Council of Electrical and Electronic Engineering in the Senate Hall of Vilnius Gediminas Technical University at **3 p. m. on 8 July 2021**.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.
Tel. +370 5 274 4956; fax +370 5 270 0112; e-mail: doktor@vilniustech.lt

A notification on the intend defending of the dissertation was send on 7 June 2021. A copy of the doctoral dissertation is available for review at Vilnius Gediminas Technical University repository http://dspace.vgtu.lt, at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Wroblewski Library of the Lithuanian Academy of Sciences (Žygimantų st. 1, LT-01102, Vilnius, Lithuania).

Vilnius Gediminas Technical University scientific book No. 2021-030-M
doi: 10.20334/2021-030-M

paulius.tumas@vilniustech.lt

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Paulius TUMAS

INTELEKTUALIŲJŲ METODŲ PĖSTIESIEMS APTIKTI TOLIMOSIOS INFRARAUDONOSIOS SPINDULIUOTĖS VAIZDUOSE TOBULINIMAS

DAKTARO DISERTACIJA

TECHNOLOGIJOS MOKSLAI,
ELEKTROS IR ELEKTRONIKOS INŽINERIJA (T 001)

Vilnius, 2021

Disertacija rengta 2016–2021 metais Vilniaus Gedimino technikos universitete.

Vadovas

prof. dr. Artūras SERACKIS Vilniaus Gedimino technikos universitetas, technologijos mokslai, elektros ir elektronikos inžinerija – T 001.

Vilniaus Gedimino technikos universiteto Elektros ir elektronikos inžinerijos mokslo krypties disertacijos gynimo taryba:

Pirmininkas

doc. dr. Vitalij NOVICKIJ (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001).

Nariai:

habil. dr. Pawel FORCZMANSKI (Ščecino Vakarų Pamario technologijos universitetas, Lenkija, elektros ir elektronikos inžinerija – T 001),

prof. dr. Rytis MASKELIŪNAS (Kauno technologijos universitetas, informatikos inžinerija – T 007),

doc. dr. Raimondas POMARNACKI (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001),

prof. dr. Nerija ŽURAUSKIENĖ (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001).

Disertacija bus ginama viešame Elektros ir elektronikos inžinerijos mokslo krypties disertacijos gynimo tarybos posėdyje **2021 m. liepos 8 d. 15 val.** Vilniaus Gedimino technikos universiteto senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.
Tel. +370 5 274 4956; fax +370 5 270 0112; el. paštas: doktor@vilniustech.lt

Pranešimai apie numatomą ginti disertaciją išsiųsti 2021 m. birželio 7 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto talpykloje http://dspace.vgtu.lt, Vilniaus Gedimino technikos universiteto bibliotekoje (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) bei Lietuvos mokslų akademijos Vrublevskių bibliotekoje (Žygimantų g. 1, LT-01102 Vilnius, Lietuva).

Abstract

Each year, over 1.35 million lives are tragically lost on roads, according to The World Health Organization (WHO). Even though the European Union (EU) has the safest roads in the world, 221 people are being killed on roads every day, thousands more are injured or disabled, with long-lasting effects. Each year EU introduces new safety measures in cars, lorries, and buses for advanced driver assistance systems (ADAS) to prevent accidents. One of the primary functions of ADAS systems is pedestrian detection based on intelligent systems. The recent development of convolutional neural network (CNN) based detectors has proven excellent results in object detection. However, not many studies have been performed with a low resolution far-infrared spectrum images. Since CNN based object detection training requires many images, a new FIR domain dataset is introduced captured during severe weather conditions called ZUT-FIR-ADAS (ZUT). This dataset is the second biggest open-access FIR dataset containing Controller Area Network (CAN) bus data synchronized with the FIR images. Then state of the art YOLO (You Only Look Once) detector is modified and trained on this newly introduced dataset, reaching 89.1 mAP (mean Average Precision). However, the dataset and detectors comparison revealed that DNN detectors tend to adapt to specific conditions and features from captured images and do not work accurately when different dataset images are provided. For this reason, ZUT and SCUT (the biggest open access FIR domain dataset) datasets were merged, and two parallel experiments were done. The first experiment aimed to find a training approach and optimize detector structure for speed and performance. The experiment showed that it is possible to increase accuracy by more than five mAP units by retraining the detector on images where the detector fails to detect pedestrians the most. The first experiment also revealed a possibility to minimize detector structure and decrease needed floating point operations by four times without losing accuracy. The second experiment aimed to transfer severe weather features from the ZUT dataset to the SCUT dataset. The experiment revealed that newly generated images increased the accuracy of the detector by 9.38 mAP. The thesis results were published in seven scientific publications – three in peer-reviewed scientific papers, four in conference proceedings. Additionally, the results of the research were presented in seven conferences.

Reziumė

Pasak Pasaulio sveikatos organizacijos (PSO), per metus, pasaulio keliuose tragiškai praranda gyvybę daugiau nei 1,35 milijono žmonių. Nors Europos Sąjungoje (ES) yra saugiausi keliai pasaulyje, kasdien keliuose žūva 221 žmogus, tūkstančiai suluošinami arba tampa neįgalūs, sukeliamos ilgalaikės pasėkmes. Kasmet diegiamos naujos lengvųjų automobilių, sunkvežimių ir autobusų saugos technologijos, skirtos pažangioms vairuotojo pagalbos sistemoms, siekiant išvengti nelaimingų atsitikimų. Viena pagrindinių vairuotojo pagalbos sistemų funkcijų yra pėsčiųjų aptikimas, grįstas intelektualiaisiais metodais. Neseniai sukurti aptiktuvai, grįsti sąsūkos dirbtinių neuronų tinklais (SDNT), parodė puikius objektų aptikimo rezultatus. Tačiau atlikta nedaug tyrimų su mažos raiškos tolimojo infraraudonųjų spindulių (TIS) spektro vaizdais. Kadangi SDNT pagrįstam objektų aptikimo mokymui reikalinga daugybė vaizdų, paruoštas naujas TIS domenų rinkinys, surinktas blogomis oro sąlygomis ir pavadintas ZUT-FIR-ADAS (ZUT). Šis rinkinys yra antras pagal dydį atviros prieigos TIS duomenų rinkinys, kuriame papildomai pateikti automobilio CAN magistralės duomenys, sinchronizuoti su TIS kameros vaizdais. Atrinkti moderniausių struktūrų aptiktuvai buvo modifikuojami ir mokomi naudojant šį naujai pristatytą duomenų rinkinį. Pavyko pasiekti 89,1 % vidutinį atpažinimo tikslumą, didesnį nei naudojant altenatyvius, nemodifikuotų struktūrų aptiktuvus. Analizuojant sąsajas tarp pavyzdžių sąvybių ir aptiktuvo atpažinimo rezultatų paaiškėjo, kad SDNT aptiktuvai linkę prisitaikyti prie vaizdų, surinktų vienodomis oro sąlygomis ir nebeveikia taip tiksliai, kai testuojama ant skirtingomis oro sąlygomis duomenų rinkinių. Dėl šios priežasties ZUT ir SCUT duomenų rinkiniai buvo sujungti į vieną ir atlikti du eksperimentiniai tyrimai. Pirmojo tyrimo tikslas buvo rasti mokymo metodą ir optimizuoti detektoriaus struktūrą taip, kad atpažintuvas veiktų greitai ir ne mažesniu tikslumu, nei moderniausios struktūros aptiktuvas. Tikslumą pavyko padidinti daugiau nei penkiais procentais, aptiktuvo mokymui naudojant vaizdus, kur aptiktuvas labiausiai klysta aptinkant pėstįjį po pirmojo mokymo etapo. Taip pat buvo ištirta galimybė sumažinti SDNT grįsto aptiktuvo struktūrą ir padidinti apdorojimo greitį iki keturių kartų, neprarandant pėsčiųjų aptikimo tikslumo. Antruoju eksperimentu buvo siekiama papildyti geromis oro sąlygomis surinktų duomenų rinkinį SCUT naujais vaizdais, imituojant blogų oro salygų įtaką vaizdui. Vaizdo iškraipymus pavyko automatizuotu būdu perkelti SCUT duomenų rinkinio vaizdams mokant specialios struktūros SDNT and ZUT duomenų rinkinio pavyzdžių. Eksperimentinis tyrimas atskleidė, kad mokymui naudoti naujai sugeneruoti vaizdai, padidino pėsčiųjų aptikimo tikslumą 9,38 %. Darbo rezultatai paskelbti septyniose mokslo publikacijose: trijuose recenzuojamuose mokslo žurnaluose, keturiuose kituose leidiniuose ir pristatyti septyniose mokslo konferencijose.

Notations

Abbreviations

ABS – Anti-lock Braking System;
ADAS – Advanced Driver-Assistance Systems;
AGX – Jetson AGX Xavier Single Board Computer;
AP – Average Precision;
CAN – Controller Area Network;
CNN – Convolutional Neural Network;
CPU – Central Processing Unit;
DNN – Deep Neural Network;
EU – European Union;
FIR – Far-Infrared;
FN – False Negative;
FP – False Positive;
FPR – False Positive Rate;
FPS – Frames Per Second;
GDPR – General Data Protection Regulation;
GPU – Graphics Processing Unit;
HOG – Histograms of Oriented Gradient;
ICM – Instrument Cluster Module;
IoU – Intersection over Union;

IR – Infrared;
KAIST – Multispectral Pedestrian Dataset by KAIST;
LBP – Local Binary Pattern;
LWIR – Long-Wave Infrared;
mAP – mean Average Precision;
MWIR – Mid-Wave Infrared;
NIR – Near-Infrared;
RTX – NVIDIA RTX2070Ti Graphics Card;
SCUT – South China University of Technology Pedestrian Dataset;
SD – Standard Deviation;
SDK – Software Development Kit;
SVM – Support Vector Machine;
SWIR – Short-Wave Infrared;
TI – Threshold Intensity;
TP – True Positive;
TX2 – Jetson TX2 Single Board Computer;
UDP – User Datagram Protocol;
WHO – World Health Organization;
YOLO – You Only Look Once;
ZUT – ZUT-FIR-ADAS Pedestrian Dataset.

Contents

INTRODUCTION 1
Problem Formulation 1
Relevance of the Thesis 1
The Object of the Research 2
The Aim of the Thesis 2
The Objectives of the Thesis 3
Research Methodology 3
Scientific Novelty of the Thesis 3
Practical Value of the Research Findings 4
The Defended Statements 4
Approval of the Research Findings 4
Structure of the Dissertation 5
Acknowledgements 5

1. REVIEW OF THERMAL VISION-BASED PEDESTRIAN DETECTION TECHNIQUES 7
1.1. Review of Pedestrian Detection Possibilities at Different Spectrum Bands 8
1.1.1. Pedestrian Detectors based on Near-Infrared Cameras 9
1.1.2. Pedestrian Detectors based on Short-Wave Infrared Cameras . . . 9
1.1.3. Pedestrian Detectors based on Mid-Wave Infrared Cameras 10

1.1.4. Pedestrian Detectors based on Long-Wave Infrared Cameras ... 11
1.1.5. Pedestrian Detectors based on Far-Infrared Cameras 11
1.2. Hardware Solutions for Pedestrian Detection in Modern Vehicles 12
1.3. Vision-based Detector Performance Evaluation Criteria 13
1.4. Vision-based Pedestrian Detection Techniques 15
1.4.1. Pedestrian Detectors based on Histograms of Oriented Gradient Features 15
1.4.2. Object Classification using Support Vector Machines 16
1.4.3. Object Classification using Fuzzy logic 17
1.4.4. Feature Extraction and Classification using Convolutional Neural Network 18
1.4.5. Deep Neural Network based Feature Extraction and Object Classification 19
1.4.6. Modern Deep Neural Network based Object Detectors 21
1.4.7. YOLO Architecture 23
1.5. Conclusions of the First Chapter and Formulation of Dissertation Tasks 24

2. INVESTIGATION OF THE PEDESTRIAN DETECTOR PROTOTYPE .. 25
2.1. Performance Comparison of Pedestrian Detectors 26
2.1.1. Histograms of Oriented Gradient Feature based Pedestrian Detection 26
2.1.2. Histograms of Oriented Gradient Classifier Preparation 28
2.1.3. Histograms of Oriented Gradient Classifier Performance Estimation Results 29
2.2. Analysis of Currently Available Dedicated Datasets 30
2.2.1. SCUT Dataset 31
2.2.2. KAIST Dataset 32
2.2.3. The Drawbacks of Currently Available Datasets 32
2.3. Prototype Development for Dataset Collection 34
2.3.1. Preparation of the Computing Unit 34
2.3.2. Development of Controller Area Network Bus Data Capture Solution 35
2.3.3. Developing of Image Acquisition and Pre-Labeling Solution .. 36
2.4. Introduction of the New Dataset for Pedestrian Detection 39
2.4.1. Data Preparation for Deep Neural Network based Pedestrian Detection 39
2.4.2. Analysis and Correction of Pre-Labeled Annotations 40
2.5. RAW Image Preprocessing and Darknet Modifications 43
2.5.1. Experimental Investigation of Two Candidate Architectures for the Detector 44
2.5.2. Validation of Experimental Investigation Results 45

2.6. Conclusions of the Second Chapter 50

3. IMPROVEMENT AND EXPERIMENTAL TESTS OF THE PEDESTRIAN DETECTOR 51
3.1. Fusion of the Datasets and Analysis of the Annotations' Distribution . 52
3.2. Selection of the Improved Detectors for the Prototype 53
3.2.1. Results of Detector Training on a Fused Dataset 55
3.2.2. Retraining of the Detector According to Confidence Distribution 58
3.2.3. Testing Pedestrian Detectors on the Single-Board Computers . . 61
3.2.4. Optimization of the Pedestrian Detector 63
3.3. Development of the Dataset Augmentation Algorithm 66
3.3.1. Dataset Augmentation Preparation 67
3.3.2. Selection and Modification of Deep Learning Structure for Data Augmentation 69
3.3.3. Review of Dataset Augmentation Results 69
3.4. Conclusions of the Third Chapter 72

GENERAL CONCLUSIONS 73

REFERENCES 75

LIST OF SCIENTIFIC PUBLICATIONS BY THE AUTHOR ON THE TOPIC OF THE DISSERTATION 87

SUMMARY IN LITHUANIAN 89

ANNEXES[1] 101
Annex A. Declaration of Academic Integrity 102
Annex B. The Co-authors' Agreements to Present Publications Material in the Dissertation 103
Annex C. The Copies of Scientific Publications by the Author on the Topic of the Dissertation 109

[1]The annexes are supplied in the enclosed compact disc.

Introduction

Problem Formulation

This thesis is focusing on one of the main Advanced driver-assistance systems duty – fast and accurate pedestrian detection. A pedestrian detection is well know and actively researched topic nowadays, since automotive industry is trying to build a self-autonomous vehicles on the roads. However, a prediction of possible collision with a pedestrian in time requires a very accurate and fast detection algorithms in computer vision. One of the challenges to build such algorithms is an undefined number of various traffic, pose, illumination, season and weather conditions happening in transportation system every day. Hardware is another limiting factor, since thermal cameras have a very low resolution compared with visual spectrum cameras, so that the pedestrians might be as tall as five pixels in the image. The computing power is also preventing reaching real-time performance without trading accuracy.

Relevance of the Thesis

WHO serves as the secretariat for the Decade of Action for Road Safety 2011–2020. According to the Global status report on road safety (World Health Organization 2018), in 2018, traffic deaths reached 1.35 million, where half of the traffic accidents belong to the category of road users, cyclists, and pedestrians. Road

traffic injuries are now the leading killer of people aged 5–29 years. One of the main causes of traffic accidents are speeding and distracted driving. According to the report, an increase in average speed is directly related to occurring and severity of the crash. For instance, every 1% increase in mean speed produces a 4% increase in the fatal crash risk and severity to 3%. Distracted driving like mobile phone usage result four times bigger probability to involve in a crash. It also slows reaction times to braking and traffic signals, makes it difficult to keep in the correct lane and to keep the proper following distances. To prevent accidents, a new safety measures for ADAS are introduced like intelligent speed assistance, advanced emergency braking, driver drowsiness and attention monitoring.

In 2019, an European Union regional status report on road safety was prepared (World Health Organisation 2019). This report shows that over 221 people are killed on roads every day in the European region, and thousands more are injured or disabled, with long-lasting effects. According to the research, 30% of killed road users are pedestrians and cyclists. The main reasons for fatalities are rapid urbanization and motorization, poor safety standards and infrastructure, lack of strong enforcement, drivers being distracted or under the influence of drugs or alcohol, a failure to wear seat belts or helmets, and lack of access to timely post-crash care. Speeding is another critical element causing lack of time to avoid the accident, and early-stage detection of collision could drastically minimize the chance of accident (Breen *et al.* 2020; Khan, Khan 2018; Kumar, Kushwaha 2016). Lastly, severe weather conditions like rain, snow, fog are visibility affecting factors causing drivers to adapt to the conditions. However, the study of Das *et al.* 2018 showed that fog or smoke is 3.24 times more likely to result in a severe injury and is 1.53 times more likely to cause a multiple-vehicle crash. A similar study, prepared by Sun *et al.* 2011 were analyzed rain influence for the diver and, depending on road type, the risk to have an accident increase to 2.61 times.

The Object of the Research

The main research object of the doctoral dissertation are images containing pedestrians captured by a far-infrared spectrum camera. An existing DNN structures also analyzed to optimize the structure to gain detector speed and keep accuracy. Also, data augmentation techniques are used to enrich the dataset with new features.

The Aim of the Thesis

This thesis aims to improve novelty methods for pedestrian detection in far-infrared radiation images by focusing on detector processing speed and accuracy.

The Objectives of the Thesis

In order to reach the main aim of the thesis, there were these tasks defined below:

1. Improve the processing speed of pedestrian detector based on Histogram-Oriented Gradient features, working on Central Processing Unit.
2. Develop an adaptive pedestrian detector prototype based on ambient temperature impact on Far Infrared Radiation images.
3. Optimise convolutional neural network based pedestrian detector for Edge-Computing Prototype.
4. Investigate dataset augmentation techniques to improve pedestrian detection in severe weather.

Research Methodology

The work applies digital image processing, artificial neural networks, deep learning, statistical analysis theories. Adapted and implement image preprocessing, DNN training and execution, speed and accuracy assessment. Specialized image datasets were collected and compiled for experiments to train DNN to detect a pedestrian in the FIR spectrum. Implementation was provided by software packages like Cuda, Darknet, OpenCV, Pytorch, Matlab, WolframAlpha and a training cluster made of AMD 3900X and Intel i7 8^{th} generation processors. Also, graphical accelerators were used, utilizing training stations using double NVIDIA RTX2080Ti graphics cards or training stations with double NVIDIA Geforce GTX 1080Ti graphics cards. The single-board computers like NVIDIA TX2 and Xavier AGX were utilized to perform real-time pedestrian detection.

Scientific Novelty of the Thesis

A collected dataset contains 122,000 annotations, and more than 79,000 were collected during the drizzle or the rain. The remaining annotations were collected during frosty and cloudy conditions. This dataset fills the gap in existing pedestrian detection algorithm research since, until the publication date, there was no such dataset published. Also, a dataset provides a synchronised car CAN bus data (non to existing provided such data), which can be used to create ADAS systems with conjunction of thermal image-based detectors. Also, experiments show that an extended dynamic range of FIR images from 8 bit to 16 bit is a valuable way to improve the accuracy of the DNN based pedestrian detector. The training and slimming approach show how to remove unnecessary DNN elements to gain pro-

cessing speed without losing accuracy. Augmentation by using inverse denoising neural networks also helps artificially create new situations in a controlled manner.

Practical Value of the Research Findings

A provided methodology, experiment investigation, DNN architecture, and the collected dataset are openly accessible, providing an opportunity to enhance and continue the research and development for newly developed ADAS systems. It also allows performing sensor-fusion pedestrian detection research since it contains a synchronized car data with the specific frame. The provided detector training and structure optimization techniques help to improve detector detection speeds several times without losing accuracy. The augmentation technique helps to scale small size and feature datasets by expanding detector capabilities.

The Defended Statements

1. The extended dynamic range of FIR images from 8 bit to 16 bit improves the mean average precision of the TinyV3 pedestrian detector by 11.2%.
2. Selection of training examples according to the confidence distribution increases the mean average precision of the ResNext50 pedestrian detector by 6.24%.
3. The slimming of the convolutional neural network structure decreases calculations 4.83 times, reaching a gain of 11.9 FPS running on the AGX single-board computer and obtaining an 8.38% mean average precision increase.
4. Augmentation of the dataset by simulating severe weather conditions impact to images using a deep learning network increases the ResNext50 pedestrian detector mean average precision by 9.38%.

Approval of the Research Findings

The main research contributions were published in seven scientific papers. Three of them are published in scientific journals:

- Electronics (MDPI), referenced in Clarivate Analytics Web of Science (CA WoS), with impact factor 2.110 (P. Tumas, A. Serackis, A. Nowosielski 2021).
- IEEE Access, referenced in CA WoS, with impact factor 4.098 (P. Tumas,

A. Nowosielski, A. Serackis 2020).

- International Journal of Advanced Research, referenced in Index Copernicus, (A. Jonkus, P. Tumas, A. Serackis 2018).

Four of them are published in conference proceedings, referenced in CA WoS Proceedings.

Results of dissertation were presented in seven conferences:

- RB Feature based Matching of Two-Dimensional Electrophoresis Gel Images. Biomedical Engineering‘2016, Kaunas, Lithuania.
- Mėsos / žuvies spektrinių savybių pasyvaus stebėjimo sistema. JMK‘2017, Vilnius, Lithuania.
- ROI Detection for Food Quality Inspection Systems. eSTREAM‘2017, Vilnius, Lithuania.
- Pėsčiųjų aptikimas infraraudonųjų spindulių vaizdo sraute. JMK‘2018, Vilnius, Lithuania.
- Pedestrian detection using FIR domain camera. eSTREAM‘2018, Vilnius, Lithuania.
- Automated image annotation based on YOLOv3. AIEEE‘2018, Vilnius, Lithuania.
- Lightweigth pedestrian tracker. JMK‘2019, Vilnius, Lithuania.

Structure of the Dissertation

The dissertation consists of an introduction, three chapters and general conclusions. The volume of the dissertation is 102 pages, in which are given: 43 figures, 4 equations and 37 tables. Additionally, 126 items are cited in the dissertation.

Acknowledgements

I would like to express the greatest gratitude to Prof. Dr Artūras Serackis for supervising me in this journey, for his critical and professional thinking, and dedication to science. Also, special thanks to Prof. Dr Dalius Navakauskas for supporting me in terminology and shared knowledge. West Pomeranian University of Technology for providing a costly camera and especially Dr Adam Nowosielski for hospitality and very friendly internship. A Yukon Advanced Optics for introducing me to the pedestrian detection topic, lending a thermal camera and computational hardware. But more importantly, I would like to thank my wife, Sandra Tumienė, for her contribution to the annotating process, patience, and support during challenging

moments. Of course, to my study mate and good friend Julius Skirelis for exciting discussions, knowledge sharing, and research process support. Lastly, to my family members of support during the studies.

1

Review of Thermal Vision-based Pedestrian Detection Techniques

Pedestrian detection has always been a vital problem in the domain of Computer Vision and Pattern Recognition – especially in the infrared spectrum. It is formulated as the problem of because of two main reasons: (1) the low resolution of thermal camera providing less texture information, and (2) the lack of large-scale pedestrian dataset in infrared spectrum to ensure the training of deep learning-based detectors with good generalization performance.

In addition, pedestrian detection is also challenging task because of there are so many variations in images like:

- different body attire and pose;
- occlusion;
- other illumination parameters in different scenarios;
- the presence of clutter in the background.

One of first aim in this section is to review different infrared spectrum bandwidths to determine which spectrum parts provide the best details for pedestrian detection application in thermal domain. Next, it is analyzed where is the current lack of research for pedestrian detection application. Later in this section, existing industry solutions are reviewed, and state-of-art objected detection methods are introduced to known. In the end, conclusions are drawn and formulation of dissertations tasks are created. The review, presented in this chapter is published in three scientific papers (Tumas, Serackis 2018, Tumas *et al.* 2020 and Tumas *et al.* 2021).

1.1. Review of Pedestrian Detection Possibilities at Different Spectrum Bands

Human eyesight is very narrow, where according to the Fig. 1.1, the visible spectrum is a small part of the surrounding range. This is because our eye is limited by absence of a tapetum lucidum, which is a layer of tissue that reflects distinct light back through the retina. Since the human eye is quite robust to surroundings, providing high-level contextual information to our brains during the day-time, our vision is quite limited during the night-time. An exploration to expanding human vision during the night-time started in World War II, where first night-vision systems were used. Such night vision systems depends on the low light levels of starlight and night sky brightening (called "atmospheric nightglow") to help picture the focused on the scene and its environment.

A modern touch to utilize infrared spectrum is spread through various parts of our lives. One of the mainstream usages is transportation. Since infrared cameras are sensitive to thermal energy, automakers develop ADAS systems to help orienteer the driver and avoid collisions by detecting animals and pedestrians on the road. According to the Fig. 1.1 infrared spectrum is classified into three parts:

- Near-Infrared;
- Mid-Infrared;
- Far-Infrared.

Such classification provides different features and possibilities for pedestrian detection systems.

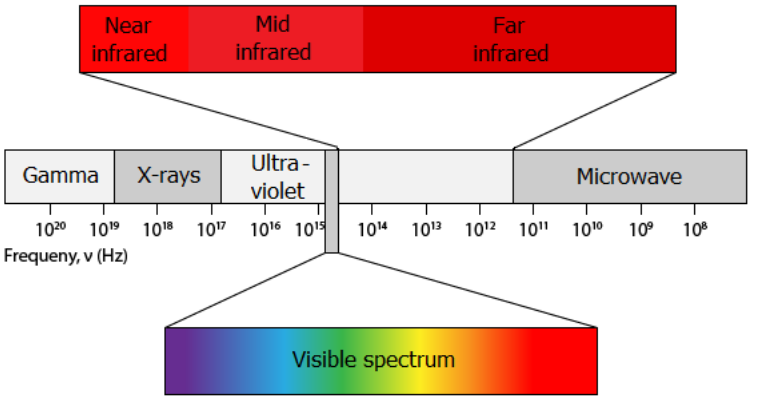


Fig. 1.1. Visualization of visible and infrared spectrum bands

1.1.1. Pedestrian Detectors based on Near-Infrared Cameras

The most highly available and cheapest cameras on the market are near-infrared (NIR) cameras. It is a visual spectrum camera without an infrared spectrum filter. Such cameras for pedestrian detection systems could provide 44–73 m of visibility during the night according to Tsimhoni *et al.* 2007. However, NIR vision's primary drawbacks include their susceptibility to glare, blooming, and streaking from active light sources such as oncoming traffic, traffic lights, streetlights, and reflective objects such as road signs. Also, NIR illuminators may cause glare to other drivers using the same type of system and may cause damage to eyes at short distances (<1 m) based on Yagi *et al.* 2003 research.

Auto manufacturers like Toyota also try to use NIR cameras to enhance visibility for the driver by emitting near-infrared light through headlight projectors. Then a camera captures that reflected radiation. However, according to Toyota Motor Asia Pacific Pte Ltd 2005 the reported visibility distance is only up to 250 m. In history, this is not the first attempt to use additional radiations source. Eichhorn *et al.* 2001 reported the development of laser-diode-based IR illuminators. Laser-diode technology may effectively increase the illuminators' power and reduce the susceptibility of the camera to glare. Time synchronization of the camera with the laser pulses opens additional options for dealing with current NIR systems' drawbacks.

1.1.2. Pedestrian Detectors based on Short-Wave Infrared Cameras

A short-wave infrared (SWIR) spectrum (part of mid-infrared range) image sensors are rarely used in pedestrian detection applications. Bertozzi *et al.* 2013 tried to use SWIR cameras in poor visibility conditions. Research has demonstrated that

a) b)

Fig. 1.2. Images acquired by Miron *et al.* (2013) outdoor: a) visible; b) SWIR bandwidths highlight similar features both for pedestrian and background

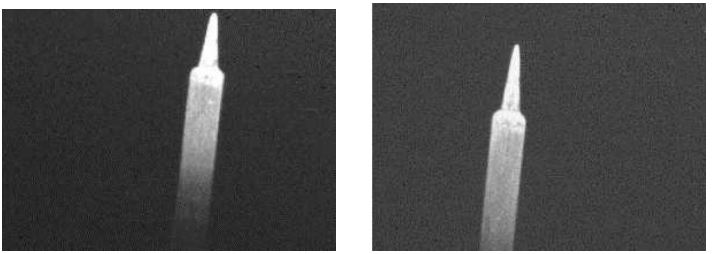
reduced visibility phenomena as haze and fog feature quite different behaviors in the SWIR spectrum but still of no practical utility for automotive applications. Whereas better visibility through haze may be achieved by employing SWIR sensors, but it is a negligible benefit for pedestrian detection, hazing a long-distance phenomenon, no improvements can be obtained in foggy conditions. Similar results observed by Miron *et al.* 2013, where authors shown that detection rates obtained are no better (see Fig. 1.2) than reported in the revised state-of-the-art works.

1.1.3. Pedestrian Detectors based on Mid-Wave Infrared Cameras

Mid-Wave infrared (MWIR) is popular bandwidth used in military applications such as guided missile technology (see Fig. 1.3). The 3000–5000 nm portion of MWIR band is the atmospheric window in which the homing heads of passive IR 'heat-seeking' missiles are designed to work, homing on to the IR signature of the target aircraft, typically the jet engine exhaust plume according to Negied *et al.* 2015.

There are not many published research works related to MWIR use for pedestrian detection use, probably because of specific military use. However, Nguyen *et al.* 2014 tried to set up a dual detector based on MWIR and Long-Wave Infrared (LWIR) spectrum. The work provided a very robust pedestrian tracking. However, none of the tested sequences provided ambient conditions such as fog, smoke, haze, and precipitation, all of which are expected to impact the two bands differently.

Another difficulty of using MWIR cameras for pedestrian detection is that the camera has large dimensions (see Fig. 1.4). Also, it is required to have cryogenic cooling, and because of that, it also happens to be a power-hungry application according to Barrière *et al.* 2012; Catanzaro *et al.* 2004; Druart *et al.* 2013.



a) b)

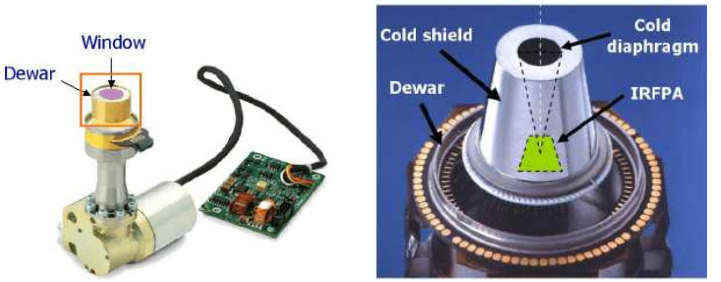
Fig. 1.3. Images from Krishna (2005) captured by the 320×256 DWELL focal plane array: a) with MWIR (3000–5000 nm) filter; b) LWIR (8000–12000 nm) filter

1.1.4. Pedestrian Detectors based on Long-Wave Infrared Cameras

LWIR cameras can provide a better fit for person detection, especially in complex outdoor scenarios with masking background texture or lack of illumination according to Teutsch *et al.* 2014. In general, the human appearance in LWIR images is not homogeneously bright due to clothes and other effects. Instead, there are smooth gray-value transitions inside the human blob and, in case of a merge also to surrounding bright background regions. Also, according to Van Beeck *et al.* 2017, LWIR images provide visible pedestrians in severe weather conditions (e.g., fog, heavy rain).

1.1.5. Pedestrian Detectors based on Far-Infrared Cameras

FIR is different from other infrared spectrum wavelength bands because it provides a detailed view of human body features, such as hotspots for candidate detection Kim, Kim 2018. Saito *et al.* 2008 also observed that during bad weather, FIR performs better than other cameras because far-infrared rays are less susceptible to moisture than other wavelength bands' rays compared and described in Table 1.1. Unlike NIR or visible-ray cameras, FIR cameras are not sensitive to disturbing light, such as oncoming headlights. The possible disadvantages of FIR according to Jeong *et al.* 2017 are that in the daytime and especially summer, there is a very little temperature difference between pedestrian and background objects (such as buildings and roads) because the sun heats background objects which makes less visible pedestrians. Also, FIR cameras are more expensive (lenses and imaging devices are especially expensive) than NIR cameras or visible-ray cameras. Similar statements were also absorbed by Gonzalez Alzate *et al.* 2016 where authors



a) b)

Fig. 1.4. Representation of Cryo cooler taken from Barrière *et al.* 2012 where: (a) external view; (b) internal view

tried to use a dual-camera setup (visual and FIR). According to them, "The detector worked the best by combining visible spectrum camera with FIR imagery and during the night capture FIR features provided the best results."

1.2. Hardware Solutions for Pedestrian Detection in Modern Vehicles

Advanced driver-assistance systems ADAS are control systems integrated into vehicles to assist the driver according to Shopovska *et al.* 2019. These systems do not take actions fully autonomously but provide relevant information to drivers and assist them in performing critical maneuvers. Combined with variety of vehicle heterogeneous sensor equipment sensors, ADAS can present the driver with additional data such as distance to objects and warnings for increased safety. When a human driver performs a complex driving operation, such as lane changing, it will focus on different areas to ensure lane safety based on Zhang *et al.* 2017 research. Current commercially available driver-assistance systems include Tesla Autopilot, Mobileye, and Openpilot that help the driver in collision avoidance, automatic lane centering, parking assistance etc.

Automakers like Audi, BMW, and Daimler offer Autoliv designed FLIR Path finder nighttime driving assistance which displays a feed from a thermal camera described in Fig. 1.5. Such a system is based on a FIR spectrum FLIR camera having resolution 324×256 with a refresh rate of 30 Hz found in (FLIR Systems Inc 2019). However, not many details are available except on publication by Forslund, Bjärkefur 2014, in which it is mentioned that the detector was based on a Cascade classifier with Haar-like feature, and the dataset was collected driving eight years, four seasons in various locations and about one million miles were driven. The official user manual (FLIR Systems Inc 2019) is also limited by details of the system. There are no exact accuracy measures and details of temperature ranges but the only found is a single statement: "Depending on conditions and ambient temperatures, the detection algorithms may work poorly or not at all during the daytime."

Some automakers like Toyota and Daimler try to enhance visibility for the driver by using NIR cameras and additionally mounted projectors by emitting near-

Table 1.1. Infrared spectrum comparison based on Saito *et al.* 2008

Performance	FIR	NIR	Visible spectrum
Day	Good	Good	Good
Night	Excellent	Good	Poor
Bad Weather	Good	Poor	Poor

infrared. A camera captures that reflected radiation and video is displayed for the driver. However, the visibility distance can be increased only up to 250 m based on Toyota Motor Asia Pacific Pte Ltd 2005 documentation.

A comparative study was performed by Nowosielski *et al.* 2020 to challenge the ADAS system effectiveness on human performance versus a thermal imaging–based automatic system in severe lighting conditions, the second turned out to be better at detecting pedestrians. In many cases, participants of conducted experiments did not notice pedestrians at all, in contrast to a computer system that analyzed the thermal data.

1.3. Vision-based Detector Performance Evaluation Criteria

In this section it is analysed the criteria how to evaluate detector performance.

One of the essential things in machine learning is the evaluation of the model. One of the most important detector indicator is precision which is defined as Equation 1.1. It is defined as the true positive (TP) ratio and the total number of predicted detections including false negative (FN).

$$P = \frac{TP}{TP + FN}, \tag{1.1}$$

where P is the precision.

Recall defined as Equation 1.2 effectively describes the completeness of TP relative to the ground truth annotations sum of TP and false positive (FP).

$$R = \frac{TP}{TP + FP}, \tag{1.2}$$

where R is the recall.

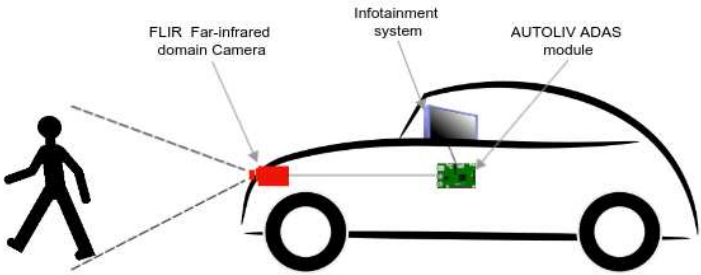


Fig. 1.5. FLIR Pathfinder nighttime driving assistance

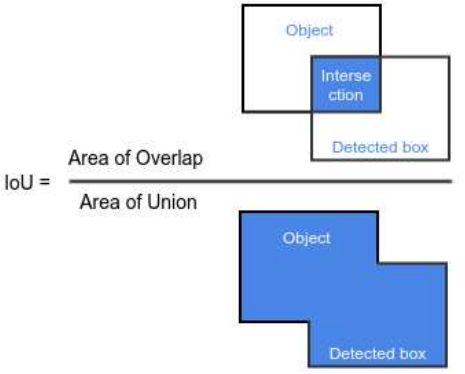


Fig. 1.6. Intersection over union

Intersection over Union (IoU) (see Fig. 1.6), also referred to as the Jaccard Index, is an evaluation metric that quantifies the similarity between the ground truth bounding box and the predicted bounding box to evaluate how good the predicted box is. In general, the IoU measures the overlap between the ground truth box and the predicted box over their union.

Two criteria are usually used to evaluate object detectors' performance: speed measured in frames per second (FPS) and accuracy evaluated by mean Average Precision (mAP). It is a metric typically used for PASCAL challenges Everingham *et al.* (2008) where Average Precision (AP) for one object class is calculated having an IoU threshold of 0.5 and the mAP is calculated by averaging AP over all object classes. Mathematically, AP can be expressed as: Equation 1.3.

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \times 100\%\,, \tag{1.3}$$

where P_n and R_n are the precision and recall at the n threshold which is typically 50% overlapping area of IoU.

There is additional metric called F1-score defined as Equation 1.4. It is the harmonic mean of Precision and Recall and this metric gives a better measure of the incorrectly classified cases.

$$F1 = \frac{2PR}{P+R}\,, \tag{1.4}$$

where P is precision and R is recall.

1.4. Vision-based Pedestrian Detection Techniques

This section analyzes pedestrian detection techniques by reviewing the most significant achievements in object and pedestrian detection. The section starts by reviewing past solutions and variations. Later, the review focus on modern state of the art object detection techniques.

1.4.1. Pedestrian Detectors based on Histograms of Oriented Gradient Features

One of the first modern attempts to create a pedestrian detection algorithm was proposed by Dalal, Triggs (2005). As a key element, histograms of oriented gradient (HOG) (see Fig. 1.7) descriptors were used, which showed a significantly outperforming existing techniques for human detection.

Authors studied the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks showed that all are important for good results. The proposed approach gives near-perfect separation on the original Fleet *et al.* 2000 pedestrian database, so they introduced a more challenging dataset called INRIA containing over 1800 annotated human images with an extensive range of pose variations and backgrounds. However, research of (Bilinski *et al.* 2009; Sidla, Rosner 2007; Taiana *et al.* 2013) showed that INRIA is not enough to validate real-life situations, the dataset barely solve pedestrian occlusion in various poses and cases, and annotations are not precise.

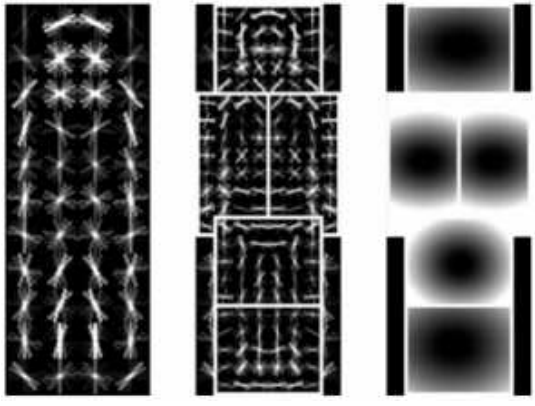


Fig. 1.7. Illustration of estimated HOG features, according to De Smedt 2015

To address occlusion and provide a wider variety of samples than INRIA a second attempt was made by Dollár *et al.* 2012. Authors addressed to properly evaluate a general pedestrian detection performance by creating a large scale visual spectrum pedestrian detection benchmark dataset containing approximately 10 hours of resolution 640×480 images captured at 30 Hz taken from a driving vehicle through regular traffic in an urban environment called Caltech Pedestrian Detection Benchmark. The dataset contains about 250,000 frames (in 137 approximately minute-long segments) with a total of 350,000 bounding boxes, and 2300 unique pedestrians were annotated. This dataset was well used in benchmarking various object detection algorithms.

One of the first attempt to test HOG detector on large scale FIR domain dataset was proposed by Hwang *et al.* 2015. Authors captured the aligned multi-spectral (RGB color + FIR) images dataset called KAIST. All the image pairs were manually annotated (person, people, cyclist) for the total of 103,128 dense annotations and 1,182 unique pedestrians. The HOG detector reached 54.40% accuracy.

1.4.2. Object Classification using Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for classification or regression problems. In computer vision SVM is wide used Bishop 2006 because of low computation efficient performance and accuracy.

Wang *et al.* 2009 was one of the first authors who reached an outstanding performance and increased HOG accuracy by using Local Binary Pattern (LBP) as the feature set. The authors propose a novel human detection approach capable of handling partial occlusion. Two kinds of detectors, i.e., a global detector for whole scanning windows and part detectors for local regions, are learned from the training data using linear SVM. For each ambiguous scanning window, the authors constructed an occlusion likelihood map using the response of each block of the HOG feature to the global detector. The Mean-shift approach then segments the occlusion likelihood map. The segmented portion of the window with a majority of negative responses is inferred as an occluded region. Suppose the partial occlusion is indicated with a high likelihood in a specific scanning window. In that case, part detectors are applied on the unoccluded areas to achieve the final classification on the current scanning window.

According to Hoang *et al.* 2014 another way for enhancing the accuracy and improve the speed of a pedestrian detection system one way is by using variant scale block based HOG features along with a hybrid of boosting and SVM techniques. The boosting technique is used as a global system, while SVM is used as a weak classifier inside of AdaBoost. The set of HOGBs is used to input of each SVM. The boosting technique is also used for selecting high discriminative

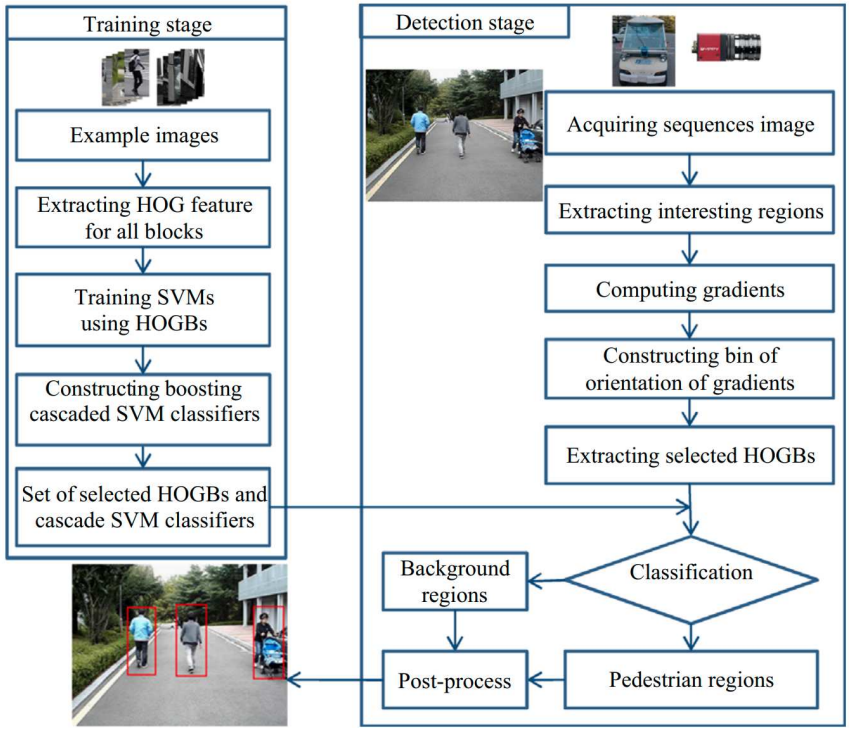


Fig. 1.8. Illustration of Boosting HOG with SVM Hoang *et al.* 2014

HOGBs to reduce the size of data in the classification stage. To reduce computational time of the system, the classification stage is constructed based on cascade boosting SVMs. An overview of the author proposed method is presented in Fig. 1.8.

Another attempt to increase HOG accuracy was revisited by Bilal, Hanif 2020, where authors found that HOG-linear SVM detector grossly underestimate. A proposed authors method accomplishes this by considering only a small set of the most relevant training examples and mitigating the class imbalance problem through manipulation of miss classification cost ratios. Authors also showed that by following their methodology, HOG can achieve a 14% lower miss rate.

1.4.3. Object Classification using Fuzzy logic

In the proposed work Mahapatra *et al.* 2013 authors used a fuzzy logic model a robust background for object detection. Three different features are extracted from the contours of the detected objects. These features were aggregated using fuzzy

inference system. Then human contour is identified using template matching. The proposed method consists of four main steps; Moving Object Detection, Feature Extraction, Feature Aggregation, and Human Contour Detection. The proposed method was tested with outdoor videos of humans and other objects and achieved a 97.8% correct detection for human. From the comparison study, it is observed that, proposed method is one among the efficient algorithms for classification of foreground objects, however the proposed method will not work if there is occlusion.

1.4.4. Feature Extraction and Classification using Convolutional Neural Network

A Convolutional Neural Network (CNN) according to Hussain *et al.* 2019 is a Deep Learning algorithm which can take in an input image, assign importance to various aspects/objects in the image and be able to differentiate one from another. The preprocessing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters/characteristics. The architecture of a CNN is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlap to cover the entire visual area.

In particular, object detection and pedestrian detection share a very similar pipeline. For both, some candidate regions have to be identified by means of a sliding window approach or more complex region proposal algorithm. Then, considering object detection, each region should be analyzed to check whether it contains an object and, if so, identify the class of such object. Instead, for pedestrian detection each proposal should be analyzed in order to check whether it contains a human shape. For both tasks such last stage of detection can be effectively accomplished resorting to a properly trained classifier.

Szarvas *et al.* 2005 shows one of the first attempts to use CNNs instead of SVM for pedestrian detection. Author method achieves high accuracy by automatically optimizing the feature representation to the detection task and regularizing the neural network. Authors evaluate the proposed method on a difficult database containing pedestrians in a city environment with no restrictions on pose, action, background and lighting conditions. The false positive rate (FPR) of the proposed CNN classifier is less than $1/5^{\text{th}}$ of the FPR of a support vector machine (SVM) classifier using Haar-wavelet features when the detection rate is 90%. The accuracy of the SVM classifier using the features learnt by the CNN is equivalent to the accuracy of the CNN.

A Comparison of human detection performances with HOG-SVM and CNN was held by Aslan *et al.* 2020. In the research was shown that CNN is more successful in human detection. It was observed that in the case of pedestrian occlusion, detection by CNN provides a stronger estimate. Similar study performed by Guo, Zhan 2018/05 revealed that CNN based detector is 17.01% more accurate than HOG.

1.4.5. Deep Neural Network based Feature Extraction and Object Classification

Deep Neural Network (DNN) is a part of CNN, used as well as the state of the art in terms of accuracy for a number of computer vision tasks such as image classification, object detection and segmentation, often outperforming the previous gold standards by a large margin according to Tomè *et al.* 2016. A main difference between CNN and DNN is that DNN having more than 4 layers. In literature DNN usually referred to CNN, since CNN is used not to classify images only but for example do speech recognition and many other according to Kim *et al.* 2020.

One of the most significant contributors to DNN was made by Krizhevsky *et al.* 2012, where authors presented AlexNet which had only five convolutional layers represented in Fig. 1.9. By staking more layers, the AlexNet neural network was able to extract more features and perform better (Krizhevsky *et al.* 2012; Li, Zhou 2019). Authors also utilised a Graphics Processing Unit (GPU) to do convolutions which made a new era in computer vision.

However, growing network depth does not work by simply accumulating layers together. It turns out that DNN is hard to train because of the vanishing gradient problem according to Hochreiter 1998. As a result, as the network goes deeper, its performance gets saturated or even degrades rapidly.

As the vanishing gradient issue became more and more apparent, researchers

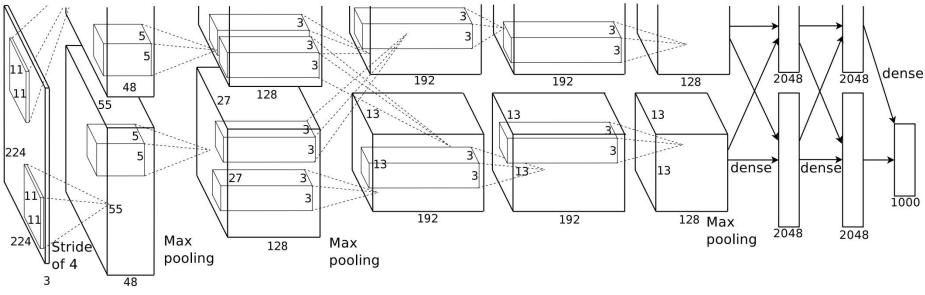


Fig. 1.9. AlexNet structure taken from Krizhevsky *et al.* 2012

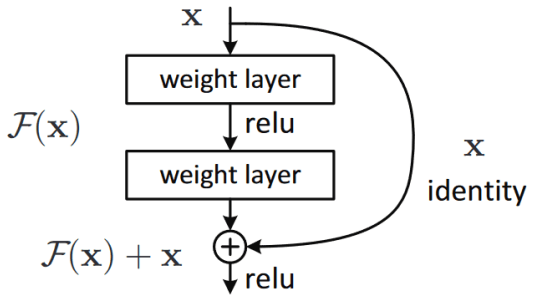


Fig. 1.10. Identity shortcut connection taken from He *et al.* 2016a

tried to add an auxiliary loss in a middle layer as extra supervision, but none seemed to stop the problem. As the solution, there was ResNet neural network introduced by He *et al.* 2016a. The core idea of ResNet is based on identity shortcut connection that skips one or more layers, as shown in the following Fig. 1.10.

The same authors introduced batch normalization, which stabilizes the learning process and dramatically reduces the number of training epochs required to train deep networks.

As a result of the major breakthrough, multiple variations came out of shortcut connection like ResNext (Xie *et al.* 2017b). The ResNext paper refers to the number of branches or groups as the cardinality of the ResNext cell represented in Fig. 1.11. It performs a series of experiments to understand relative performance gains between increasing the cardinality, depth, and network width. The experiments showed that increasing cardinality is more effective at benefiting model performance than expanding the network's width or depth.

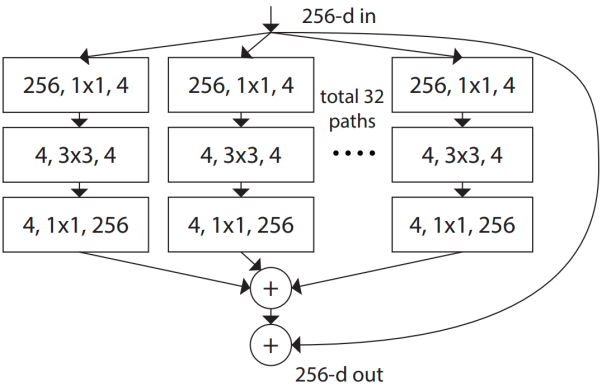


Fig. 1.11. ResNext block taken from Xie *et al.* 2017b

1.4.6. Modern Deep Neural Network based Object Detectors

Modern DNN based object detectors may be categorized into two categories according to literature. Figure 1.12 represents a schematic diagram of one-stage and two-stage detectors. The shared part between the two categories are the backbone and the feature map layer parts. At this stage, the object classification is usually implemented by using the following variations of structures:

- VGG (Simonyan, Zisserman 2014);
- ResNet;
- ResNext
- Darknet (Redmon, Farhadi 2018a).

The feature map layer is a bridge between the backbone and head where different layers are interconnected and composed of several paths. Typically, at this stage, researchers include different Feature Pyramid Networks (He *et al.* 2014; Liu *et al.* 2017a; Redmon, Farhadi 2018a) and Path Aggregation Network (Liu *et al.* 2018) (PANet). A head is the part where actual detection is taking place.

The state of the art of single-stage object detectors includes YOLO with it's several versions (Bochkovskiy *et al.* 2020a; Redmon *et al.* 2016; Redmon, Farhadi 2017, 2018a), SSD (Liu *et al.* 2016), ResNet50, ResNext50 and RetinaNet (Lin *et al.* 2017). The reason single-stage object detectors became useful in object detection applications is real-time performance and accurate enough object detection. YOLO and minimized version called TINY YOLO were one of the first detectors, working at 17–30 FPS with accuracy up to 43% mAP, outperforming SSD by 14% mAP Bochkovskiy *et al.* (2020b). Prihatmaja, Widyantoro 2019 also were evaluating SSD and YOLO DNN model variations by using KITTI (Geiger *et al.* 2012) dataset reaching 57.9% mAP and 19.61 FPS.

The state of the art two-stage detectors are R-CNN (Ren *et al.* 2015a), Faster R-CNN (Ren *et al.* 2015b), R-FCN (Dai *et al.* 2016) and FPN (Lin *et al.* 2017). These detectors are very slow because of a complex feature extraction stage and complicated structure. Two-stage detector implementations currently are able to process images reaching only from 0.1 to 5 FPS (Dai *et al.* 2016; Kaarmukilan *et al.* 2020; Le *et al.* 2016; Ren *et al.* 2015a,b). However, such detectors has comparatively excellent accuracy of up to 69% mAP according to Jiao *et al.* 2019; Soviany, Ionescu 2018 and running them on most popular benchmarking datasets (Caltech (Dollár *et al.* 2011), KITTI, ImageNet (Russakovsky *et al.* 2014), PASCAL VOC (Everingham *et al.* 2010) and MS COCO (Lin *et al.* 2014)).

One of the challenges for both detector categories' backbone and feature map layers is the multi-scale object detection. To address this problem, researchers usually stack different size layers on each other to extract features at different scales and join them later in a feature map layer. For example, the YOLOv2 backbone has

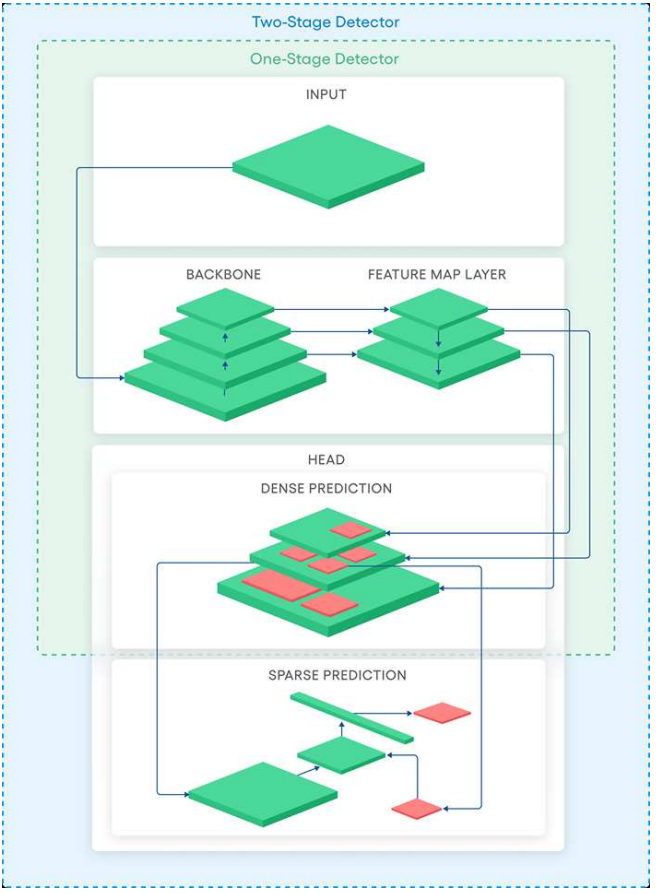


Fig. 1.12. A schematic diagram for comparison of one-stage and two-stage object detectors

19 layers, and YOLOv3 has 53 layers. Such modification has affected the detector's accuracy on the MS COCO dataset by giving a 9.8% mAP increase. However, the speed has decreased from 40 to 20 FPS.

For pedestrian detection researchers try to combine visual spectrum with FIR domain data (Jegham, Khalifa 2017; Shopovska *et al.* 2019; Takumi *et al.* 2017) and apply YOLOv1, YOLOv2, YOLOv3, Faster R-CNN, R-FCN or slightly modified versions where accuracy is ranging from 66% mAP to 79% mAP. An actual FIR domain accuracy was tested by thermal camera manufacturer FLIR and SSD detector and reached 79.4% mAP (FLIR Systems Inc 2020). A modified version of SSD was also tested by Chen, Shin 2020 and reached from 87.68% mAP to 97.5% mAP and YOLOv2 ranged from 58.5% mAP to 80.5% mAP.

1.4.7. YOLO Architecture

YOLOv1 is a 19-layer neural network detector also known as Darknet-19 which one of the first single-stage detectors allowing real-time object detection. YOLOv2 used am architecture borrowed from Darknet-19 with 11 more object detection layers. With a 30-layer architecture, YOLOv2 often struggled with small object detections caused by the loss of fine-grained features on downsampled input. YOLOv2 used identity mapping to solve this, concatenating feature maps from a previous layer to capture low-level features. However, YOLOv2 architecture lacked some of the essential elements that are now staples in most state-of-the-art algorithms, like residual blocks, skip connections upsampling.

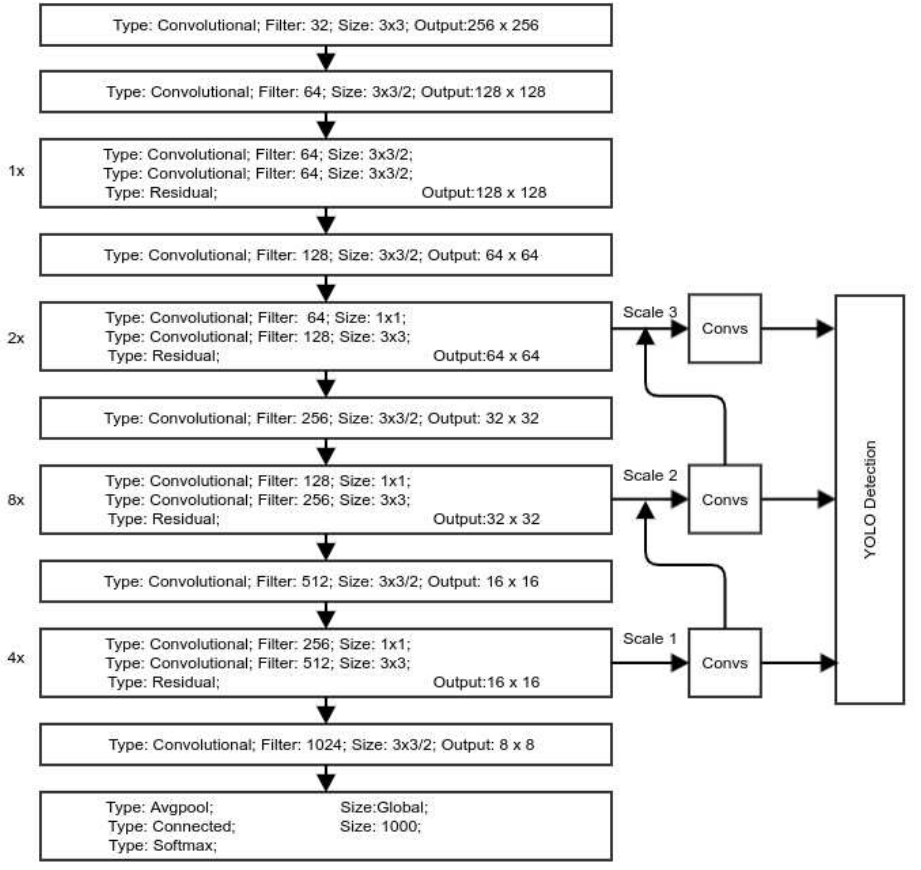


Fig. 1.13. Structure of the YOLOv3 detector

The newer architecture of YOLOv3 (see Fig. 1.13) incorporates residual skip connections and upsampling. The most salient feature of v3 is that it makes detections at three different scales. YOLOv3 is a fully convolutional network, and its eventual output is generated by applying a 1×1 kernel on a feature map. In YOLOv3, the detection is done using 1×1 detection kernels on feature maps of three different sizes at three other places in the network.

1.5. Conclusions of the First Chapter and Formulation of Dissertation Tasks

As the literature review showed there basically three main things to conclude:

1. The analysis of NIR, SWIR, MWIR, FIR sensor sensitivity to thermal radiation showed that the FIR domain provides the best-detailed view of human body features. These wavelength bands are less susceptible to moisture so that ADAS systems can perform more accurately during bad weather conditions.
2. FIR domain images are sensitive to background temperature, for example during daytime, there is a very little temperature-related pixel intensity difference between pedestrian and background objects.
3. The analysis of different pedestrian detection techniques showed, that the detectors based on HOG features provide up to 54.4% pedestrian detection accuracy.
4. The analysis of modern object detectors' performance, stated in the literature showed, that DNN based systems showed 66%–97.5% mAP in object detection tasks.
5. The comparison of published experimental investigations, related to pedestrian detection, showed that there is a lack of FIR domain research in specific weather conditions like rain, snow, fog and etc.

According to the result of the review, the following tasks were formulated:

1. Improve the processing speed of pedestrian detector based on Histogram-Oriented Gradient features, working on Central Processing Unit.
2. Develop an adaptive pedestrian detector prototype based on ambient temperature impact on Far Infrared Radiation images.
3. Optimise convolutional neural network based pedestrian detector for Edge-Computing Prototype.
4. Investigate dataset augmentation techniques to improve pedestrian detection in severe weather.

2

Investigation of the Pedestrian Detector Prototype

According to the literature review, currently, two options need to be investigated for pedestrian detection. Firstly, the HOG-based pedestrian detector is analyzed, and performance acceleration is proposed to see if the HOG detector should be considered the primary research detector. Since HOG experiment reveals drawbacks, the next and all other studies are focused on DNN type pedestrian detectors. For this reason, existing datasets were analyzed for DNN training. There were two significant datasets found. Their annotations were carefully investigated, and inconsistencies in annotations and lack of data diversity were discovered. For this reason, a prototype was created to collect a new pedestrian dataset, and a new dataset was published for pedestrian detection in severe weather conditions. Then, two real-time detectors were trained on this new dataset, and accuracy improvements were suggested. Finally, identical detectors were trained on a previously analyzed dataset and compared with the newly collected dataset. The section ends with conclusions and recommendations for the following experiments. The research results presented in this chapter are published in four author publications (Tumas, Serackis 2018, Tumas *et al.* 2018, Jonkus *et al.* 2018 and Tumas *et al.* 2020).

2.1. Performance Comparison of Pedestrian Detectors

A typical HOG features-based detector utilizes a sliding window approach that moves across the image, and pedestrians are detected in each window. Such a pedestrian detection method makes computations very inefficient because thousands of shifts need to cover the entire image. However, thermal imagery provides an essential feature because all warm objects in the captured background are visible as brighter objects. Such property can be used for finding pedestrians in these regions only. Firstly in this section, the HOG-based classifiers using hand-crafted features are analyzed and modified for real-time pedestrian detection using a Central Processing Unit (CPU) on FIR domain images. Later, deep learning-based detectors are tested on publicly available datasets, and performance is evaluated using a default configuration. Finally, a modification is provided to HOG detector for pedestrian detection in real-time, and a study is executed using empirically defined background subtraction solution versus automated solutions.

2.1.1. Histograms of Oriented Gradient Feature based Pedestrian Detection

In this subsection a HOG based detector is evaluated. Since a traditional way of running HOG is based of sliding window which causes a noticeable delay and is computationally inefficient. An optimization is proposed to gain more FPS. The proposed algorithm (see Fig. 2.1) consists of four processing steps. In the *first step* of the algorithm, video frames are captured form FIR camera and converted to single channel images. At the *second step*, a background subtraction algorithm is applied. Background subtraction extracts the brightest spots (objects with most intensive thermal radiation) in the image.

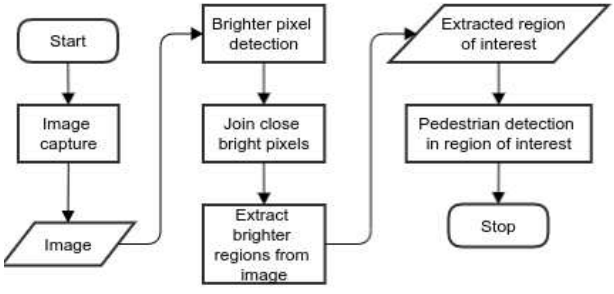


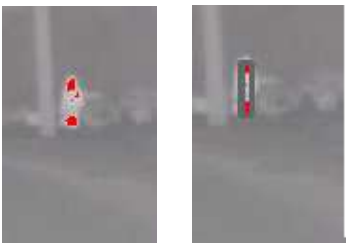
Fig. 2.1. Proposed algorithm based on background subtraction

Fig. 2.2. Illustration of a frame with thresholded bright spots coloured in red

Two automatic background subtraction algorithms were tested and compared with an empirical selection of the threshold in our experimental investigation. For comparison of automatic background subtraction algorithms, we have selected method proposed by Kurita *et al.* 1992 which is known as Otsu and Gaussian adaptive threshold algorithm, where the threshold value is a Gaussian-weighted sum of the neighbourhood values minus the manually defined constant. Additional empirical threshold selection was also considered because the dynamic range of FIR images is highly related to the temperature outside.

Background subtraction, applied on FIR images, usually gives a set of distributed spots with clusters belonging to the same object (see Fig. 2.2). To connect image pixels that belong to the same object, the *third step* of the algorithm includes morphological operations like erosion and dilation for connecting the neighboring pixels in the image based on Bradski 2008 research. An illustration of this step is given in Fig. 2.3.

The third step of the pedestrian detector is essential from the performance perspective. It connects closely situated image regions, which are received after



a) b)

Fig. 2.3. Morphological operations applied on an image: (a) initial image on the left; (b) operation applied is on the right

Fig. 2.4. Detected 4 contours which are cropped and passed to the detector

background subtraction. The number of final image segments (treated as a potential pedestrian) is directly related to the processing time required for validation of each image segment (candidate).

At the *fourth step* of the algorithm, a rectangular window centered on the previously extracted image segment is used to crop the image visible in Fig. 2.4. A new smaller image is used as an input to the trained detector for validation of pedestrian presence in the image. The size of the rectangular window is always matched to the size of the classifier input.

2.1.2. Histograms of Oriented Gradient Classifier Preparation

The performance evaluation of the pedestrian detectors on FIR images was performed using three selected classifiers: HOG feature descriptor based pedestrian detector, F-RCNN and YOLOv2. Since the aim of the investigation was not evaluating accuracy but mainly focus on processing speed, the datasets, selected for testing, were different. The HOG detector was trained on INRIA dataset, the F-RCNN was trained on VOC2007 dataset and YOLOv2 was trained on COCO dataset images. None of the datasets contained FIR camera images.

The following parameters were set for the detectors: the HOG detector used 64×128 window size, 16×16 block size, 8×8 block stride, 8×8 cell size, and the L2-Hys (Liu *et al.* 2014) histogram normalization method with threshold of 0.2; the F-RCNN was setup up with the confidence threshold of 50% and image size of 640×480; the YOLOv2 was setup up with the confidence level of 50% and image size of 416×416.

To compare the efficiency of the FIR image pre-processing algorithm proposed in this dissertation, to the initial workflow of the pedestrian detectors, the experimental investigation was divided into two parts. In the first part, the whole image from FIR camera was sent to unmodified HOG, F-RCNN and YOLOv2

algorithms. In the second part, the sliding window approach was removed from HOG and F-RCNN pedestrian detectors' pipeline and our proposed modification was integrated into these algorithms. Since YOLOv2 detector is using grid instead of windowing, we were not able to perform out proposed modifications on this detector. Therefore, YOLOv2 was used only for performance comparison.

2.1.3. Histograms of Oriented Gradient Classifier Performance Estimation Results

The performance estimates (average FPS and Standard deviation (SD)) of the unmodified pedestrian detectors are presented in Table 2.1. All experiments were organized to run only on CPU (Intel i7-7820). The best result (7 FPS on average for whole test set) was achieved with the Tiny YOLO detector. However, this model was able to detect fewer classes and was not so accurate as YOLOv2. The classical HOG based pedestrian detector was the slowest one and reached only 0.5 FPS on average. In the second part of our experiment, there was evaluated a modified HOG and F-RCNN detectors. The core of modification was the elimination of the sliding window based technique and inclusion of the advanced window selection. The window selection used background subtraction and clustering of pixels in the neighborhood. Three different background subtraction algorithms were tested experimentally. The performance of the modified pedestrian detectors is given in the Table 2.2.

Since the proposed algorithm performance depends on background subtraction, it is represented in the Table 2.2 how Otsu, Gaussian and empirically selected threshold algorithms influence the whole processing time. There are three columns which represents each background subtraction algorithm processed when 4, 6, 8 and 10 objects are detected in the view. The best result was achieved with HOG detector, reaching 6 FPS with the empirically selected threshold (101 out of 256). Comparing to initial version of the HOG detector represented in the Table 2.1, the modified one performed 12 times faster. The manual selection of the threshold was important to be tested, because the threshold depends on the environment temperature which is not changing fast. Therefore, it is possible to include an automatic threshold detection (e.g., during startup) and apply this step to every frame.

Table 2.1. Average image processing speed of unmodified detectors

Detector	FPS	SD
HOG	0.5	0.17
Faster R-CNN	1.1	0.08
YOLOv2	2.3	0.21
Tiny YOLO	7.0	0.05

Table 2.2. Average performance of modified pedestrian detectors

Detector	Otsu (FPS)	Gausian (FPS)	Empirical (FPS)
4 objects + HOG	4.0	3.2	**6.0**
4 objects + Faster RCNN	1.3	1.2	1.6
Only 4 objects	10.0	8.0	13.0
6 objects + HOG	3.0	2.7	5.1
6 objects + Faster RCNN	0.7	0.9	1.3
Only 6 objects	8.0	7.1	11.0
8 objects + HOG	2.1	1.9	3.9
8 objects + Faster RCNN	0.5	0.6	0.9
Only 8 objects	5.1	5.3	8.8
10 objects + HOG	3.0	2.7	5.1
10 objects + Faster RCNN	0.4	0.3	0.7
Only 10 objects	2.4	2.1	3.5
No objects	12.0	9.0	15.0

The experimental investigation also showed that the performance of the algorithm with proposed modifications depends on the number of detected objects in the video frame (number of pixel clusters). It means that if there are no objects found, the algorithm can perform up to 15 FPS (see Table 2.2, "No objects"). Otherwise, if there were more than four objects found in the single frame, the performance decreased to 6 FPS. The experiment concludes that, such an approach is not applicable for real-time pedestrian detection since the algorithm depends on the high intensity of the pixel. The more high-intensity pixels are present, the more time is needed to process the frame. However, such algorithm could useful during the night, because images contains minimal number of intensive pixels.

2.2. Analysis of Currently Available Dedicated Datasets

According to research Sze *et al.* 2017 and Zhao *et al.* 2019, the accuracy of the detector vary according to the variety of dataset samples, detector input/type used, and implementation details. Currently, the use of DNN showed that to train the object detector, it is needed to have thousands of various pose and featured images. There are hundreds of available datasets for pedestrian detection in the visual spectrum containing thousands of images with various body poses, shapes, and different contextual features like PASCAL VOC 2012 (Everingham *et al.* 2011), KITTI, GM-ATCI (Silberstein *et al.* 2014), NICTA (Overett *et al.* 2008), INRIA and others. However, to train DNN on thermal imagery data, it is only possible to find just up to ten datasets which are free for use:

- CVC-09 (Socarras *et al.* 2013);
- CVC-14 (Gonzalez Alzate *et al.* 2016);
- FLIR-ADAS (FLIR Systems Inc 2018);
- KAIST (Choi *et al.* 2018);
- KMU (Jegham, Ben Khalifa 2017);
- LSIFIR (Khellal *et al.* 2015);
- OTCBVS (Davis, Keck 2005);
- RISWIR (Miron 2014);
- Terravic Motion IR (Davis, Keck 2005);
- SCUT (Xu *et al.* 2019).

KAIST and SCUT are the most representative datasets designed for ADAS use. Other datasets contains images from cameras installed in buildings, tripods, UAVs for the purpose of tracking or surveillance systems. Images, obtained with such camera platforms are unsuitable for ADAS according to Kim, Kim 2018.

2.2.1. SCUT Dataset

SCUT dataset is the biggest publicly available dataset. Unlike other datasets, it contains images captured from driving a car in areas like downtown, suburbs, campuses, and expressway roads. SCUT dataset consist of about 11 hours-long image sequences at a rate of 25 Hz. The image sequences were collected from 11 road sections under four kinds of scenes including downtown, suburbs, expressway and campus in Guangzhou, China. Currently it is the biggest publicly available dataset in terms of a number of frames and annotations (containing 216,000 frames and 448,000 annotations).

Fig. 2.5. Illustration of sample taken from SCUT dataset

The image resolution is another important aspect where the SCUT provides images captured with 384×288 sensor FIR camera and captured image is interpolated to 720×576 resolution. Finally, SCUT keeps strictly predefined labeling protocol (see Fig. 2.5) which is followed by six classes (walk person, squat person, ride person, people, person, and combined annotation person/people). As it is visible in illustration, the SCUT dataset also contains various densely populated frames with pedestrians in the crowd, which benefit in various object tracking challenges and detector accuracy measurements.

2.2.2. KAIST Dataset

KAIST is the second biggest available dataset containing multi-spectral images captured in visible and thermal domains visible in Fig. 2.6, having 95,000 frames of resolution 640×480 taken from on a vehicle-mounted camera. Recordings were taken in multiple areas like campuses, cities, and outskirts. All image domains were manually annotated with three classes (person, people, and cyclist) for a total of 103,128 annotations and 1,182 unique pedestrians. The picture shows day and night situations where pedestrians in thermal and visual data are represented. It is important to emphasize, that during the day thermal images might not be sensitive to pedestrians since ambient temperature might be higher that human body temperature.

2.2.3. The Drawbacks of Currently Available Datasets

SCUT and KAIST datasets were collected for ADAS applications. The main research object in the papers, where these datasets were used is detection of pedestrians. However, each of these datasets has some limitations. First of all, none of the datasets contain information about exact weather conditions while recording. For example, when it rains, the water and dirt coats cameras lenses, which causes an effect of a blurry image without precise contours and lowered intensity

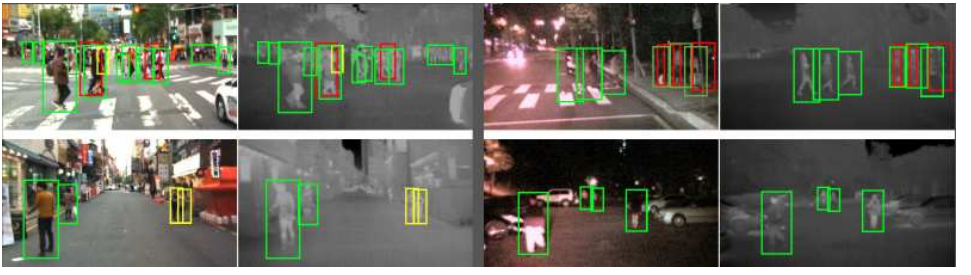


Fig. 2.6. Sample from KAIST dataset

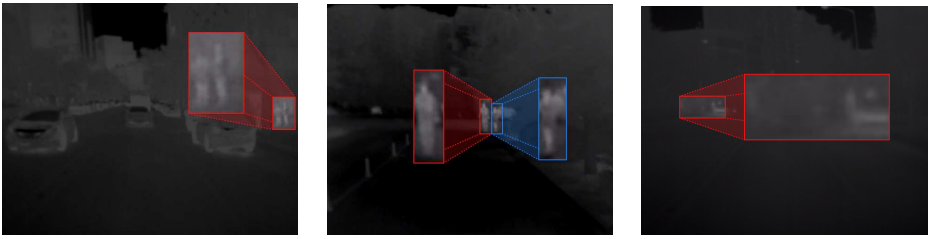
of pedestrian. Also, since ADAS systems uses sensor fusion, there is no information provided about the car parameters itself: speed, brake pedal status, steering wheel angle, outside temperature and many others.

While analysing SCUT dataset it was found, that there are many cases when several pedestrians touched with the hand is marked as a single annotation visible in Fig. 2.7. Also, when there is a case of group people (two pedestrians visible, others not), it was marked as a single group of people annotation, including partially visible pedestrians. This situation should be taken in to account, since it might cause accuracy issues.

The KAIST dataset also contains annotation issues. The most important factor in dataset is to keep annotation quality by complying with the same annotation protocol. It was checked more than 50,000 annotations of the KAIST dataset and found that in some cases a group of people is marked as one annotation visible in Fig. 2.8 (a), sometimes as separate visible in Fig. 2.8 (b) or some annotations looked miss-labeled visible in Fig. 2.8 (c) having no context related to pedestrian shape, outlook and body features.

In addition to existing findings, none of the datasets contain information about exact weather conditions while recording. For example, when it rains, the water and dirt coats cameras lenses, which causes an effect of a blurry image without precise contours and lowered intensity of pedestrian. SCUT authors mentioned that they were recording during December in Guangzhou, China, where the average rainfall is only 32 mm AM Online Projects (2019). Secondly, there is no tracking of the outside the temperature, which affects the image details. For example, during the cold winter days, the pedestrians look very bright, however during the hot summer nights, the pedestrians are blended with the background. In addition to this, none of the datasets reported in review contain a sufficient variety of samples

Fig. 2.7. Group of people as single annotation



a) b) c)

Fig. 2.8. Examples of images in KAIST dataset when pedestrians represented: a) in one annotation; b) in two annotations; c) miss-annotated

to create a detector, capable of detecting pedestrians in severe weather conditions. For these reasons it was decided to collect a new dataset called ZUT-FIR-ADAS (in this dissertation a short ZUT name is used), which was used for investigations, presented in this dissertation.

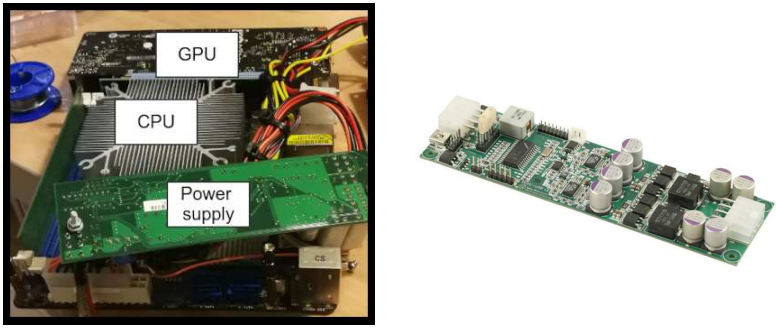
2.3. Prototype Development for Dataset Collection

To collect a new FIR domain dataset, it is not enough to have a traditional laptop in the car and a camera connected to it. Since the vehicle should be used on public roads – distractions might happen, and accidents might occur. In the case of an accident, the airbags might blow, and all the hardware lying on the seat can hurt the driver or the passenger. In order to prevent possible hurt and be as much as possible compliant within safety, it was decided to build and integrate a specially designed computing unit within the car.

2.3.1. Preparation of the Computing Unit

The computing unit is based on a Mini-ITX motherboard with an LGA1151 socket and I7 eighth-generation processor with the addition of an NVIDIA RTX2070 GPU, shown in Fig. 2.9 (a). A graphics card was connected with 90 degrees PCI-e adapter to minimize the height of the chassis. The computing unit was installed in Skoda Fabia MK2 Green Line 1.4 TDI car.

For powering the system, it is crucial to have an efficient and intelligent power supply. Firstly, the battery might be drained very quickly because the GPU and the CPU together consume more than 250 W. Secondly, during the engine's start, the voltage might drop to less than 6 V, which will be halting the system. To solve these problems, the power supply was constructed using two DC-DC (120 W and 200 W), shown in Fig. 2.9 (b) converters, which can boost the voltage during the



a) b)

Fig. 2.9. On-board computer: a) computing unit being assembled;
b) DCDC-USB-200 Intelligent DC-DC Converter

engine start, have protection over-discharge, and has the small form factor. The First 120 W power supply was dedicated to the CPU, and another 200 W power supply was for the GPU.

2.3.2. Development of Controller Area Network Bus Data Capture Solution

In order to collect data from CAN bus an Atmel SAM3X8E 32 bit ARM microprocessor with a clock speed of 84 MHz was used. An MCP2515 CAN Controller was used with an SPI interface connected directly to the instrument cluster CAN bus wiring. For network interface ENC28J60 Ethernet SPI interface microcontroller was used, and CAN message was converted to UDP packet and sent through a gigabit network. The UDP packets were selected to minimize packet size and gain speed, but the drawback is that some of the packets might be lost. In Fig. 2.10 is represented the application flow chart diagram, which shows the message broadcast throttling strategy dynamically adjusted on a basis. Since Anti-lock braking system (ABS) and Instrument Cluster Module (ICM) are streaming messages at a different rate, additional logic was implemented to persisting new messages in the memory and keep the UDP broadcast rate ten frames per second. However, on sudden data change, like the brake pedal is pressed, the broadcast rate is overridden, and a UDP message with changed data is sent. This strategy was chosen to minimize network overflow and overhead in thread synchronization.

The CAN data includes car speed, the brake status (brake released, foot on the brake, brake press), which was captured from the ABS module, and the outside temperature from the instruments cluster. The CAN message broadcast baud rate was 500 kbit/s where 1 bit error might accrue every 0.7 s.

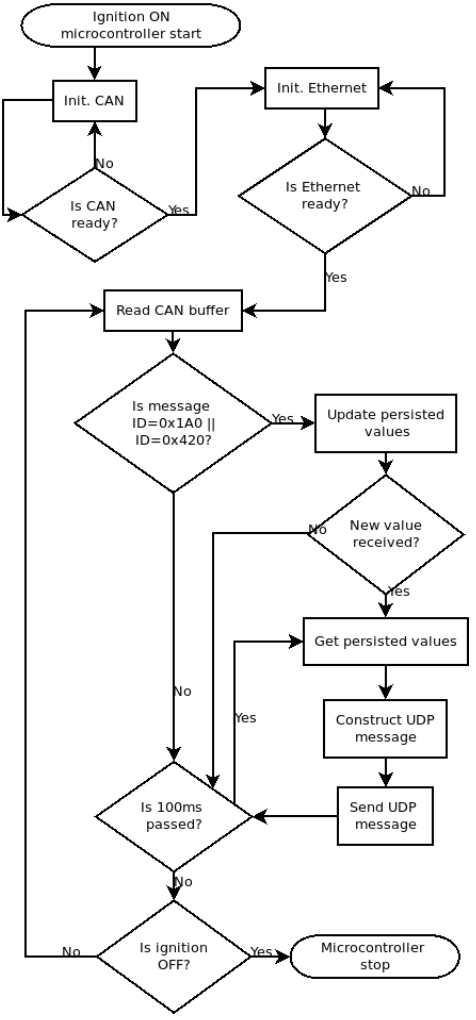


Fig. 2.10. The flowchart of CAN bus capturing program

2.3.3. Developing of Image Acquisition and Pre-Labeling Solution

The camera used for dataset image capturing was FLIR SC320 thermal camera with a spatial resolution of 320×240, capturing 16 bit frames at 30 FPS. To protect the camera and its lens against dust, rain, and dirt, every session, the camera has been wrapped (see Fig. 2.11 (a)) with a very thin plastic film used in the food industry. Experimentally, it was found that very thin plastic film allows thermal

a) b)

Fig. 2.11. Prototype installation: a) wrapped camera and placement; b) recording control monitor inside the car

energy to pass through it with a minor distortion represented in Fig. 2.12. The center of the roof was chosen for the location of the mounting point of the camera, in order to minimize the dirt coming from cars in front and to allow equal left/right side visibility. Additionally, the camera was re-calibrated for a better view before each recording.

In car's cockpit the touchscreen monitor was installed which is represented in Fig. 2.11 (b). The main purpose of it was to monitor and control the recording process and handle recorded data synchronization to the cloud. Also monitor helped organise the process and do not disturb while driving.

There was second application developed for data recording (see Fig. 2.13) and installed on the car computer to record data. The image capturing thread reads data from the camera using the GenICam protocol. Since this protocol is generic, there are many Software Development kit (SDK) providers for camera image capturing manipulation. In our case, we used the Baumer GAPI SDK (Baumer Group Corporation 2019), because it allowed us to get access to raw data, and SDK is compatible with OpenCV. After successfully grabbing the frame, we convert it to a single-channel 16 bit Mat object scaled to 640×480 resolution. Then, CAN data from UDP capture was copied by a thread locking semaphore. To minimize anno-

a) b)

Fig. 2.12. A comparison of camera view: a) with protection; b) without protection

tation work, we used the pre-annotation approach based on the TinyV3 (Redmon, Farhadi 2018b) neural network. The detector for pre-annotations was trained by 50,000 of images scaled down to 640×480 resolution taken fromSCUT dataset with excluded class named "People-?", reaching 67% mAP. Finally, we save corresponding frame ID with CAN data to comma-separated CSV file and the captured frame with YOLO-type annotations. A 4G modem was also installed for the network with four 1 Gbit ports used for CAN packets, FLIR camera video stream

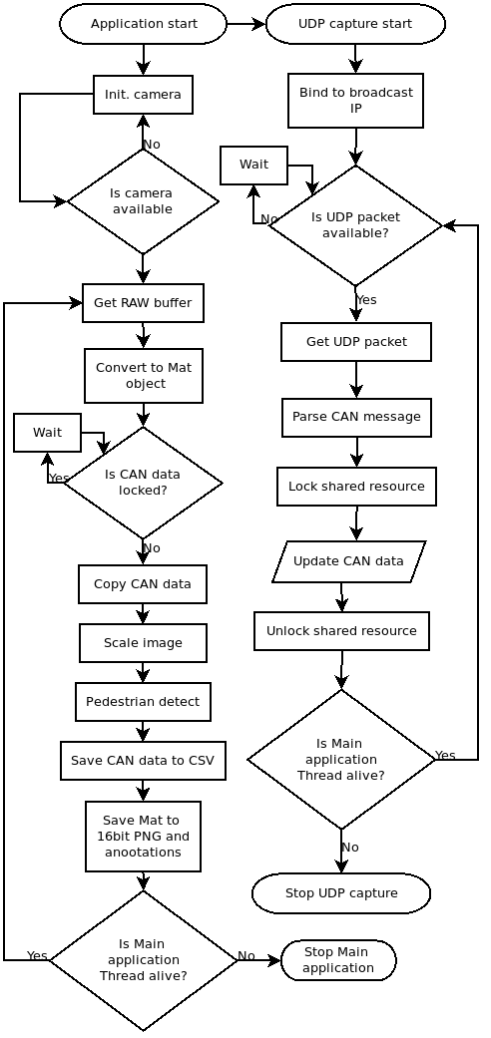


Fig. 2.13. The flowchart of data recording application

broadcast, and communication with computing unit. It was also deployed a cloud solution for synchronizing recorded data to the open source cloud-based storage service called Seafile.

2.4. Introduction of the New Dataset for Pedestrian Detection

For the recording location, we selected four European Union countries: Denmark, Germany, Poland, and Lithuania starting in the middle of autumn and finishing in the middle of winter. Law limitations were based on the selection criteria and General Data Protection Regulation (GDPR) rules, typical weather conditions based on season, car accident statistics, and traffic infrastructure. For example, we wanted to record in Austria. However, it is entirely illegal to use a camera there, and it is possible to face fines of 26,000 Euros (Reach PLC 2019; Schindler IT-Solutions 2019). Fortunately, Denmark, Germany, Poland, and Lithuania are camera friendly. Only in Germany is it required to mask car number plates, and people face for data publication. Fortunately, these restrictions do not apply to thermal vision spectrum imagery.

Denmark was selected because it has up to 19 days of rainfall in November, is the top country in traffic safety records, and has traffic infrastructure designed for cyclists. Germany is also one of the top countries in traffic safety, but it is rich in traffic infrastructure/regulations and has unlimited speeds on the Autobahn. The remaining countries, Poland and Lithuania, are rich by nature, have forest-surrounded roads with a high probability of animals' being present, are surrounded by small villages across the main highways, and are traffic heavy on transitive routes. The dataset contains frames driven through various weather conditions like fair weather, cloudy with a chance of rain, mild rain, heavy rain, and fog. The dataset includes ten road scenes: city center, old town, roundabouts, tunnels, city outskirts, one-way roads, two-way roads, highways, Autobahns. The ZUT also includes images driving with speeds up to 180 km/h, driving through capital cities (Berlin, Copenhagen, Warsaw, and Vilnius) and during morning and evening rush hours. The temperature is ranging from -0.5°C to 12°C. The detailed comparison is visible in Table 2.3.

2.4.1. Data Preparation for Deep Neural Network based Pedestrian Detection

For data preparation, the ideal approach is to have an equal train and test dataset. However, data annotation is a time-consuming process since it requires manual

labeling of data objects to use them as annotations. There are alternative automated approaches which use DNN, retrained on the minimal set of specific domain data, and then DNN performs pre-annotation of the dataset. However, this process still requires human verification.

Suppose there is no possibility to have an equal train and test dataset. In that case, the dataset is divided by the 80/20 (Afzal *et al.* 2017; Lanka *et al.* 2020; Mohanty *et al.* 2016) ratio also known as Pareto's rule (Tanabe 2018) or other variations like 70/30 (Polat *et al.* 2008) or 90/10 (Mustaqeem *et al.* 2018) depending on situation. The division is based on annotation height, width, and position in the image, and the goal is to cover the same feature space by test and train dataset division.

Typically annotation file is represented by class id and rectangular (top left and bottom right corner points) coordinates. However Darknet uses rectangular center point coordinate system. For example <classId> <x> <y> <width> <height>, where <x> is rectangle center point x coordinate, <y> is rectangle center point y coordinate, <width> is rectangle width and <height> is rectangle height. Also, it is important to mention that all these parameters are float and absolute values meaning that <x> and <width> vales are also divided by frame resolution width where <y> and <height> are divided by frame resolution height.

2.4.2. Analysis and Correction of Pre-Labeled Annotations

The annotation started on pre-labeled data. We have found that generally pedestrians were pre-annotated well, but the main work was to divide objects into different classes. For this process, we used Ybat: YOLO BBox Annotation Tool (Tech 2019). The ZUT dataset contains two sets of annotations. The first set is made of annotations used for the training. The benchmarking set was used to measure the accuracy of the detector in places where it was not trained.

Each set of annotations includes fine-grained labels divided into nine classes:

Table 2.3. Comparison of data diversity on KAIST, SCUT and ZUT datasets

Parameter	KAIST	SCUT	ZUT
frames	95,000	216,000	110,000
classes	4	6	9
annotations	103,000	448,000	122,000
road scene	3	4	10
pedestrian distance	2.4–61 m	4.6–132 m	10–100 m
data depth	8 bit	8 bit	16 bit
temperature	not measured	not measured	–0.5 to 12 °C
maximum speed	not measured	80 km/h	180 km/h
spectral range	7.5–13.0 µm	8.0–14.0 µm	7.5–13.0 µm

Pedestrian, Occluded, Body-parts, Cyclist, Motorcyclist, Scooterist, Unknowns, Baby carriage, Pets, and Animals. The cardinality of each class or available annotations are presented in the Table 2.4 and Table 2.5. An individual person is labeled as a pedestrian when it is walking, running, standing, or when at least 60% of the individual person's body is visible. Occluded class is used when it is impossible to distinguish an individual from a group of people. The Body-parts class is mainly used when less than 40% of the individual person's body is visible. For example, individual person is behind a car, and his/her head is visible, then the individual person is considered part of the Body-parts class. The same strategy is used with legs and hands. Cyclists, Motorcyclists, and Scooterist are labeled separately because there are a lot of scooterists and motorcyclists in Germany, but cyclists are dominant in Denmark. The Unknowns class is mainly used for objects similar to pedestrians, like a tilted tree or a hot traffic sign or traffic light. The Baby Carriage class doesn't contain many annotations, but we saw that there is some visibility of children in it. The last class Pets and Animals, includes domestic cats and dogs mainly, but there are several foxes and rabbits annotated as well.

The database was collected during hours of driving in different real-life scenarios without any artificial arrangements. For this reason, the database reflects real situations encountered on the road, and the number of object instances is fewer in some classes than in others. For example, compared to pedestrians pets and animals are extremely rare in the city environment.

In Fig. 2.14 presents pedestrian visibility changes through different weather conditions. When the recording started, the camera is clear, and pedestrians are very clearly visible in Fig. 2.14 (a). However, the visibility becomes indistinct after driving several minutes in the rain visible in Fig. 2.14 (b), where only the warmest are visible. Similarly, the view looks darker during the mild rain visible in Fig. 2.14 (e) and driving in the fog represented in Fig. 2.14 (f). During the heavy rain, there are almost no thermal objects visible except exhaust pipe and some contours of the car represented in Fig. 2.14 (d). The opposite situation is visible

Table 2.4. Annotation count per pedestrian related class

	Pedestrian	**Occluded**	**Cyclist**	**MotorCyclist**	**Scooterist**
Train	59,649	4,008	7,908	173	94
Test	21,083	1,112	2,355	49	0
Total	**80,732**	5,120	**10,263**	**222**	**94**

Table 2.5. Annotation count per specific other than pedestrian class

	Body Parts	**Unknowns**	**Baby Carriage**	**Pets and Animals**
Train	16,611	14	27	140
Test	9,091	1	10	107
Total	25,702	15	37	**247**

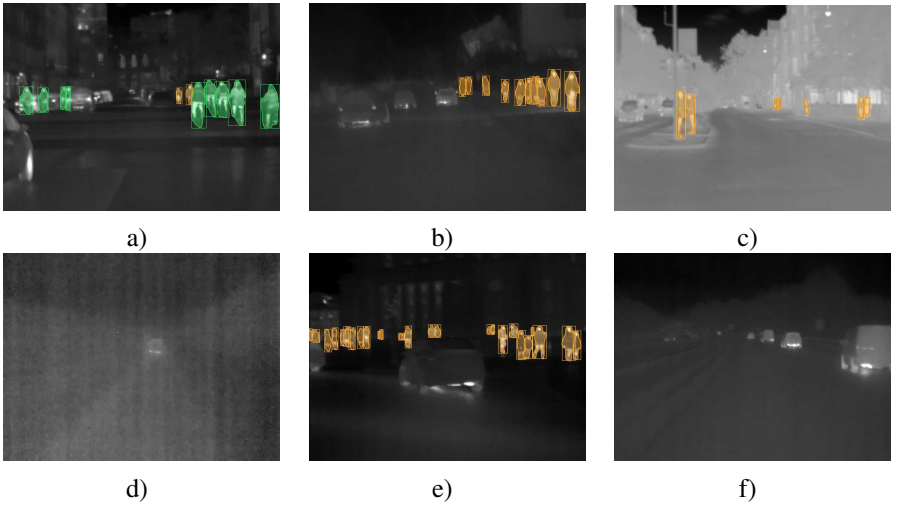


Fig. 2.14. ZUT Dataset examples: a) cyclists in Denmark; b) driving in the rain after 2 minutes; c) pedestrian visibility in the frost; d) heavy rain; e) mild rain; f) driving in Autobahn during the fog

during the frost, because the background objects and the people are very bright in the image, which makes hard to distinguish them further than 60 m represented in Fig. 2.14 (c).

Some pedestrian detection applications, like detecting pedestrians in the crowd is need to have an indicator of pedestrian annotation intersection (Chan *et al.* 2015; Zhou, Yuan 2017). For this reason, a histogram was created and represented in Fig. 2.15 which shows intersected area distribution through the dataset. As fig-

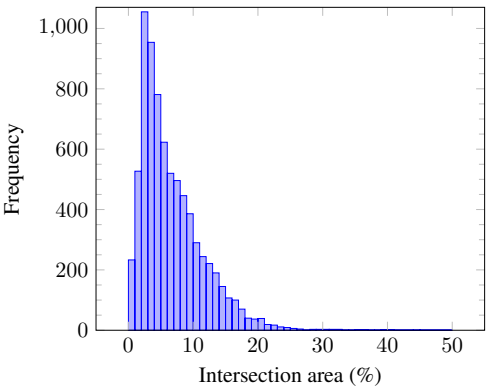


Fig. 2.15. Pedestrian annotation intersection histogram

ure shows, the majority of intersection is concentrated next to 8% of overlapping area. After this peak, the remaining intersection area drops drastically, and close to 20% there are a very minimal number of annotations.

2.5. RAW Image Preprocessing and Darknet Modifications

For the training, there was decided to use two versions of YOLO DNN: YOLOv3 and TinyV3, because theoretical research showed that both detectors can be used for real-time performance and are regarded as state-of-the-art detection and recognition approaches. In order to increase the accuracy of YOLO, it was decided to train DNN by using 16 bit depth images data, which provide up to 256 times more information than a regular 8 bit images. To support this feature, the latest Darknet implementation was used (maintained by AlexeyAB (Bochkovskiy 2019a)) and the following changes were applied to the framework:

- the image loading function;
- the normalization function;
- the data augmentation function.

The image loading function was changed in training and testing phases to load 16 bit images. The normalization function was changed by dividing the intensity value of over 65,535 instead of 255. Finally, the augmentation function was changed by adding a threshold mechanism to cut-off high intensity pixels (TI) by using outside temperature captured from CAN bus. To design a cut-off mechanism (see Fig. 2.16 (a)), all dataset annotations were used. The corresponding temperature averaged the maximum and minimum pixel intensity, and the quadratic function was fitted upon the lowest maximum intensity points. In case the maximum intensity was below the function, the value is kept without applying the threshold. This image processing resulted in hot objects like tires, disk brakes, exhaust pipes, windows, and chimneys to be less bright and provide more contour and pattern information. This also enhanced the visibility of "hot" pedestrians who, for example (see Fig. 2.16 (c)), had driven the long distances in warm cars.

The remaining YOLO configuration was left unchanged except the input resolution was changed to 640×480, the anchors were recalculated using k-means algorithm, and the dimensions of the input channel changed to 1. The ZUT training set was divided into two parts, where 80% of images used for training and 20% for testing. Additionally, we excluded classes like Occluded, Unknowns, Baby carriage, Pets and Animals from the dataset and merged the remaining classes into single category.

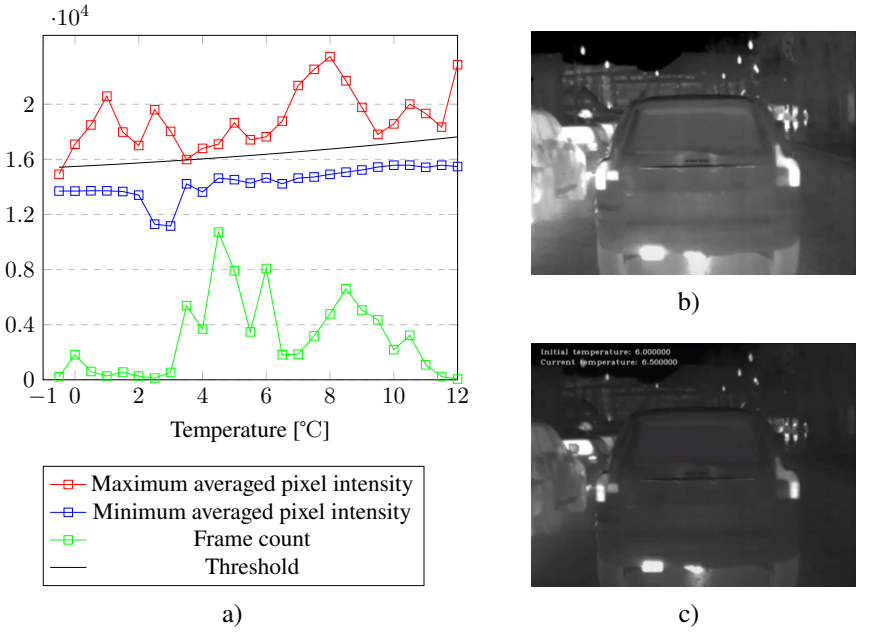


Fig. 2.16. Visual representation of: a) intensity and temperature distribution per dataset; b) initial capture; c) processed capture

2.5.1. Experimental Investigation of Two Candidate Architectures for the Detector

The training results are presented in Table 2.6. In this table the information on experiment configuration and obtained results is provided. Two variants of mAP (for different IoU: 50 and 25), loss and number of iteration are referenced. Initially, the best performance was registered for unmodified YOLOv3 DNN, reaching the accuracy of 80.5% mAP. The TinyV3 acheived only 66.3% mAP after 243,000 iterations, which indicates that the network cannot extract more features from the dataset. In this case, we have increased input resolution to 640×480, and YOLOv3 improved by 4.9% mAP reaching 85.4% mAP. However, the TinyV3 increased accuracy by 12.8% mAP which is almost the same as YOLOv3 with an input resolution of 416×416. Furthermore, the TI proposed method additionally increased accuracy by 3.7% mAP for YOLOv3 and by 3.2% mAP for a TinyV3. Finally, we converted the training set to 8 bit images with intensity threshold method to compare the 16 bit images versus 8 bit images. The 8 bit TinyV3 reached 71.1% mAP and we stopped the training at 78,000 iterations because the loss stopped to decreasing. The YOLOv3 reached 79.6% mAP at 123,000 iterations and the loss stopped decreasing as well. To sum up, by using TI and 16 bit images versus

8 bit images the accuracy was increased by 9.5% mAP for YOLOv3 detector and 11.2% mAP for TinyV3 detector.

The visibility distance is another crucial aspect for evaluation of detection accuracy(see Fig. 2.17) shows, that pedestrians whose height is 1.88 m in the 100 m distance would be equal to 21 pixels height in captured frame. Pedestrians, situated at distances more than 100 m are very poorly visible. However, Table 2.9 and Table 2.10 shows, that the training set contains about 39,000 of annotations at a distance between 61–80 m where the YOLOv3 with 16 bit input and applied intensity threshold method reached 82.9% mAP on the training set and 52.4% mAP on the test set. The TinyV3 with 16 bit input and applied intensity threshold method performed worse and entered 71.5% mAP for the training set and 51.7% mAP for the test set. The second biggest interval is from 41–61 m. This interval was the best for both detectors gaining 92.9% mAP for YOLOv3 (16 bit and intensity threshold method), 75.6% mAP for TinyV3 (16 bit and intensity threshold method) for the training set. On the test set, the detectors showed similar results where YOLOv3 (16 bit and intensity threshold method) got 76.5% mAP and TinyV3 (16 bit and intensity threshold method) 59.4% mAP. The worst results were obviously from 81 m going to infinity, where the object becomes very small in the image. In addition to this, there was also another interest to measure the detection precision on classes excluded from the training. Table 2.7 shows that the YOLOv3 and TinyV3 are detecting Body parts, Unknowns and Baby carriages with very low rate of mAP, which means that YOLOv3 and TinyV3 are trained sufficiently having very minimal miss-rate.

2.5.2. Validation of Experimental Investigation Results

In validating our training results, we faced two challenges. The first one is that there is no 16 bit thermal dataset used for pedestrian detection application, primarily used in competitions like VOT Challenge (Kristan *et al.* 2016). The second

Table 2.6. ZUT training results with 8 bit and 16 bit images at IoU 50 and 25

Resolution	Version	Depth, bits	Loss	mAP50	mAP25	Iteration
416 × 416	YOLOv3	16	0.2448	80.5	91.5	251,000
416 × 416	TinyV3	16	0.2954	66.3	86.0	243,000
640 × 480	YOLOv3	16	0.2514	85.4	92.5	220,000
640 × 480	TinyV3	16	0.2681	79.1	92.3	250,000
640 × 480	YOLOv3 + TI	16	**0.1514**	**89.1**	**95.4**	383,000
640 × 480	TinyV3 + TI	16	**0.1681**	**82.3**	**94.2**	420,000
640 × 480	YOLOv3 + TI	8	0.1914	**79.6**	92.3	123,000
640 × 480	TinyV3 + TI	8	0.2414	**71.1**	89.1	78,000

Fig. 2.17. Pedestrian height distribution per distance

issue is annotation methodology, since pedestrians can be annotated in many ways and poses, the dataset used for direct comparison should be annotated in the same way to have the most accurate results.

For those reasons, we decided to use our dataset 8 bit version with processed images by TI and compare detectors YOLOv3 and TinyV3 against the SCUT test dataset. Besides, we had deeply analyzed the SCUT dataset annotation methodology and found that there are many cases when two pedestrians touched with the hand is marked as a single annotation. Also, when there is a case of group people

Table 2.7. Detector evaluation on classes which were excluded from training

Version	Dataset	mAP	IoU, %	TP	FP	FN	Avg. IoU, %	Prec.	Rcl.	F1
YOLOv3	Train	**13.2**	50	743	1,021	3,657	27.64	0.42	0.17	0.24
TinyV3	Train	**3.3**	50	368	1,600	4,032	11.37	0.19	0.08	0.12
YOLOv3	Train	21.3	25	186	166	694	32.73	0.53	0.21	0.30
TinyV3	Train	21.6	25	940	824	3,460	32.32	0.53	0.21	0.30
YOLOv3	Test	**3.1**	50	118	394	2,109	14.81	0.23	0.05	0.09
TinyV3	Test	**0.8**	50	83	665	2,144	6.80	0.11	0.04	0.06
YOLOv3	Test	5.8	25	159	353	2,068	18.20	0.31	0.07	0.12
TinyV3	Test	3.8	25	178	570	2,049	11.64	0.24	0.08	0.12

Table 2.8. Annotation distribution according to weather conditions

Country	Dataset	Drizzle	Frost	Rain	Cloudy	Fog	Clear sky
Denmark	Train	**20,886**	0	**37,064**	3,051	0	0
Denmark	Test	**16,291**	0	0	0	0	0
Germany	Train	0	10,206	13	0	1,535	0
Poland	Train	212	0	0	15,657	0	0
Poland	Test	0	0	1,153	6,441	0	752
Lithuania	Test	3,687	0	25	5,459	0	0

Table 2.9. YOLOv3 precision comparing pedestrian distance from camera

Pedestrian distance	Annotation in test set	Annotations in train set	mAP50 in test	mAP50 in train
81 m to inf.	7,050	6,725	**38.7**	1.8
61 m to 80 m	**39,339**	**15,790**	**82.9**	52.4
41 m to 60 m	**24,689**	**7,104**	**92.9**	76.5
21 m to 40 m	8,212	2,085	91.4	79.5
0 m to 20 m	9,334	2,104	91.9	69.7

Table 2.10. TinyV3 precision comparing pedestrian distance from camera

Pedestrian distance	Annotation in test set	Annotations in train set	mAP50 in test	mAP50 in train
81m to inf.	7,050	6,725	**38.6**	0.6
61m to 80m	**39,339**	**15,790**	**71.5**	22.8
41m to 60m	**24,689**	**7,104**	**75.6**	59.4
21m to 40m	8,212	2,085	70.9	51.7
0m to 20m	9,334	2,104	68.9	23.2

(two pedestrians visible, others not), it was marked as a single group of people annotation, including partially visible pedestrians. This part was taken with the care in the ZUT dataset. We marked clearly visible pedestrians as pedestrians, and only the occluded and partially visible were marked as an occluded annotation.

To solve this incompatibility in annotation methodology, we have iterated through all SCUT dataset and excluded frames containing a group of people annotations and people annotations similar to the square shape. Also, to make a competition fair, we reused our YOLO 8 bit configuration and trained it on the modified SCUT dataset. Important to mention, we have also scaled down all the images to 640×480 resolution, since the original source resolution was lower and merged with other classes to people class. The dataset shrunk to 78,942 frames (118,377 annotations) for the training and 76,381 frames (122,537 annotations) for the testing.

In the Table 2.11 it is presented the training results of YOLOv3 and TinyV3 detectors of the modified SCUT dataset. We have used two thresholds for IoU, which were set to 50% and 25%. The YOLOv3 version reached 86.4% mAP, which was very close to our 16 bit version and outperformed our 8 bit version. The TinyV3 reached up to 79.3% mAP, better than the 8 bit version but still not enough to compete with 16 bit modification.

In the Table 2.12 and Table 2.13 it is provided mAP, Average IoU (Avg.IoU), as well as True Positive (TP), False Positive (FP) and False Negative (FN) and Recall (Rcl.), Precision (Prc.), F1-score (F1) measures having two thresholds (25% & 50%) of IoU of detectors testing one dataset against another. The strategy of comparison was to take SCUT and compare it against ZUT training. Then benchmark sets and after that do oppositely for ZUT. Such comparison revealed that both sets are not performing verywell against each other since both of them were collected in different weather conditions and location and surrounding do not have much of a familiar context. The best results for SCUT were on the ZUT training set, having 37.7% mAP of 50% IoU by YOLOv3. The ZUT performed similarly on the SCUT dataset, reaching 39.1% mAP with YOLOv3, at the most.

Since both sets performed similarly, it was additionally decided to join them into one set and experiment on retrained YOLOv3 & TinyV3 detectors. For this reason the same validation performed once again and the Table 2.14 it is visible that the precision improved a lot. YOLOv3 on the ZUT training set reached 82.7% mAP, and on the ZUT benchmark 69.2% mAP and 80.8% mAP. TinyV3 also improved and achieved 73.9% mAP on the ZUT training set, 58.6% mAP on SCUT dataset.

Table 2.11. Results of YOLOv3 and TinyV3 trained on SCUT dataset

Resolution	Version	Depth, bit	Loss	mAP50	mAP25	Iteration
640×480	YOLOv3	8	0.1456	**86.4**	89.3	269,000
640×480	TinyV3	8	0.1786	**79.3**	83.3	168,000

Table 2.12. SCUT trained detectors testing on ZUT dataset

Detector	Validation	mAP	IoU, %	TP	FP	FN	Avg. IoU, %	Rcl.	Prc.	F1
YOLOv3	ZUT train	**37.7**	50	3,647	2,534	5,874	39.27	0.38	0.59	0.46
TINYv3	ZUT train	**32.4**	50	3,022	2,265	6,500	39.12	0.32	0.57	0.41
YOLOv3	ZUT train	55.0	25	4,503	1,678	5,019	45.02	0.57	0.73	0.57
TINYv3	ZUT train	45.8	25	3,633	1,654	5,889	43.87	0.38	0.69	0.49
YOLOv3	ZUT test	**27.7**	50	3,796	3,543	9,280	34.39	0.29	0.52	0.37
TINYv3	ZUT test	**25.4**	50	3,483	3,994	9,593	32.01	0.27	0.47	0.34
YOLOv3	ZUT test	37.8	25	4,495	2,844	8,581	38.31	0.34	0.61	0.44
TINYv3	ZUT test	33.9	25	4,092	3,385	8,984	35.38	0.31	0.55	0.40

Table 2.13. ZUT trained detectors testing on SCUT dataset

Detector	Validation	mAP	IoU, %	TP	FP	FN	Avg. IoU, %	Rcl.	Prc.	F1
YOLOv3	SCUT	**39.1**	50	38,059	7,008	84,478	56.08	0.31	0.84	0.45
TINYv3	SCUT	**32.6**	50	29,597	10,514	92,940	46.96	0.24	0.74	0.36
YOLOv3	SCUT	55.7	25	43,996	1,071	78,541	61.86	0.36	0.98	0.52
TINYv3	SCUT	57.0	25	38,011	2,100	84,526	55.97	0.31	0.95	0.47

Table 2.14. Detectors trained on combined dataset (ZUT and SCUT) testing on ZUT and SCUT datasets separately

Detector	Validation	mAP	IoU, %	TP	FP	FN	Avg. IoU, %	Rcl.	Prc.	F1
YOLOv3	ZUT train	**82.7**	50	6,530	1,009	2,992	62.29	0.69	0.87	0.77
TINYv3	ZUT train	**73.9**	50	6,882	1,806	2,642	56.67	0.72	0.79	0.76
YOLOv3	ZUT train	92.1	25	6,967	572	2,555	64.70	0.73	0.92	0.82
TINYv3	ZUT train	89.8	25	7,715	971	1,807	60.57	0.81	0.89	0.85
YOLOv3	ZUT test	**69.2**	50	6,705	1,059	6,371	60.24	0.51	0.86	0.64
TINYv3	ZUT test	**58.6**	50	6,880	2,008	6,196	54.02	0.53	0.77	0.63
YOLOv3	ZUT test	79.7	25	7,193	571	5,883	62.92	0.55	0.93	0.69
TINYv3	ZUT test	72.5	25	7,692	1,196	5,384	57.86	0.59	0.87	0.70
YOLOv3	SCUT	**80.8**	50	86,512	7,156	36,025	71.19	0.71	0.92	0.80
TINYv3	SCUT	**78.1**	50	85,661	12,426	36,876	66.04	0.70	0.87	0.78
YOLOv3	SCUT	83.9	25	87,991	5,677	34,546	71.87	0.72	0.94	0.81
TINYv3	SCUT	83.7	25	88,500	9,587	34,037	67.22	0.72	0.90	0.80

To sum up, in this section there was provided a ZUT dataset that contains 122,000 annotations and more than 79,000 (see Fig. 2.8) collected during the drizzle or the rain. The remaining annotations were collected during frosty and cloudy conditions. Only 752 annotations were observed when the sky was clear. In addition to this, the dataset includes car CAN data, which can be used for creating ADAS systems for thermal image based detectors.

Furthermore, the proposed modifications show that using 16 bit images instead of 8 bit improves detection accuracy 9.5–11.2% mAP. The adaptive threshold intensity method also gives improvements in increasing accuracy by four percent, however the more complex structure could improve the accuracy further because the current proposition is based on average results, and in this dataset, there was not enough samples in the temperature range between –1.5 to 4 ℃. Also, the onboard precipitation sensor would help in adjusting the intensity of the image, because currently we cannot apply any further real-time enhancements in regards to the rain or fog. Finally, the comparison of SCUT and ZUT databases showed that a wider variety of annotations made a much stronger detector, which is capable of work in severe and good weather conditions.

2.6. Conclusions of the Second Chapter

Several attempts to create the pedestrian detector using different approaches that would be able to work in real-time gave new important results. The following conclusions were formulated:

1. It is possible to speed-up pedestrian detection up to 12 times when a sliding window-based approach is changed to background subtraction.
2. The detection speed of the detectors, based on background subtraction depends on the number of the analyzed objects in the image and are not suitable for robust pedestrian detection in real-time.
3. Automatic threshold estimation techniques for background subtraction have only up to 80% of detection accuracy comparing to the empirically selected threshold values.
4. Incorrect annotation of the objects in two biggest FIR image datasets does not allow us to use these datasets for training and as a benchmark of deep-learning based pedestrian detectors.
5. The extended dynamic range of FIR images from 8 bit to 16 bit improves the accuracy of the convolutional neural network-based pedestrian detector to 9.5% mAP for YOLOv3 detector and 11.2% mAP for TinyV3 detector.
6. Temperature based adaptive threshold applied for histogram equalization reduces the intensity of hot objects and is able to increase pedestrian detector precision by 3.2% mAP for YOLOv3 detector and 4.3% mAP for TinyV3 detector in ZUT dataset.
7. The Environmental condition's impact on the ZUT dataset images reduce mAP by 58.7% for YOLOv3 detector and 53.9% for TinyV3 detector when detectors are trained on the SCUT dataset.

3

Improvement and Experimental Tests of the Pedestrian Detector

In this chapter several techniques to improve the performance of the pedestrian detector are introduced and experimentally tested. The experiments starts from merging a two datasets (ZUT and SCUT). Having a large dataset, six most popular detectors were trained and tested. Later, each detector was analysed to compare how well it was able to transfer the learning data. The dataset was split into two datasets were one contains data of low confidence data and another contains high confidence data. Using dataset with low confidence data, the detector was re-trained until the best accuracy level was received. Later, the selected detectors were tested in two low power single-board computers. According to the received results, the structure of detectors were optimised to obtain real-time performance by implementing the proposed methodology of training and slimming procedures. During the second experiment a SCUT dataset images were augmented in the way to reflect severe weather features. The same six detectors were trained and accuracy of detecting pedestrians was analysed.

The research results, presented in this chapter are published in one scientific paper (Tumas *et al.* 2021).

3.1. Fusion of the Datasets and Analysis of the Annotations' Distribution

Numerous research showed that a key to successful deep learning-based image object detection is a rich dataset with diverse labeled examples used for model training. By having a rich dataset it is possible to create an accurate detector working in various conditions. To improve performance of the pedestrian detection in our prototype, various ways to extend the training dataset were used. Two FIR datasets (SCUT and ZUT) were combined into single dataset for training. From ZUT it was taken an 8 bit + TI version frames, containing only "Pedestrian", "Occluded", "Cyclists", "Motorcyclist" and "Scooterist" classes, which were merged into single category. During the next stage, the preparation of the SCUT dataset was performed by three stages:

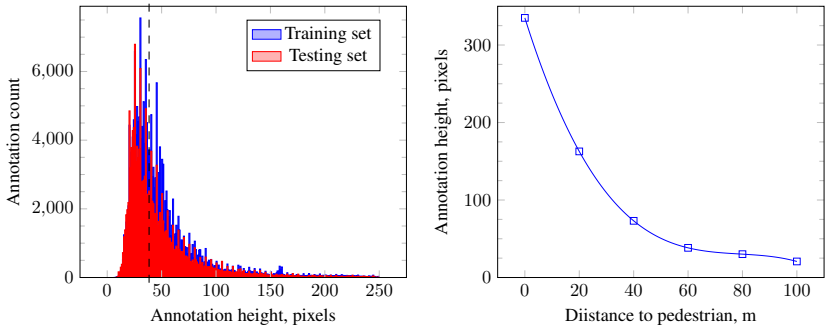
- an inspection of the annotations in the SCUT training dataset was performed by removing frames containing a group of people and people annotations similar to the square shape, because annotation methodology in ZUT and SCUT differs ZUT;
- all classes were merged to a single People class;
- all images were scaled down to 640×480 resolution, since SCUT has interpolated resolution of 720×576.

In the Table 3.1, the number of annotations and frames prepared for training and testing of the detector is showed. There were 69,455 frames extracted for training (88,624 annotations) and 40,103 frames extracted for testing (33,808 annotations) from ZUT dataset. The SCUT dataset provided a bit more examples: 78,942 frames (118,377 annotations) were used for training and 76,381 frames (122,537 annotations) were used for testing.

To show the spatial distribution of image annotations among these datasets, a heat map was generated (see Fig. 3.2), representing annotations' location and size in the image. From this heat map, it is visible, that most annotations are located on right side of the road (red and dark red color), because China and Europe (where these FIR images were collected) are left-side drive countries and pedestrians usually walks on the right side. Also, it is illustrated in Fig. 3.1, the spatial distribution of annotations remains similar in the train and test subsets.

Table 3.1. Dataset used for the training and testing

Dataset	Train frames	Train annotations	Test frames	Test annotations
ZUT	69,455	88,624	40,103	33,808
SCUT	78,942	118,377	76,381	122,537
Total	**148,397**	**207,001**	**116,484**	**156,345**



a) b)

Fig. 3.1. Data diversity in the combined dataset: a) test and train dataset; b) annotation height per pedestrian distance

The accuracy of pedestrian detector depends on the distance to the pedestrian (the number of pixels covering the pedestrian in the image). In order to understand the combined dataset annotation distribution according to the distance to pedestrian, a histogram was prepared (see Fig. 3.1a). It is visible in the histogram that the annotation height count distribution across training (blue bars) and testing (red bars) sets has only minor differences. Also the balancing boundary (dashed line) is shown in Fig. 3.1a, where both sides have equal number of annotations (at 41 px). From blue and red column it is also visible that both datasets (training and testing) are well balanced between each other, because the shape of training set and testing set are very similar and there are no major gaps which would influence imbalance of the dataset and mAP calculation. From the Fig. 3.1b it is visible that when pedestrian is within 60–100 m range to the camera, the height of the bounding box (annotation) changes almost in linear manner. However, in the range from 40 to 60 m, the height of annotation starts to increase rapidly.

3.2. Selection of the Improved Detectors for the Prototype

Six well-known convolutional neural network architectures were selected to be investigated for the most accurate and real-time ready pedestrian detector: a conventional TinyV3 (Redmon, Farhadi 2018c); a TinyV3 with additional head, named TinyL3 (Bochkovskiy 2019c); YOLOv3; YOLOv4; ResNet50 (He *et al.* 2016b; Wong 2020) and Cross Stage Partial Network, named CSPNet (Wang *et al.* 2020), applied on ResNeXt50 (Bochkovskiy 2019b; Xie *et al.* 2017a). Also, the following modifications to the neural network configuration were made:

- the input layer was modified to accept monochromatic inputs of 640×480 resolution;
- annotations were auto-rotated by 5 degrees;
- contrast and brightness left unchanged;
- filter sizes were recalculated for single class use;
- anchor ratios for the datset were recalculated using the *k*-means algorithm;
- the learning rate was set to 0.001.

To select the best one, each detector was trained in 300,000 steps by saving training weights after each 1,000 steps. After the training, mAP at IoU=50 were measured for each saved step and the result showing the best precision was used for further evaluation.

To evaluate the performance of the detector, it was decided to measure the most common metrics used for image object detectors evaluation: FPS, mAP, Average IoU (Avg. IoU), Recall (Rcl.), Precision (Prc.), F1-score (F1). Also, there were calculated True Positive (TP), False Positive (FP) and False Negative (FN) rates, using 50% of the IoU threshold. The FPS measurements were performed on Intel i7-8750H eighth-generation processor and nVidia RTX2070Ti 8GB graphics card. Training and validation were performed using Darknet DNN framework (Bochkovskiy 2019a).

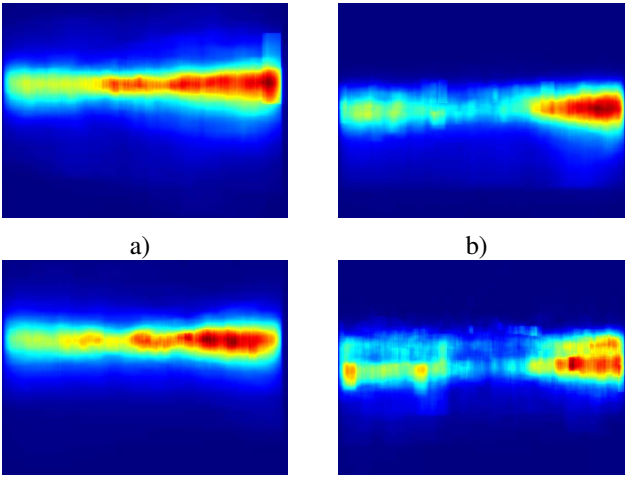


Fig. 3.2. Spatial distribution of annotations in: a) SCUT train dataset; b) ZUT train dataset; c) SCUT test dataset; d) ZUT test dataset

3.2.1. Results of Detector Training on a Fused Dataset

The initial results of pedestrian detector training on a fused dataset are presented in Table 3.2. The most accurate detector, tested on the fused dataset, was YOLOv4. It reached 86.05% mAP and achieved 15.97 FPS on average. The second highest precision was received by ResNet50, with reached 81.00% mAP and achieved 19.82 FPS on average. Detector structures with a minimized Backbone, such as TinyL3, outperformed YOLOv3 (the predecessor of YOLOv4) by 0.34% mAP and reached 43.1 FPS on average. The fastest detector was based on TinyV3 architecture. It worked at 55.57 FPS on average. However, it was the least accurate and showed 73.25% mAP. The ResNext50 have not outperformed tested detectors in any metrics.

In order to better understand where each pedestrian detector has weak detection points, it was decided to evaluate each detectors confidence (probability) distribution in six different regions based on Fig. 2.17:

- from 100 m to infinity (height is from 21 px to 0 px);
- from 80 m to 100 m (height is from 29 px to 22 px);
- from 60 m to 80 m (height is from 40 px to 20 px);
- from 40 m to 60 m (height is from 72 px to 41 px);
- from 20 m to 40 m (height is from 136 px to 73 px);
- from 0 m to 20 m (height is from 480 px to 137 px).

The training set was used for the evaluation here, because it is important to verify if all annotations were learned from the training dataset by the detector. The detection evaluation results of TinyV3 detector are presented in Fig. 3.3. The graph shows, that the detector was capable to learn the majority of annotations, since the highest annotation peaks are concentrated next to 90% probability. However, ranges 100 m to infinity, 80–100 m, 60–80 m and 0–20 m reveal that the detector was not able to learn many other annotations, for which the output of the detector gave a probability lower than 90%.

Table 3.2. Results of Detector Training on a Fused Dataset

Detector	FPS	mAP	TP	FP	FN	Avg. IoU, %	Rcl.	Prc.	F1
TinyV3	**55.57**	73.25	86,009	11,109	49,604	66.57	0.63	0.89	0.74
TinyL3	43.10	80.14	96,265	18,311	39,348	62.52	0.71	0.84	0.77
YOLOv3	17.88	80.48	95,201	8,978	40,412	68.95	0.70	0.91	0.78
YOLOv4	15.97	**86.05**	113,778	32,318	21,835	59.28	0.84	0.78	0.81
ResNet50	**19.82**	**81.00**	101,059	21,578	34,554	62.34	0.75	0.82	0.78
ResNext50	17.70	77.07	87,850	13,755	47,763	65.57	0.65	0.86	0.74

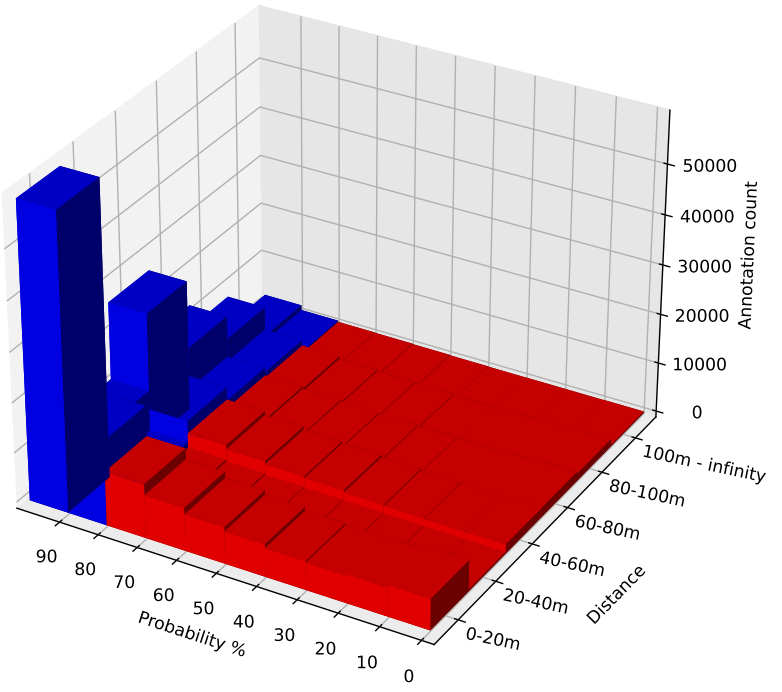


Fig. 3.3. TinyV3 detector confidence distribution based on pedestrian distance from the camera

Despite the fact that TinyL3 detector outperformed TinyV3, the Fig. 3.4 shows that the detector was able to learn the majority of annotations in ranges 100 m to infinity; 60–80 m and 20–40 m, but in the remaining distance ranges the detector was not confident enough. For example in 80–100 m range it is visible that the majority of annotation is concentrated next to 20–50% confidence range. The 40–60 m range shows that detector still was not capable to learn all annotation, but there are more annotations with higher probability next to the right side of the graph (closer – 90%). Lastly, in the range from 0–20 m it is visible that the probability is almost evenly distributed for all annotations, which means that detector would miss detect pedestrians.

The pedestrian detection probability estimates for YOLOv3, visible in the Fig. 3.5, showed very similar results to TinyV3, where the majority of annotations has the probability close to 100%, but in rages of 80–100 m and from 100 m to infinity a higher number of annotations with lower probability were received.

Even though the YOLOv4 was the leader in detection, the Fig. 3.6 shows that

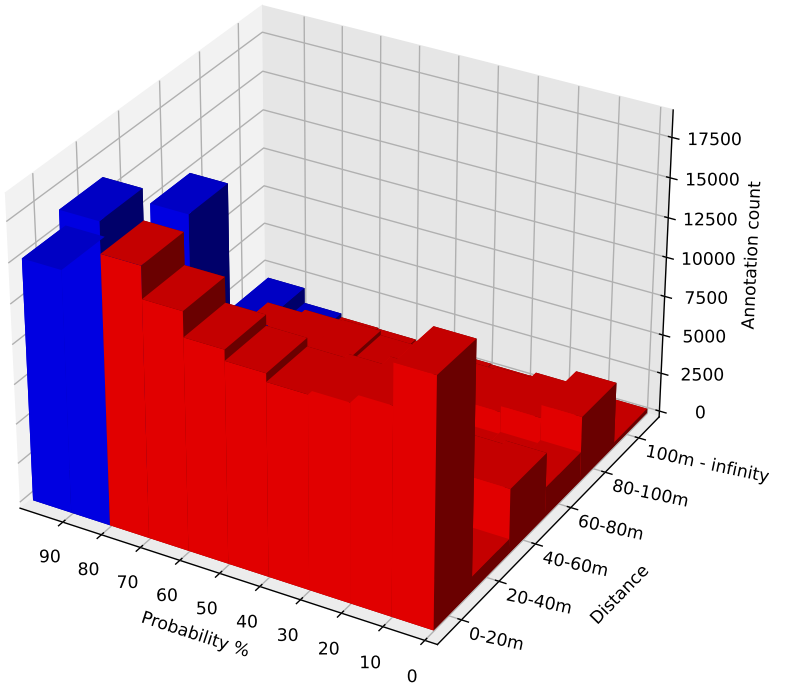


Fig. 3.4. TinyL3 detector confidence distribution based on pedestrian distance from the camera

the detector has close to excellent detection mostly at near distances (0–80 m). However, at far distance with small pedestrians the YOLOv4 has a challenge to detect pedestrians confidently, that can be observed in ranges 80–100 m and 100 m to infinity.

The ResNet50 showed similar results (see Fig. 3.7) to TinyV3 detector, where the outputs of the detector, related to the majority of annotations had high probability (close to 100%). However, the detector output probabilities were slightly higher for distance to pedestrian rages of 60–80 m, 80–100 m and from 100 m to infinity.

Finally, the results of ResNext50 detector, represented in Fig. 3.8, showed close to excellent performance in the ranges of 0–20 m, 20–40 m and 40–60 m. In the range from 60–80 m, the detector showed very similar results to YOLOv4, but in the ranges 80–100 m and from 100 m to infinity, detector struggles to extract small pedestrian features and learn to have higher confidence.

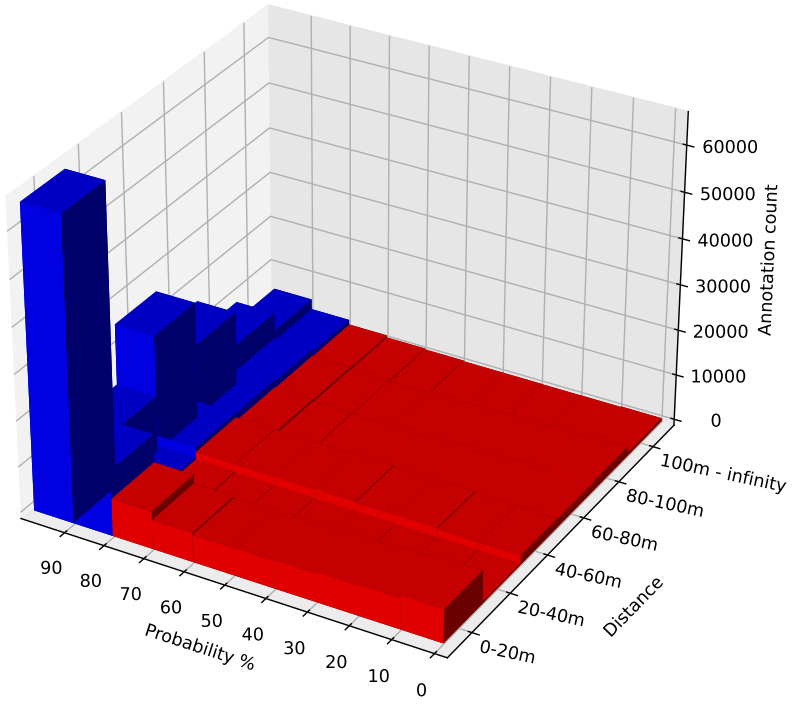


Fig. 3.5. YOLOv3 detector confidence distribution based on pedestrian distance from the camera

3.2.2. Retraining of the Detector According to Confidence Distribution

Once probability distribution histograms were made, the next step was to adjust the training dataset by selecting regions where detector is not accurate enough. For TinyV3 detector, all six example subsets (see Fig. 3.3 marked columns in red) were selected that had a probability lower than 80%. This decision was made because the TinyV3 has a consistent and highest confidence distribution from 80% to 100% in all ranges.

Next, for TinyL3 detector (see Fig. 3.4 marked columns in red), five regions (from 0–20 m, from 20–40 m, from 40–60 m, from 60–80 m and from 100 m to infinity) were selected with criteria where detector confidence is less than 80% and a full annotation set from region of 80–100 m since here detector makes most of inaccurate detections.

For YOLOv3, the same methodology (see Fig. 3.5 marked columns in red) was used like in TinyV3 case – where all six regions were selected that had a detection probability lower than 80%.

For YOLOv4 (see Fig. 3.6 marked columns in red), four regions (0–20 m, 20–40 m, 40–60 m and 60–80 m) were selected with probability less than 90%. In addition, from 80 m to 100 m all annotations were selected which gave detector probability bellow 80% and all annotations were selected from the region 100 m to infinity.

For ResNet50 (see Fig. 3.7 marked columns in red), in the same manner like for TinyV3 detector, all six regions were selected that had a probability, lower than 80%. Lastly, for ResNext50 (see Fig. 3.8 marked columns in red) two regions (from 40–60 m and from 20–40 m) were selected with probability lower than 90%. Then, two regions (from 60–80 m and 0–20 m) were selected with probability lower than 80%. For remaining regions (from 80–100 m and 100 m to infinity) all annotations were selected.

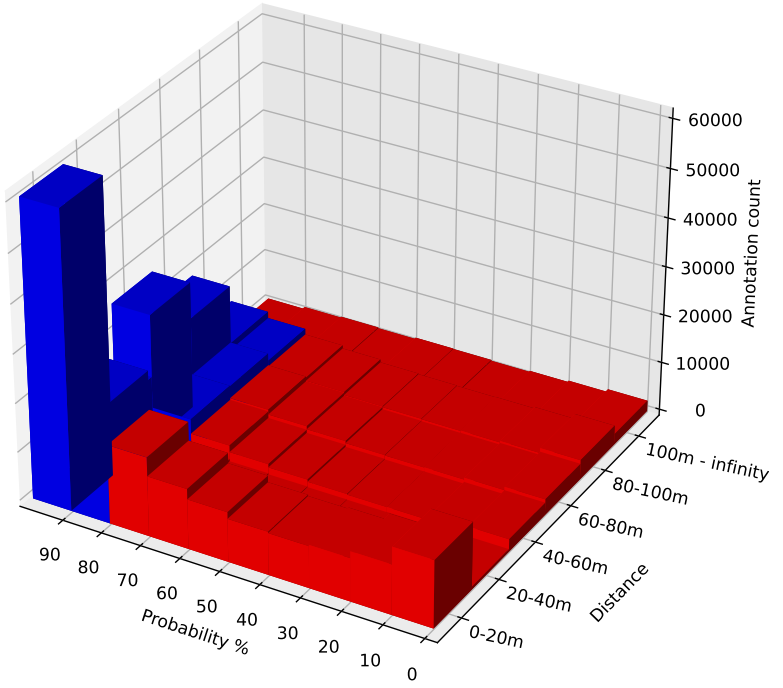


Fig. 3.6. YOLOv4 detector confidence distribution based on pedestrian distance from the camera

Then all detectors were trained on their dedicated datasets and the performance evaluation results are represented in Table 3.3. The best performance was observed by YOLOv4 detector which reached the accuracy of 86.69% mAP and gained 0.64% mAP improvement. The next most accurate detector was YOLOv3 which reached 84.41% mAP and which showed 3.93% mAP improvement. The ResNext50 detector was the third most accurate detector which was able to reach 83.31% mAP accuracy and gained the highest accuracy boost with the confidence training reaching 6.24% mAP. The ResNet50 was the fourth most accurate detector (82.84% mAP and gained 1.84% mAP). TinyL3 was fifth most accurate detector which unfortunately was not able to gain any improvements in this experiment. The TinyV3 was the least accurate detector which reached 78.77% mAP accuracy and showed quite significant improvement which is 5.52% mAP. In addition to training results it is also important to mention that selection for the data using confidence distribution was applied only single time as well as all detectors were retrained once.

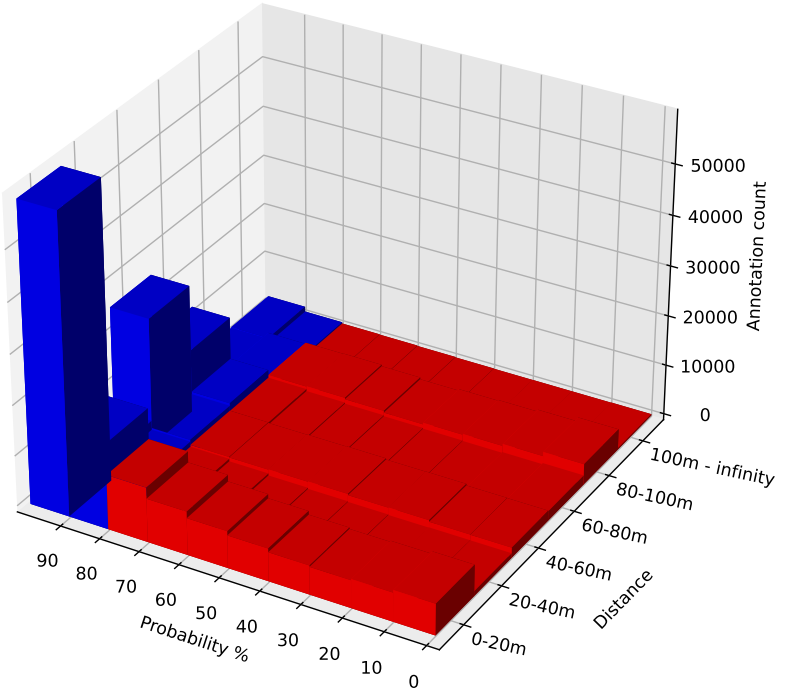


Fig. 3.7. ResNet50 detector confidence distribution based on pedestrian distance from the camera

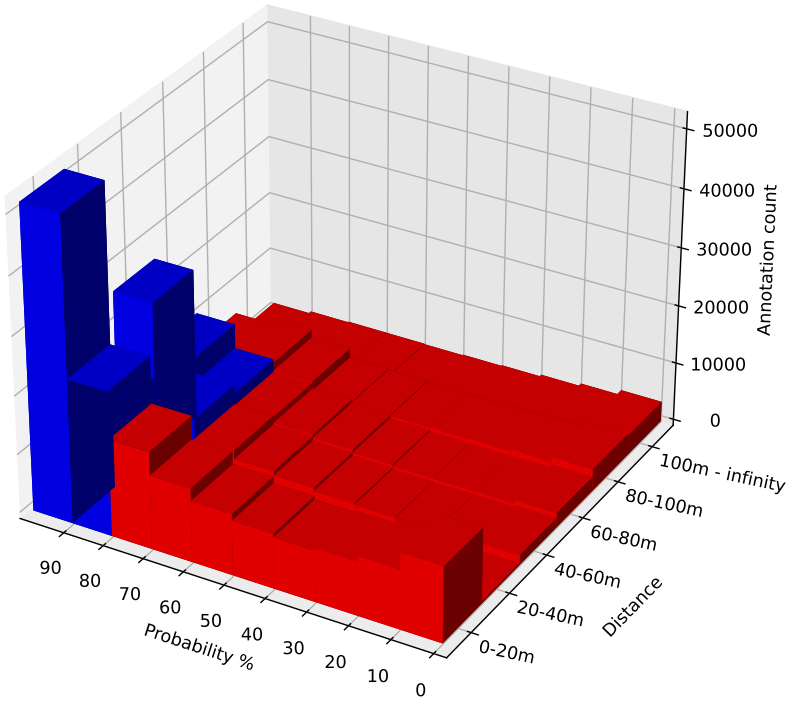


Fig. 3.8. ResNext50 detector confidence distribution based on pedestrian distance from the camera

Table 3.3. Retraining results by using confidence distribution

Detector	mAP	TP	FP	FN	Avg. IoU, %	Rcl.	Prc.	F1
TinyV3	78.77	102,303	24,893	33,310	59.92	0.75	0.80	0.78
TinyL3	80.13	96,140	17,915	39,473	62.83	0.71	0.84	0.77
YOLOv3	84.41	104,451	8,978	40,412	68.95	0.70	0.91	0.78
YOLOv4	**86.69**	118,996	60,752	16,617	49.56	0.88	0.66	0.71
ResNet50	82.84	105,311	23,340	30,302	61.55	0.78	0.82	0.80
ResNext50	83.31	112,030	51,038	23,583	50.57	0.83	0.69	0.75

3.2.3. Testing Pedestrian Detectors on the Single-Board Computers

Since all detectors can run in real-time on a powerful GPU and are pretty accurate so far, a low-power solution should be investigated next for practical applica-

tion. For such applications, two single-board computers were taken into consideration in this section. Currently, the top-notch deep neural network devices are made by Nvidia, and two single board computers were tested: Jetson TX2 (TX2) and Jetson AGX Xavier (AGX) (a detailed comparison of device specifications is presented in the Table 3.4). Both computers are low-power devices dedicated to edge computing, end-to-end AI robotics applications for manufacturing, delivery, retail, agriculture, and are compatible with the CUDA, cuDNN, and TensorRT frameworks, which means that already tested detectors can be implemented using straight-forward approach.

For hardware preparation an Ubuntu 18.10 operating system was installed, with CUDA 10.2 framework, OpenCV with CUDA support enabled and Darknet framework compiled to utilise OpenCV and CUDA capabilities. Also, for TX2 and AGX devices the maximum performance profiles were activated: TX2 was utilising all 6 cores and was consuming 15 W; the AGX was utilising all 8 cores and was consuming 30 W of power. Before starting the initial test, the number of computation operations (needed for running the detector) were measured for each detector individu-

Table 3.4. Comparison of TX2 and AGX hardware specifications

Parameter	Jetson TX2	Jetson AGX Xavier
CPU	4-core ARM Cortex-A57 @ 2 GHz, 2-core Denver2 @ 2 GHz	8-core ARM Carmel v.8.2 @ 2.26 GHz
GPU	256-core Pascal @ 1.3 GHz	512-core Volta @ 1.37 GHz
Memory	8 GB 128-bit LPDDR4 58.3 GB/s	16 GB 256 bit LPDDR4 137 GB/s
Storage	32 GB eMMC 5.1	32 GB eMMC 5.1
Tensor cores	0	64
Video encoding	1x 4K60 (H.265) 3x 4K30 (H.265) 4x 1080p60 (H.265)	4x 4K60 (H.265) 16x 1080p60 (H.265) 32x 1080p30 (H.265)
Video decoding	2x 4K60 (H.265) 7x 1080p60 (H.265) 14x 1080p30 (H.265)	2x 8K30 (H.265) 6x 4K60 (H.265) 26x 1080p60 (H.265) 72x 1080p30 (H.265)
USB	(1x) USB 3.0 (1x) USB 2.0	(3x) USB 3.1 (4x) USB 2.0
PCI Express lanes	5 lanes PCIe Gen 2	16 lanes PCIe Gen 4
Power	7.5W / 15W	10W / 15W / 30W

Table 3.5. FPS measures running detectors on TX2 and AGX computers

Detector	BFLOPS	TX2	AGX
TinyV3	9.67	**10.1**	**42.8**
TinyL3	12.60	8.9	35.0
YOLOv3	115.93	2.2	5.7
YOLOv4	105.73	2.1	5.3
ResNet50	86.52	3.1	6.5
ResNext50	82.59	2.3	5.1

ally in billion floating point operations per second (BFLOPS). The results, showing how each low-power computer performed and how many operations were needed to run a detector, are presented in Table 3.5. The best performance (10.1 FPS) for TX2 was received with TinyV3 detector. The next most accurate detector was TinyL3, which was capable to run at 8.9 FPS. The most accurate YOLOv4 detector on TX2 board was running only at 2.1 FPS. The AGX computer showed much better results. There the most fastest was TinyV3 detector, running at 42.8 FPS and TinyL3 detector, reaching 35.0 FPS. The most accurate detector YOLOv4 reached only 5.3 FPS but still outperformed version running on TX2 by two times.

3.2.4. Optimization of the Pedestrian Detector

Since only AGX showed real-time performance running TinyV3 and TinyL3 detector, there was a need to search for the possibilities to optimize the structure of the detectors, to gain more inference speed. In case of success, a real-time performance utilising TX2 computer could be achieved.

During the first attempts to optimize the structure of DNN for inference speeds, a pruning (also known as slimming) introduced by (Liu *et al.* 2017b) was applied. A network slimming takes a wide and large networks as input models, but during training, the insignificant channels are automatically identified and pruned afterwards, yielding thin and compact models with comparable accuracy. For this reason, the pruning methodology was utilised to find inactive channels, applying additionally the previously introduced DNN retraining methodology (retraining based on confidence distribution).

An approach of the DNN slimming procedure is presented in Fig. 3.9. First, one of the detectors was selected for training and slimming. During the training, an analysis was performed to find inactive or close to zero channels of the network. Then, identified channels were removed and a new detector configuration (and detector weights) were generated. After this procedure, the low and high confidence training datasets were generated. Only a low confidence dataset was used to train the slimmed detector. During the training, mAP was measured every 1000 steps and the training was continued only if mAP started to increase. If the mAP did

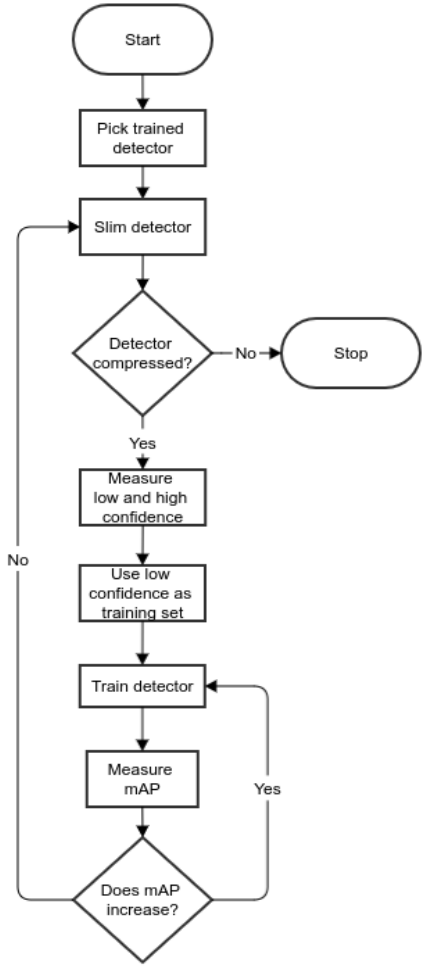


Fig. 3.9. A DNN slimming algorithm

not increased after 1000 steps, the training process was terminated and a slimming procedure started. The procedure was repeated continuously until there were no new inactive or close to zero channels found.

In the Table 3.6 there are presented results of the first DNN slimming iteration, initially tested on NVIDIA RTX2070Ti (RTX) graphics card, then on TX2 and AGX boards. As it is shown in the table, TinyV3 detector was able to perform at 11.8 FPS on TX2, reaching 58.1 FPS on AGX and reaching 223.3 FPS on RTX (7 times fasted than required for real-time processing of video stream).

Looking from perspective of DNN structure optimization, the size of the network decreased 2.18 times, while the precision dropped only by 0.49% mAP. TinyL3 detector was able to slim by 2.07 times and but it gained precision by 1.61% mAP and was able to perform 10.2 FPS on TX2, 53.3 FPS on AGX and 169.8 FPS on RTX. YOLOv3 reflected similarly where it was slimmed by 1.97 times, but lost precision by 0.95% mAP and was able to perform 32.2 FPS, however no real-time performance was observed on single board computers. YOLOv4 also showed similar results as other detectors, where it was slimmed 2.40 times and it also lost precision by 2.96% mAP. Looking from performance point of view, YOLOv4 was able to reach 54.4 FPS on RTX. ResNet50 slimmed the most by the first iteration 4.47 times, however, it lost only 0.35% mAP and was able to perform 7.7 FPS on TX2, 16.2 FPS on AGX and 82.2 FPS on RTX, which is 2.5 than real-time. Finally, ResNext50 slimmed by 2.79 imes and improved it's accuracy by 1.55% mAP. It also showed the best accuracy from all the other detectors, outperforming the YOLOv4. Performance wise TX2 reached 4.5 FPS, AGX 10.4 FPS and 57.6 FPS on RTX.

Results of the second slimming itteration are presented in Table 3.7. TinyV3 was able to shrink additionally by 1.81 to 2.45 BFLOPS operations maintaining 78.10% mAP, which differs from the original configuration only by 0.67% mAP. The detector speed running on AGX reached 87.9 FPS, 12.3 FPS when running on TX2 and 229 FPS on RTX. In the similar manner, the slimming of TinyL3 structure decreased the computational load additionally 1.58 times from the first iteration result. However, the accuracy of TinyL3 after second slimming iteration decreased by 1.47% mAP, but the accuracy still remained 0.14% mAP higher, comparing to the original configuration. The speed of TinyL3 increased the most on RTX, reaching 251.6 FPS. YOLOv3 was able to shrink 1.72 times more and showed its best accuracy by comparing original result (0.92% mAP increase). YOLOv4 was also to shrink by 2.37 times. However, it had a critical impact to accuracy which dropped by 15.78% mAP from initial setup. ResNet50 also decreased in size by 1.80 times, but it also affected accuracy by 3.91% mAP. The ResNext50 showed best results, after second slimming iteration by 1.73, and gained accuracy

Table 3.6. Detector evaluation results after the first slimming iteration

Detector	TX2	AGX	RTX	mAP	BFLOP	TP	FP	FN
TinyV3	11.8	58.1	**223.2**	78.28	4.44	99,055	21,510	36,558
TinyL3	10.2	53.5	**169.8**	**81.74**	6.08	100,096	19,864	35,517
YOLOv3	3.0	8.3	32.2	**83.46**	58.93	101,432	12,813	34,181
YOLOv4	3.2	9.2	54.4	83.73	43.99	108,110	26,814	27,503
ResNet50	7.7	**16.2**	82.2	82.34	19.36	93,339	8,920	42,274
ResNext50	4.5	10.4	**57.6**	**84.64**	29.60	108,972	28,976	26,641

Table 3.7. Detector evaluation results after the second slimming iteration measuring FPS for different computers

Detector	TX2	AGX	RTX	mAP	BFLOP	TP	FP	FN
TinyV3	**12.3**	**87.9**	229.0	78.10	2.45	98,440	21,795	37,173
TinyL3	10.8	49.5	**251.6**	80.27	3.85	97,922	22,298	37,691
YOLOv3	4.3	13.0	35.8	**85.33**	34.37	104,257	18,185	31,356
YOLOv4	5.8	14.3	77.2	70.27	18.60	93,977	48,850	41,636
ResNet50	11.5	21.8	112.8	78.58	10.72	91,610	14,232	44,003
ResNext50	6.3	17.0	**78.6**	**85.45**	17.09	115,213	49,614	20,400

by 0.81% mAP. However, to compare with initial results, the ResNext50 was only by 0.6% mAP less precise, but outperformed YOLOv4 almost 4 times with very close precision.

The attempt to continue slimming by running the third iteration was not successful. The training pipeline became very unstable, the loss was jumping and there were observed infinities during training. An attempt to decrease learning rate did not solve the problem and no positive result was received.

To conclude the results of the experimental investigation, it is reasonable to note, that the dataset fusion requires modification of neural network training pipeline, because the detector tends to learn the most general samples. By utilising the confidence distribution concept, it possible to increase the detector accuracy up to 6.24% mAP. Next, by utilising the slimming proceedure, first introduced by (Liu *et al.* 2017b) and a confidence distribution based training introduced in this dissertation, it is possible to decrease the number of floating point operations by 3.95 times. It helps to increase the inference time by 7 times without loosing precision of pedestrian detection (TinyL3 case with iteration 1). In some cases we may expect even the increase of the accuracy, e.g., by 8.38% mAP as in ResNext50 case compared initial training with slimming iteration 2.

3.3. Development of the Dataset Augmentation Algorithm

A key to successful machine learning-based image object detection is a rich dataset with diverse labeled examples used for model training. However, it is not easy to collect a dataset with many examples covering various situations captured by the FIR image sensor. The first reason – severe weather conditions prevent data collection due to rain and dirt which covers the sensor. The second reason – an image from the sensor is captured after pre-processing, where the charge collected in the FIR sensor matrix is equalized over the range to form an image. This preprocessing causes different images achieved at various environmental tempera-

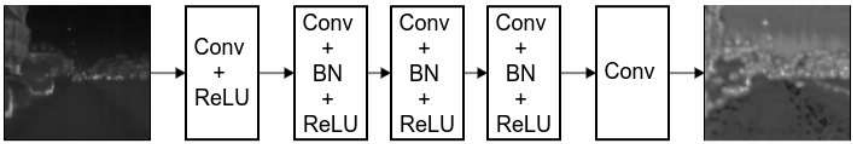


Fig. 3.10. A structure diagram of the DnCNN inspired DNN architecture for severe weather feature augmentation

tures, and the same objects might look differently on a hot summer day and on cold winter day. Also, the annotation process is time-consuming and requires manual and repetitive work, which usually introduces errors. Finally, data diversification does not cover all situations. For these reasons, application-specific data augmentation could be an advantageous technique to generate additional unique samples. Data augmentation makes the dataset full of different samples, more balanced, transferring annotation and filling the dataset gaps.

A typical data augmentation is a collection of methods used to automatically generate new data samples via the combination of existing examples and prior domain knowledge. Each particular augmentation method is usually designed to support model performance invariance (i.e., unchanging) on corresponding cases of possible inputs. Usually, data augmentation consists of following operations:

- flipping horizontally / vertically;
- rotation (5–15 degrees);
- scaling (15–20%) or cropping;
- distortions (geometrical);
- intensity jittering;
- edge-enhancement;
- addition of Gaussian white noise.

3.3.1. Dataset Augmentation Preparation

Since the manual collection of additional images in severe weather conditions is complicated and time-consuming process, in this section synthesis option for FIR images is investigated. By investigating the spatial distribution of annotations on the left side is less intensive on a heat map visible in Fig. 3.2, we flipped images and their annotations horizontally. After mirroring the dataset images, the next step was to enrich a training dataset with severe weather samples.

A typical way to generate a wider variety of samples would be by using a Generative adversarial network (GAN). However, we have investigated an alterna-

tive approach. Since we are aiming to generate a severe weather features into an image that is visually close to noise and contains linear predefined features, we tried to invert the functionality of denoising DNN. DnCNN (Xu *et al.* 2015) is well known DNN-based denoiser, providing excellent Gaussian denoising, super-resolution transformation capabilities. For this reason we took DnCNN architecture, given in Fig. 3.10, and performed a modification of training input function $y = x + v$ to $y = x$, where x is an input image and v is a random Gaussian noise.

To train a neural network dedicated to adding severe weather-related FIR image distortions, we took naturally distorted FIR images from the ZUT dataset. These natural images contain sequences of heavy rain, drizzle, and fog. We trained the slightly modified DnCNN until the loss function stopped converging. We applied DnCNN to generate distorted samples from images in the combined dataset, including the flipped ones.

The generated samples are visually similar to original situations where rain and dirt distort captured image. In Fig. 3.11 we have provided samples used for the training: input images (Fig. 3.11c and Fig. 3.11d), real images captures in heavy rain with wet and dirty sensor (Fig. 3.11a and Fig. 3.11b) and the output, modified by the DnCNN (Fig. 3.11e and Fig. 3.11f). From the sample, shown in Fig. 3.11c, it visible that the pedestrian is cold and there is not much thermal radiation visible, but the corresponding sample in Fig. 3.11e, generated by DnCNN, shows that features of a cold pedestrian are kept. Similar behavior is notices with warm pedestrians: Fig. 3.11d shows input image and Fig. 3.11f shows the generated output.

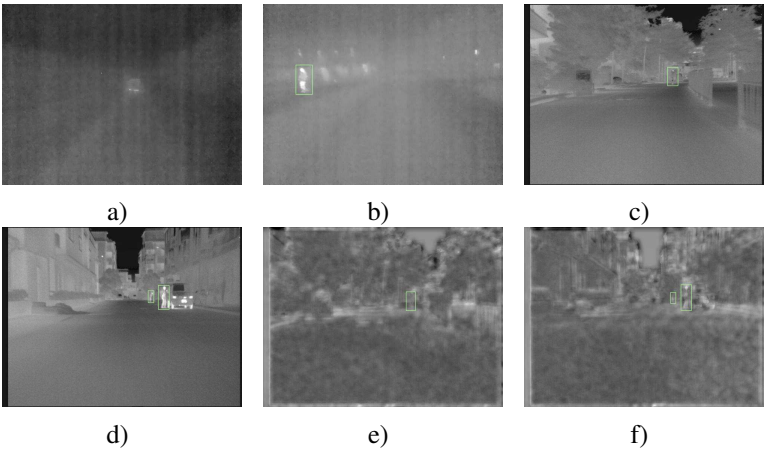


Fig. 3.11. Comparison of captured and samples, augmented with DnCNN where: a) Image during heavy rain; b) Wet and dirty sensor; c) FIR image with a cold pedestrian; d) FIR image with a warm pedestrian; e) Sample (c) modified with DnCNN; f) Sample (d) modified with DnCNN

Table 3.8. Results of the detectors evaluation after training on augmented dataset

Detector	mAP	TP	FP	FN	Avg. IoU	Rcl.	Prc.	F1
TinyV3	78.22	98,158	19,810	37,455	61.98	0.72	0.83	0.77
TinyL3	81.98	98,603	14,445	37,010	65.84	0.73	0.87	0.79
YOLOv3	83.87	101,814	13,676	33,799	67.50	0.78	0.88	0.81
YOLOv4	**87.02**	114,212	32,051	21,401	58.95	0.84	0.78	0.81
ResNet50	82.72	103,827	20,451	31,786	63.89	0.77	0.84	0.80
ResNext50	**86.45**	112,292	31,797	23,321	59.37	0.83	0.78	0.80

3.3.2. Selection and Modification of Deep Learning Structure for Data Augmentation

For data augmentation six DNN architectures were selected (initially pre-trained during the first experiment) to investigate the most accurate pedestrian detector: TinyV3, a TinyV3, YOLOv3, YOLOv4, ResNet50 and ResNeXt50. Also, there were the following modifications made to the neural network configuration training pipeline:

- the input was set to 640×480;
- annotations were auto-rotated by 5 degrees;
- contrast and brightness left unchanged;
- filter sizes were recalculated for single class use;
- anchor ratios were recalculated using the k-means algorithm;
- learning rate set to 0.001.

For selecting the best detector, we trained each detector until 300,000 steps by saving training weights every 1,000 step. After the training, the mAP was measured at IoU = 50 for each saved step and the best result was used for further evaluation. The same methodology was used after augmentation, but the training steps were increased to 500,000 steps. For evaluation of the detector performance we decided to measure the following metrics: Frames Per Second (FPS), mAP, Average IoU, Recall, Precision, and F1-score. Also, there was calculated True Positive (TP), False Positive (FP), and False Negative (FN) rates, using 50% of the IoU threshold. The FPS measurements were performed on Intel i7-8750H eighth-generation processor and NVIDIA RTX2070Ti 8 GB graphics card. Training and validation were performed on Darknet framework.

3.3.3. Review of Dataset Augmentation Results

The performance of the detectors, trained on an augmented dataset are presented in Table 3.8. The best performance was observed by YOLOv4, reaching an accuracy

of 87.02% mAP. The next most accurate detector is ResNext50, which reached 86.45% mAP and gained the highest accuracy boost with the augmented dataset (9.38% mAP). YOLOv3 is the third most accurate detector (83.87% mAP), which outperformed ResNet50 by 1.15% mAP and TinyL3 by 1.89% mAP. The ResNet50 is the fourth most accurate detector (82.72% mAP), which outperformed TinyL3 by 0.74% mAP. The TinyL3 detector showed 81.98% mAP and is more accurate than TinyV3 by 3.73% mAP.

A spatial detector confidence distribution of FP and FN visible in Fig. 3.12 revealed locations where the detector is failing to identify pedestrians or rather makes a false positive detection. This information could be a powerful way to dynamically adjust the detector's confidence threshold based on the image object location. For this reason, the experiment was made to slightly modify the detector's head to adjust the confidence threshold based on combined FP and FN heatmaps.

In Fig. 3.13 it is represented confidence heat-map where a decision to accept or reject detection is made by taking a detected pedestrian rectangle center coordinates and compare probability in this heat-map. If detected object probability is lower than a value of the heat-map – the detector ignores detection, but if the

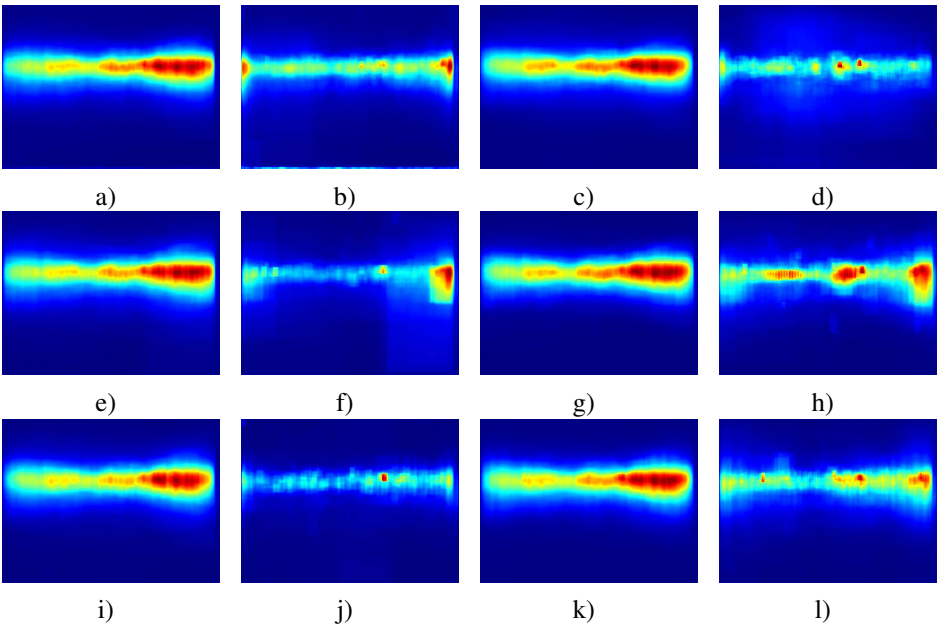


Fig. 3.12. Spatial detectors' confidence distribution of FP and FN: a) YOLOv4 FN; b) YOLOv4 FP; c) YOLOv3 FN; d) YOLOv3 FP; e) ResNet50 FN; f) ResNet50 FP; g) ResNext50 FN; h) ResNext50 FP; i) TinyV3 FN; j) TinyV3 FP; k) TinyL3 FN; l) TinyL3 FP

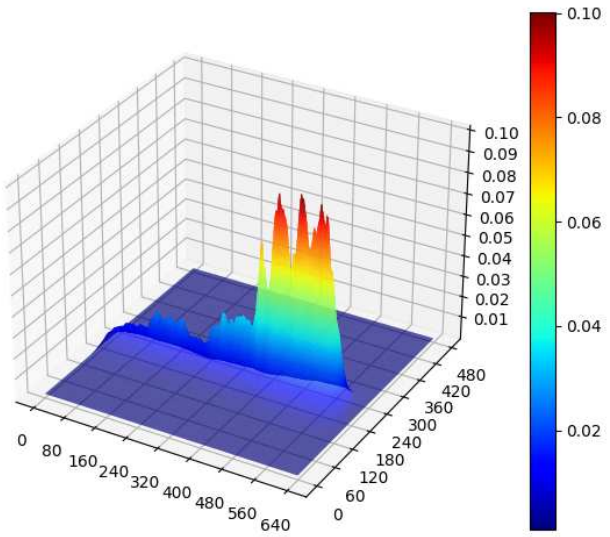


Fig. 3.13. An illustration of a confidence heat-map

probability in the output of the detector is higher – we accept detection. As it is shown in the Table 3.9, the detector, trained on the augmented dataset additionally gained 0.18–1.02% mAP (without any performance loss) where the most significant boost was observed by using YOLOv3 detector and the least observed by YOLOv4. Also, such confidence heat-map could be utilised with CAN information where e.g., car speed is used as threshold to control detectors sensitivity based on the location (highway or city) where the car is driven and detected pedestrian location on the captured image.

To summarise the experiment in this section, the DnCNN image augmentation application proved its value in pedestrian detection using state-of-the-art detectors, evaluated under a wider variety of situations. More complex samples of anno-

Table 3.9. mAP evaluation using heatmap head

Detector	**mAP**
TinyV3	78.74
TinyL3	82.73
YOLOv3	**84.91**
YOLOv4	87.20
ResNet50	83.44
ResNext50	86.67

tations resulted in a more robust detector capable of working in a broader range of weather conditions and situations, resulting in readiness for real-time pedestrian detection application. From the results presented, it can be concluded that data synthesis could contribute to other distortions generation, reflecting different weather conditions.

3.4. Conclusions of the Third Chapter

1. Re-training of the deep learning based pedestrian detector using examples selected according to the confidence distribution may increase the precision by 6.24% mAP to ResNext50 detector.
2. By utilising slimming introduced by (Liu *et al.* 2017b) and confidence distribution training, it is possible to decrease 4.83 times needed floating point operations, gain 11.9 FPS running detector on AGX as well as increasing accuracy by 8.38% mAP (ResNext50 case compared initial training with slimming iteration 2).
3. The proposed image augmentation approach based on DcDNN is able to gain the precision of 9.38% mAP to ResNext50 detector.
4. The proposed confidence heat-map is able to gain the precision of pedestrian detection by adding from 0.18% mAP to 1.02% mAP (observed by YOLOv3 detector) on unknown images.

General Conclusions

1. Classical pedestrian detectors based on application of Histogram Oriented Gradients for feature extraction could be accelerated by changing the sliding-window based analysis stage:
   1.1. It is possible to speed-up pedestrian detection up to 12 times when a sliding window-based approach is changed to background subtraction.
   1.2. The detection speed of the detectors, based on background subtraction depends on the number of the analyzed objects in the image and are not suitable for robust pedestrian detection in real-time.
   1.3. Automatic threshold estimation techniques for background subtraction reaches only up to 80% of detection speed comparing to the empirically selected threshold values.
2. Newly introduced dataset of Far-Infrared Radiation Images leads to development of more precised and more robust deep learning based pedestrian detectors:
   2.1. Incorrect annotation of the objects in two biggest FIR image datasets does not allow us to use these datasets for training and as a benchmark of deep-learning based pedestrian detectors.
   2.2. 16bit image modification for Darknet framework is able to increase pedestrian detector precision by 9.5% mAP to YOLOv3 detector and 11.2% mAP to TinyV3 detecor.

2.3. Temperature based adaptive threshold applied for histogram equalization reduces the intensity of hot objects and is able to increase pedestrian detector precision by 3.2% mAP to YOLOv3 detector and for TinyV3 detector by 4.3% mAP in ZUT dataset.

2.4. The Environmental condition's impact on the ZUT dataset images reduce mAP by 58.7% for YOLOv3 detector and 53.9% for TinyV3 detector when detectors are trained on the SCUT dataset.

3. The improvement of the pedestrian detectors could be successfully done by combining several datasets, additionally boosting the speed using network architecture slimming procedures:

3.1. Re-training of the deep learning based pedestrian detector using examples selected according to the confidence distribution may increase the precision by 6.24% mAP of ResNext50 detector.

3.2. By utilising slimming introduced by (Liu *et al.* 2017b) and confidence distribution training, it is possible to decrease 4.83 times needed floating point operations, gain 11.9 FPS running detector on AGX as well as increasing accuracy by 8.38% mAP for ResNext50 detector.

4. Applying deep learning based image dataset augmentation it is possible to gain precision in the way:

4.1. The proposed image augmentation approach based on DcDNN is able to gain the precision of 9.38% mAP to ResNext50 detector.

4.2. The proposed confidence heat map is able to gain the precision of pedestrian detection by adding from 0.18% mAP to 1.02% mAP on unknown images.

References

Afzal, M. Z.; Kölsch, A.; Ahmed, S.; Liwicki, M. 2017. Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 883–888. [see 40 p.]

AM Online Projects 2019. *Guangzhou Weather by Months*. Accessed: 2020-01-03, Available online at: <https://en.climate-data.org/asia/china/guangdong/guangzhou-2309/>. [see 33 p.]

Aslan, M. F.; Durdu, A.; Sabanci, K.; Mutluer, M. A. 2020. CNN and HOG based comparison study for complete occlusion handling in human tracking, *Measurement* 158: 107 704. ISSN 0263-2241. [see 19 p.]

Barrière, F.; Druart, G.; Guerineau, N.; Lasfargues, G.; Fendler, M.; Lhermet, N.; Taboury, J. 2012. Compact infrared cryogenic wafer-level camera: Design and experimental validation, *Applied optics* 51: 1049–60. [see 10, 11 p.]

Baumer Group Corporation 2019. *Baumer SDK*. Accessed: 2020-01-03, Available online at: <https://www.baumer.com/ch/en/product-overview/industrial-cameras-image-processing/software/baumer-gapi-sdk/c/14174>. [see 37 p.]

Bertozzi, M.; Fedriga, R. I.; Miron, A.; Reverchon, J.-L. 2013. Pedestrian detection in poor visibility conditions: would swir help?, in *International conference on image analysis and processing*, Springer, 229–238. [see 9 p.]

Bilal, M.; Hanif, M. S. 2020. Benchmark revision for HOG-SVM pedestrian detector through reinvigorated training and evaluation methodologies, *IEEE Transactions on Intelligent Transportation Systems* 21(3): 1277–1287. [see 17 p.]

Bilinski, P.; Bremond, F.; Kaâniche, M. 2009. Multiple object tracking with occlusions using HOG descriptors and multi resolution images, in *3rd International Conference on Imaging for Crime Detection and Prevention*, London, United Kingdom. [see 15 p.]

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN 0387310738. [see 16 p.]

Bochkovskiy, A. 2019a. *Darknet YOLO implementation*. Accessed: 2020-01-03, Available online at: <https://github.com/AlexeyAB/darknet/commit/dcfeea30f195e0ca1210d580cac8b91b6beaf3f7>. [see 43, 54 p.]

Bochkovskiy, A. 2019b. *ResNext50 configuration*. Accessed: 2020-05-03, Available online at: <https://github.com/AlexeyAB/darknet/blob/master/cfg/csresnext50-panet-spp.cfg>. [see 53 p.]

Bochkovskiy, A. 2019c. *Tiny YOLOv3 configuration with 3 layers*. Accessed: 2020-05-03, Available online at: <https://github.com/AlexeyAB/darknet/blob/master/cfg/yolov3-tiny_3l.cfg>. [see 53 p.]

Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y. M. 2020a. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. Accessed: 2020-12-29, Available online at: <http://arxiv.org/abs/2004.10934>. [see 21 p.]

Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y. M. 2020b. *YOLOv4: Optimal speed and accuracy of object detection*. Accessed: 2020-05-03, Available online at: <http://arxiv.org/abs/2004.10934>. [see 21 p.]

Bradski, A. 2008. *Learning OpenCV, [Computer Vision with OpenCV Library ; software that sees]*. 1st edition. O'Reilly Media. ISBN 0-596-51613-4. Gary Bradski and Adrian Kaehler. [see 27 p.]

Breen, J. M.; Næss, P. A.; Hansen, T. B.; Gaarder, C.; Stray-Pedersen, A. 2020. Serious motor vehicle collisions involving young drivers on norwegian roads 2013–2016: Speeding and driver-related errors are the main challenge, *Traffic injury prevention* 21(6): 382–388. [see 2 p.]

Catanzaro, B. E.; Dombrowski, M.; Hendrixson, J.; Hillenbrand, E. 2004. Design of dual-band SWIR/MWIR and MWIR/LWIR imagers, in *Infrared Technology and Applications XXX*, vol. 5406, ed. by Andresen, B. F.; Fulop, G. F., International Society for Optics and Photonics, SPIE, 829 – 835. [see 10 p.]

Chan, K. C.; Ayvaci, A.; Heisele, B. 2015. Partially occluded object detection by finding the visible features and parts, in *2015 IEEE International Conference on Image Processing (ICIP)*, 2130–2134. ISSN null. [see 42 p.]

Chen, Y.; Shin, H. 2020. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network, *Applied Sciences* 10(3). ISSN 2076-3417. [see 22 p.]

Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J. S.; An, K.; Kweon, I. S. 2018. Kaist multi-spectral day/night data set for autonomous and assisted driving, *IEEE Transactions on Intelligent Transportation Systems* 19(3): 934–948. ISSN 1524-9050. [see 31 p.]

Dai, J.; Li, Y.; He, K.; Sun, J. 2016. R-FCN: Object detection via region-based fully convolutional networks, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*: NIPS'16, Red Hook, NY, USA: Curran Associates Inc., 379–387. ISBN 9781510838819. [see 21 p.]

Dalal, N.; Triggs, B. 2005. Histograms of oriented gradients for human detection, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 886–893 vol. 1. [see 15 p.]

Das, S.; Brimley, B. K.; Lindheimer, T. E.; Zupancich, M. 2018. Association of reduced visibility with crash outcomes, *IATSS research* 42(3): 143–151. [see 2 p.]

Davis, J. W.; Keck, M. A. 2005. A two-stage template approach to person detection in thermal imagery, in *Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05) - Volume 1 - Volume 01*: WACV-MOTION '05, Washington, DC, USA: IEEE Computer Society, 364–369. ISBN 0-7695-2271-8-1. [see 31 p.]

Davis, J. W.; Keck, M. A. 2005. A two-stage template approach to person detection in thermal imagery, in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, vol. 1, 364–369. [see 31 p.]

De Smedt, F. 2015. *Pedestrian detection for real-life applications*. [see 15 p.]

Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. 2012. Pedestrian detection: An evaluation of the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4): 743–761. [see 16 p.]

Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. 2011. Pedestrian detection: An evaluation of the state of the art, *IEEE transactions on pattern analysis and machine intelligence* 34: 743–61. [see 21 p.]

Druart, G.; Barrière, F. D. L.; Chambon, M.; Guérineau, N.; Lasfargues, G.; Fendler, M. 2013. Cryogenic wafer-level MWIR camera: laboratory demonstration, in *Infrared Technology and Applications XXXIX*, vol. 8704, ed. by Andresen, B. F.; Fulop, G. F.; Hanson, C. M.; Norton, P. R.; Robert, P., International Society for Optics and Photonics, SPIE, 652 – 661. [see 10 p.]

Eichhorn, K.; Abel, B.; Burg, M. 2001. Improvement of night vision using infrared headlamps, *ATZ worldwide* 103. [see 9 p.]

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; Zisserman, A. 2010. The pascal visual object classes (voc) challenge, *International journal of computer vision* 88(2): 303–338. [see 21 p.]

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. 2008. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. Available online at: <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. [see 14 p.]

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. 2011. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. Available online at: <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. [see 30 p.]

Fleet, D. J.; Black, M. J.; Yacoob, Y.; Jepson, A. D. 2000. Design and use of linear models for image motion analysis, *Int. J. of Computer Vision* 36(3): 171–193. [see 15 p.]

FLIR Systems Inc 2018. *FLIR Thermal Sensing for ADAS*. Accessed: 2019-06-11, Available online at: <https://www.flir.com/oem/adas/adas-dataset-form/>. [see 31 p.]

FLIR Systems Inc 2019. *FLIR Path Finder kit*. Accessed: 2019-06-11, Available online at: <http://www.safetyvision.com/sites/safetyvision.com/files/FLIR_PathFindIRII_User_Guide_1.pdf>. [see 12 p.]

FLIR Systems Inc 2020. *FREE FLIR Thermal Dataset for Algorithm Training*. Accessed: 2021-02-15, Available online at: <https://www.flir.com/oem/adas/adas-dataset-form/>. [see 22 p.]

Forslund, D.; Bjärkefur, J. 2014. Night vision animal detection, in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. ISSN 1931-0587. [see 12 p.]

Geiger, A.; Lenz, P.; Urtasun, R. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 3354–3361. [see 21 p.]

Gonzalez Alzate, A.; Fang, Z.; Socarras, Y.; Serrat, J.; Vázquez, D.; Xu, J.; López, A. 2016. Pedestrian detection at day/night time with visible and fir cameras: A comparison, *Sensors* 16: 820. [see 11, 31 p.]

Guo, C.; Zhan, Y. 2018/05. Fast deformation part model with CNN for face detection, in *Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, Atlantis Press, 408–413. ISBN 978-94-6252-517-7. ISSN 1951-6851. [see 19 p.]

He, K.; Zhang, X.; Ren, S.; Sun, J. 2014. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*. Accessed: 2018-02-15, Available online at: <http://arxiv.org/abs/1406.4729>. [see 21 p.]

He, K.; Zhang, X.; Ren, S.; Sun, J. 2016a. Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. [see 20 p.]

He, K.; Zhang, X.; Ren, S.; Sun, J. 2016b. Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. [see 53 p.]

Hoang, V.-D.; Le, M.-H.; Jo, K.-H. 2014. Hybrid cascade boosting machine using variant scale blocks based HOG features for pedestrian detection, *Neurocomputing* 135: 357–366. ISSN 0925-2312. [see 16, 17 p.]

Hochreiter, S. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6: 107–116. [see 19 p.]

Hussain, M.; Bird, J. J.; Faria, D. R. 2019. A study on CNN transfer learning for image classification, in *Advances in Computational Intelligence Systems*, ed. by Lotfi, A.; Bouchachia, H.; Gegov, A.; Langensiepen, C.; McGinnity, M., Cham: Springer International Publishing, 191–202. ISBN 978-3-319-97982-3. [see 18 p.]

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I. S. 2015. Multispectral pedestrian detection: Benchmark dataset and baselines, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [see 16 p.]

Jegham, I.; Ben Khalifa, A. 2017. Pedestrian detection in poor weather conditions using moving camera, in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, 358–362. [see 31 p.]

Jegham, I.; Khalifa, A. B. 2017. Pedestrian detection in poor weather conditions using moving camera, in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 358–362. [see 22 p.]

Jeong, M.; Ko, B. C.; Nam, J. 2017. Early detection of sudden pedestrian crossing for safe driving during summer nights, *IEEE Transactions on Circuits and Systems for Video Technology* 27(6): 1368–1380. [see 11 p.]

Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. 2019. A survey of deep learning-based object detection, *IEEE Access* 7: 128 837–128 868. [see 21 p.]

Kaarmukilan, S.; Poddar, S.; *et al.* 2020. FPGA based deep learning models for object detection and recognition comparison of object detection comparison of object detection models using FPGA, in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 471–474. [see 21 p.]

Khan, M. A.; Khan, S. F. 2018. Iot based framework for vehicle over-speed detection, in *2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, IEEE, 1–4. [see 2 p.]

Khellal, A.; Ma, H.; Fei, Q. 2015. Pedestrian classification and detection in far infrared images, in *Intelligent Robotics and Applications*, ed. by Liu, H.; Kubota, N.; Zhu, X.; Dillmann, R.; Zhou, D., Cham: Springer International Publishing, 511–522. ISBN 978-3-319-22879-2. [see 31 p.]

Kim, B.; Yuvaraj, N.; Ramasamy, S.; Santhosh, R.; Sabari, A. 2020. Enhanced pedestrian detection using optimized deep convolution neural network for smart building surveillance, *Soft Computing* . [see 19 p.]

Kim, T.; Kim, S. 2018. Pedestrian detection at night time in fir domain: Comprehensive study about temperature and brightness and new benchmark, *Pattern Recognition* 79: 44–54. ISSN 0031-3203. Available online at: <https://www.sciencedirect.com/science/article/pii/S0031320318300414>. [see 11, 31 p.]

Krishna, S. 2005. Quantum dots-in-a-well infrared photodetectors, *Infrared Physics & Technology* 47(1): 153–163. ISSN 1350-4495. QWIP 2004. [see 10 p.]

Kristan, M.; Matas, J.; Leonardis, A.; Vojir, T.; Pflugfelder, R.; Fernandez, G.; Nebehay, G.; Porikli, F.; Čehovin, L. 2016. A novel performance evaluation methodology for single-target trackers, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(11): 2137–2155. ISSN 0162-8828. [see 45 p.]

Krizhevsky, A.; Sutskever, I.; Hinton, G. 2012. Imagenet classification with deep convolutional neural networks, *Neural Information Processing Systems* 25. [see 19 p.]

Kumar, T.; Kushwaha, D. S. 2016. An efficient approach for detection and speed estimation of moving vehicles, *Procedia Computer Science* 89: 726–731. [see 2 p.]

Kurita, T.; Otsu, N.; Abdelmalek, N. 1992. Maximum likelihood thresholding based on population mixture models, *Pattern recognition* 25(10): 1231–1240. [see 27 p.]

Lanka, P.; Rangaprakash, D.; Gotoor, S. S. R.; Dretsch, M. N.; Katz, J. S.; Denney, T. S.; Deshpande, G. 2020. Malini (machine learning in neuroimaging): A MATLAB toolbox for aiding clinical diagnostics using resting-state fMRI data, *Data in Brief* 29: 105 213. ISSN 2352-3409. [see 40 p.]

Le, T.; Zheng, Y.; Zhu, C.; Luu, K.; Savvides, M. 2016. Multiple scale Faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 46–53. [see 21 p.]

Li, C.; Zhou, Z. 2019. Visual question answering with dynamic parameter prediction using functional hashing, in *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence*: RICAI 2019, New York, NY, USA: Association for Computing Machinery, 330–335. ISBN 9781450372985. [see 19 p.]

Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. 2017. Focal loss for dense object detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. [see 21 p.]

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. 2017. Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. [see 21 p.]

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. 2014. Microsoft coco: Common objects in context, in *Computer Vision – ECCV 2014*, ed. by Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T., Cham: Springer International Publishing, 740–755. ISBN 978-3-319-10602-1. [see 21 p.]

Liu, S.; Huang, D.; Wang, Y. 2017a. *Receptive Field Block Net for Accurate and Fast Object Detection*. Accessed: 2018-08-13, Available online at: <http://arxiv.org/abs/1711.07767>. [see 21 p.]

Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. 2018. *Path Aggregation Network for Instance Segmentation*. Accessed: 2019-09-15, Available online at: <http://arxiv.org/abs/1803.01534>. [see 21 p.]

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A. C. 2016. Ssd: Single shot multibox detector, *Lecture Notes in Computer Science* 21–37. ISSN 1611-3349. [see 21 p.]

Liu, Y.; Zeng, L.; Huang, Y. 2014. An efficient HOG–ALBP feature for pedestrian detection, *Signal, Image and Video Processing* 8: 125–134. [see 28 p.]

Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. 2017b. Learning efficient convolutional networks through network slimming, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [see 63, 66, 72, 74 p.]

Mahapatra, A.; Mishra, T. K.; Sa, P. K.; Majhi, B. 2013. Background subtraction and human detection in outdoor videos using fuzzy logic, in *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–7. [see 17 p.]

Miron, A. D. 2014. *Multi-modal, Multi-Domain Pedestrian Detection and Classification: Proposals and Explorations in Visible over StereoVision, FIR and SWIR*: Theses. INSA de Rouen; Universitatea Babeș-Bolyai (Cluj-Napoca, Roumanie). [see 31 p.]

Miron, A. D.; Bensrhair, A.; Fedriga, R. I.; Broggi, A. 2013. Swir images evaluation for pedestrian detection in clear visibility conditions, *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* 354–359. [see 9, 10 p.]

Mohanty, S.; Hughes, D.; Salathe, M. 2016. Using deep learning for image-based plant disease detection, *Frontiers in Plant Science* 7. [see 40 p.]

Mustaqeem, A.; Anwar, S.; Majid, M. 2018. Multiclass classification of cardiac arrhythmia using improved feature selection and svm invariants, *Computational and Mathematical Methods in Medicine* 2018: 1–10. [see 40 p.]

Negied, N. K.; Hemayed, E. E.; Fayek, M. B. 2015. Pedestrians' detection in thermal bands–critical survey, *Journal of Electrical Systems and Information Technology* 2(2): 141–148. [see 10 p.]

Nguyen, C. T.; Havlicek, J. P.; Fan, G.; Caulfield, J. T.; Pattichis, M. S. 2014. Robust dual-band mwir/lwir infrared target tracking, in *2014 48th Asilomar Conference on Signals, Systems and Computers*, IEEE, 78–83. [see 10 p.]

Nowosielski, A.; Małecki, K.; Forczmański, P.; Smoliński, A. 2020. Pedestrian detection in severe lighting conditions: Comparative study of human performance vs thermal-imaging-based automatic system, in *Progress in Computer Recognition Systems*, ed. by Burduk, R.; Kurzynski, M.; Wozniak, M., Cham: Springer International Publishing, 174–183. ISBN 978-3-030-19738-4. [see 13 p.]

Overett, G.; Petersson, L.; Brewer, N.; Andersson, L.; Pettersson, N. 2008. A new pedestrian dataset for supervised learning, in *2008 IEEE Intelligent Vehicles Symposium*, 373–378. [see 30 p.]

Polat, K.; Akdemir, B.; Güneş, S. 2008. Computer aided diagnosis of ECG data on the least square support vector machine, *Digital Signal Processing* 18(1): 25–32. ISSN 1051-2004. [see 40 p.]

Prihatmaja, P. A.; Widyantoro, D. H. 2019. Improving performance of YOLOv3 for vehicle detection, in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. [see 21 p.]

Reach PLC 2019. *Driver WARNING - Your dash cam could land you up to £9,000 fine and see you JAILED abroad*. Accessed: 2020-01-03, Available online at: <https://www.express.co.uk/life-style/cars/998528/Dash-cam-car-Europe-fines-prison/>. [see 39 p.]

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. 2016. You only look once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. [see 21 p.]

Redmon, J.; Farhadi, A. 2017. YOLO9000: Better, faster, stronger, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525. [see 21 p.]

Redmon, J.; Farhadi, A. 2018a. *YOLOv3: An Incremental Improvement*. Accessed: 2019-02-15, Available online at: <http://arxiv.org/abs/1804.02767>. [see 21 p.]

Redmon, J.; Farhadi, A. 2018b. *YOLOv3: An Incremental Improvement*. Accessed: 2020-05-03, Available online at: <http://arxiv.org/abs/1804.02767>. [see 38 p.]

Redmon, J.; Farhadi, A. 2018c. *YOLOv3: An incremental improvement*. Accessed: 2020-05-03, Available online at: <http://arxiv.org/abs/1804.02767>. [see 53 p.]

Ren, S.; He, K.; Girshick, R.; Sun, J. 2015a. Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39. [see 21 p.]

Ren, S.; He, K.; Girshick, R.; Sun, J. 2015b. Faster R-CNN: Towards real-time object detection with region proposal networks, in *Advances in Neural Information Processing Systems*, vol. 28, ed. by Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; Garnett, R., Curran Associates, Inc. [see 21 p.]

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L. 2014. *ImageNet Large Scale Visual Recognition Challenge*. Accessed: 2020-02-15, Available online at: <http://arxiv.org/abs/1409.0575>. [see 21 p.]

Saito, H.; Hagihara, T.; Hatanaka, K.; Sawai, T. 2008. Development of pedestrian detection system using far-infrared ray camera, *SEI Technical Review* 112–117. [see 11, 12 p.]

Schindler IT-Solutions 2019. *Fine against individual in Austria*. Accessed: 2020-01-03, Available online at: <https://easygdpr.eu/gdpr-incident/strafe-gegen-privatperson-wegen-dashcam/>. [see 39 p.]

Shopovska, I.; Jovanov, L.; Philips, W. 2019. Deep visible and thermal image fusion for enhanced pedestrian visibility, *Sensors* 19(17). ISSN 1424-8220. [see 12, 22 p.]

Sidla, O.; Rosner, M. 2007. HOG pedestrian detection applied to scenes with heavy occlusion, in *Intelligent Robots and Computer Vision XXV: Algorithms, Techniques, and Active Vision*, vol. 6764, ed. by Casasent, D. P.; Hall, E. L.; Röning, J., International Society for Optics and Photonics, SPIE, 88 – 98. [see 15 p.]

Silberstein, S.; Levi, D.; Kogan, V.; Gazit, R. 2014. Vision-based pedestrian detection for rear-view cameras, in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 853–860. [see 30 p.]

Simonyan, K.; Zisserman, A. 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Accessed: 2018-02-15, Available online at: <http://arxiv.org/abs/1409.1556>. [see 21 p.]

Socarras, Y.; Ramos, S.; Vázquez, D.; López, A.; Gevers, T. 2013. Adapting pedestrian detection from synthetic to far infrared images, in *Computer Vision in Vehicle Technology*. [see 31 p.]

Soviany, P.; Ionescu, R. T. 2018. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction, in *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 209–214. [see 21 p.]

Sun, X.; Hu, H.; Habib, E.; Magri, D. 2011. Quantifying crash risk under inclement weather with radar rainfall data and matched-pair method, *Journal of Transportation Safety & Security* 3(1): 1–14. [see 2 p.]

Szarvas, M.; Yoshizawa, A.; Yamamoto, M.; Ogata, J. 2005. Pedestrian detection with convolutional neural networks, in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, 224–229. [see 18 p.]

Sze, V.; Chen, Y.; Yang, T.; Emer, J. S. 2017. Efficient processing of deep neural networks: A tutorial and survey, *Proceedings of the IEEE* 105(12): 2295–2329. ISSN 1558-2256. [see 30 p.]

Taiana, M.; Nascimento, J. C.; Bernardino, A. 2013. An improved labelling for the INRIA person data set for pedestrian detection, in *Pattern Recognition and Image Analysis*, ed. by Sanches, J. M.; Micó, L.; Cardoso, J. S., Berlin, Heidelberg: Springer Berlin Heidelberg, 286–295. ISBN 978-3-642-38628-2. [see 15 p.]

Takumi, K.; Watanabe, K.; Ha, Q.; Tejero-De-Pablos, A.; Ushiku, Y.; Harada, T. 2017. Multispectral object detection for autonomous vehicles, in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 35–43. [see 22 p.]

Tanabe, K. 2018. Pareto's 80/20 rule and the gaussian distribution, *Physica A: Statistical Mechanics and its Applications* 510: 635–640. ISSN 0378-4371. [see 40 p.]

Tech, D. S. 2019. *Ybat - YOLO BBox Annotation Tool*. Accessed: 2020-01-03, Available online at: <https://github.com/drainingsun/ybat>. [see 40 p.]

Teutsch, M.; Muller, T.; Huber, M.; Beyerer, J. 2014. Low resolution person detection with a moving thermal infrared camera by hot spot classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 209–216. [see 11 p.]

Tomè, D.; Monti, F.; Baroffio, L.; Bondi, L.; Tagliasacchi, M.; Tubaro, S. 2016. Deep convolutional neural networks for pedestrian detection, *Signal Processing: Image Communication* 47: 482–489. ISSN 0923-5965. [see 19 p.]

Toyota Motor Asia Pacific Pte Ltd 2005. *Night View*. Accessed: 2020-01-03, Available online at: <http://www.toyota-myanmar.com/innovation/safety-technology/safety-technology-2/safety-technology-3/radar-cruise-control-2/night-view>. [see 9, 13 p.]

Tsimhoni, O.; Bärgman, J.; Flannagan, M. J. 2007. Pedestrian detection with near and far infrared night vision enhancement, *Leukos* 4(2): 113–128. [see 9 p.]

Van Beeck, K.; Van Engeland, K.; Vennekens, J.; Goedemé, T. 2017. Abnormal behavior detection in lwir surveillance of railway platforms, in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 1–6. [see 11 p.]

Wang, C.; Mark Liao, H.; Wu, Y.; Chen, P.; Hsieh, J.; Yeh, I. 2020. CSPNet: A new backbone that can enhance learning capability of CNN, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1571–1580. [see 53 p.]

Wang, X.; Han, T. X.; Yan, S. 2009. An HOG-LBP human detector with partial occlusion handling, in *2009 IEEE 12th International Conference on Computer Vision*, 32–39. [see 16 p.]

Wong, K.-Y. 2020. *ResNet50 configuration*. Accessed: 2020-05-03, Available online at: <https://github.com/WongKinYiu/CrossStagePartialNetworks/blob/master/cfg/csresnet50-panet-spp.cfg>. [see 53 p.]

World Health Organisation, W. 2019. *European regional status report on road safety 2019*. Accessed: 2021-01-03, Available online at: <https://www.euro.who.int/en/publications/abstracts/european-regional-status-report-on-road-safety-2019/>. [see 2 p.]

World Health Organization, W. 2018. *Global status report on road safety 2018*. Accessed: 2020-01-03, Available online at: <https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/>. [see 1 p.]

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. 2017a. Aggregated residual transformations for deep neural networks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500. [see 53 p.]

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. 2017b. Aggregated residual transformations for deep neural networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987–5995. [see 20 p.]

Xu, Q.; Zhang, C.; Zhang, L. 2015. Denoising convolutional neural network, in *2015 IEEE International Conference on Information and Automation*, 1184–1187. [see 68 p.]

Xu, Z.; Zhuang, J.; Liu, Q.; Zhou, J.; Peng, S. 2019. Benchmarking a large-scale FIR dataset for on-road pedestrian detection, *Infrared Physics & Technology* 96: 199–208. ISSN 1350-4495. [see 31 p.]

Yagi, S.; Kobayashi, S.; Inoue, T.; Hori, T.; Michiba, N.; Okui, K. 2003. *The Development of Infrared Projector*. Report, SAE Technical Paper. [see 9 p.]

Zhang, X.; Gao, H.; Xie, G.; Gao, B.; Li, D. 2017. Technology and application of intelligent driving based on visual perception, *CAAI Transactions on Intelligence Technology* 2(3): 126–132. [see 12 p.]

Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. 2019. Object detection with deep learning: A review, *IEEE Transactions on Neural Networks and Learning Systems* PP: 1–21. [see 30 p.]

Zhou, C.; Yuan, J. 2017. Multi-label learning of part detectors for heavily occluded pedestrian detection, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 3506–3515. ISSN 2380-7504. [see 42 p.]

List of Scientific Publications by the Author on the Topic of the Dissertation

Papers in the Reviewed Scientific Journals

Tumas, P.; Serackis, A.; Nowosielski, A. 2021. Augmentation of Severe Weather Impact to Far-Infrared Sensor Images Improve Pedestrian Detection System, *Electronics* Vol. 10, p. 1–16. ISSN 2079-9292.

Tumas, P.; Nowosielski, A.; Serackis, A. 2020. Pedestrian Detection in Severe Weather Conditions, *IEEE Access* 62775–62784. ISSN 2169-3536.

Jonkus, A.; Tumas, P.; Serackis, A. 2018. Investigation of convolutional neural networks for visual tracking of pedestrians, *International journal of advanced research (IJAR).* Vol. 6, iss. 5. p. 668–674. ISSN 2320-5407; DOI: 10.21474/IJAR01/7082

Papers in Other Editions

Tumas, P.; Serackis A. 2018. Automated Image Annotation based on YOLOv3, in *Proceedings of AIEEE Conference,* 1–4. DOI:10.1109/AIEEE.2018.859216

Tumas, P.; Jonkus, A.; Serackis, A. 2018. Acceleration of HOG based Pedestrian Detection in FIR Camera Video Stream, in *Proceedings of eSTREAM Conference,* 1–4. DOI: 10.1109/eStream.2017.7950322

Tumas, P.; Serackis, A. 2017. Effective Background Subtraction Algorithm for Food Inspection using a Low-Cost Near Infrared Camera, in *Proceedings of eSTREAM Conference,* 1–4. ISSN 12169-3536.

Tumas, P.; Serackis, A. 2016. ORB Feature based Matching of Two-Dimensional Electrophoresis Gel Images. in *Proceedings of 20th international conference*
Biomedical Engineering, 117–120. ISSN 2029-3380

Summary in Lithuanian

Įvadas

Problemos formulavimas

Disertacijoje sprendžiama pėsčiųjų aptikimo vaizde esant nepalankioms aplinkos ir oro sąlygoms problema. Disertacijoje tiriami vaizdai, gauti naudojant infraraudonosios spinduliuotės jutiklius, gebančius suformuoti žmogaus akimi aiškiai interpretuojamą vaizdą prietemoje ir net tais atvejais, kai lyja, aplinką gaubia rūkas ar į vaizdo kameros pusę yra nukreiptas šviesos šaltinis. Matomojo spektro jutikliai dėl tiesioginio šviesos srauto ženkliai susiaurina dinaminį diapazoną ir aplinkoje esantys objektai tampa nematomais. Deja infraraudonosios spinduliuotės jutiklių formuojami vaizdai tiesiogiai priklauso nuo aplinkoje esančių objektų temperatūros, o snygio ar lietaus užteršiami vaizdo kamerų optikos elementai iškraipo šiluminės spinduliuotės poveikį jutikliui ir gaunami mažiau kontrastingi vaizdai, kurių automatizuota analizė kelia papildomų iššūkių.

Šiuo metu rinkoje naudojamos masinės gamybos infraraudonosios spinduliuotės vaizdo kameros taip pat apsunkina pėsčiųjų aptikimą, nes jose naudojamų jutiklių skyra yra labai maža lyginant su matomo spektro jutikliais ir pėstysis gali būti vos penkių jutiklio taškų aukščio. Mažos galios centrinių procesorinių įrenginių ar net grafikos spartintuvų greitaveikos vis dar nepakanka, norint atlikti giliojo mokymo dirbtinių neuronų tinklais grįstų aptiktuvų skaičiavimus realiuoju laiku, nors šio tipo aptiktuvų veikimo tikslumas šiuo metu yra didžiausias. Disertacijoje siekiama ne tik patobulinti giliojo mokymo dirbtinių neuronų tinklais grįstų aptiktuvų tikslumą esant įvairioms oro sąlygoms, tačiau ir pasiūlyti būdų, kaip tokio tipo aptiktuvus paspartinti ir pritaikyti veikimui realiuoju laiku automobiliuose.

Darbo aktualumas

Jutikliai, jautrūs šiluminei spinduliuotei, gali būti naudojami kaip puiki alternatyva matomos šviesos jutikliams pažangios pagalbos vairuotojams sistemose, sankryžų stebėjimo sistemose ir kituose taikymuose, kur svarbu išlaikyti efektyvų aplinkos vizualinį stebėjimą bet kokiomis oro sąlygomis ir bet kuriuo paros metu.

Remiantis 2018 metais pateikta Pasaulio sveikatos organizacijos (PSO) ataskaita apie saugumą keliuose (World Health Organization 2018), mirčių skaičius eismo įvykiuose siekė 1,35 mln. Pusė šių eismo įvykių priklauso dviratininkų ir pėsčiųjų eismo dalyvių kategorijai. Viena pagrindinių eismo įvykių priežasčių yra greičio viršijimas ir išsiblaškęs vairuotojų elgesys. Anot PSO pranešimo, vidutinio greičio viršijimas daro tiesioginę įtaką avarijų tikimybei. Pavyzdžiui, padidinus vidutinį greitį 1%, mirtinos avarijos rizika padidėja 4 %, o išsiblaškęs vairavimas, kurį, pavyzdžiui, lemia mobiliojo telefono naudojimas dar papildomai padidina ją iki 3 %. Telefono naudojimas vairuojant tai pat lėtina reakcijos stabdymui bei į eismo signalus laiką, apsunkina važiavimą teisinga eismo juosta ir tinkamo atstumo laikymąsi.

Siekdama užkirsti kelią nelaimingiems atsitikimams, Europos Sąjunga (ES) įvedė pažangios pagalbos vairuotojams sistemoms skirtus saugos standartus, su kuriais būtų automobiliuose privalomai diegiamos naujos funkcijos: pažangus greičio išlaikymas, avarinio stabdymo ir eismo juostų laikymosi sistemos, vairuotojo mieguistumo, dėmesio stebėjimo sistemos. Be to, PSO kiekvienais metais ruošia ES regioninę kelių eismo saugumo ataskaitą (World Health Organisation 2019). Ši ataskaita rodo, kad kasdien Europos regionų keliuose žūva daugiau nei 221 žmogus, dar tūkstančiai yra sužeidžiami ar tampa neįgalūs. Tyrimo duomenimis, daugiau nei 30 % žuvusiųjų eismo dalyvių yra pėstieji ir dviratininkai. Pagrindinės žuvusiųjų priežastys yra didėjantis transporto priemonių skaičius keliuose, prasti saugumo standartai ir infrastruktūra, vairuotojų išsiblaškymas ar apsvaigimas nuo narkotikų, alkoholio, saugos diržų ar šalmų nenaudojimas. Greičio viršijimas yra dar vienas kritinis elementas, dėl kurio pritrūksta laiko išvengti nelaimingo atsitikimo (Breen *et al.* 2020; Khan, Khan 2018; Kumar, Kushwaha 2016). Galiausiai, prastos oro sąlygos, tokios kaip lietus, sniegas ir rūkas yra matomumą bloginantys faktoriai. Rūkas ir dūmai didina tikimybę patirti sunkią traumą iki 3,24 karto, o sukelti kelių eismo įvykį 1,53 karto (Das *et al.* 2018). Panašiame tyrime buvo analizuota lietaus ir kelio dangos tipo įtaka eismo įvykiams (Sun *et al.* 2011). Šis tyrimas parodė, kad rizika jog eismo įvykis įvyks esant blogoms oro sąlygoms padidėja iki 2,61 karto.

Tyrimų objektas

Pagrindinis daktaro disertacijos tyrimo objektas yra vaizdai, kuriuose yra pėstieji, užfiksuoti tolimosios infraraudonosios spinduliuotės (TIS) vaizduose. Taip pat giliųjų neuroninių tinklų (GNT) struktūros bei duomenų apdorojimo metodai.

Darbo tikslas

Disertacijos tikslas – patobulinti pėsčiųjų aptikimo tolimosios infraraudonosios spinduliuotės vaizduose metodus, sukuriant naujus mokymams skirto duomenų rinkinio papildymo būdus, ištiriant aptiktuvo struktūros optimizavimo ir efektyvaus mokymo strategijas.

Darbo uždaviniai

Disertacijos tikslui pasiekti suformuluoti keturi uždaviniai:

1. Paspartinti orientuotų gradientų histograma pagrįsto pėsčiųjų aptikimo algoritmo veikimą naudojant centrinį procesorinį įrenginį.

2. Sukurti adaptyvų pėsčiųjų aptiktuvą, įvertinantį temperatūros įtaką tolimųjų infraraudonųjų spindulių spinduliuotės vaizdams.

3. Optimizuoti krašto kompiuterijai skirtą pėsčiųjų aptiktuvą, grįstą sąsūkos dirbtinių neuronų tinklais.

4. Ištirti duomenų rinkinio padidinimo būdus, siekiant pagerinti pėsčiųjų aptikimą esant blogoms oro sąlygoms.

Tyrimų metodika

Darbe naudojamos skaitmeninio vaizdo apdorojimo, dirbtinių neuronų tinklų, giliojo mokymosi, statistinės analizės teorijos. Pritaikytas ir įgyvendintas vaizdo apdorojimas, GNT mokymas ir vykdymas, greičio ir tikslumo įvertinimas. Surinkti duomenys važiuojant automobiliu, naudojant sukurtą prototipą. Atlikti eksperimentiniai tyrimai, skirti mokyti GNT, siekiant aptikti pėsčiuosius TIS spektre. Naudoti programinės įrangos paketai, tokie kaip: „Cuda", „Darknet", „OpenCV", „Pytorch", „Matlab", „WolframAlpha". Taip pat buvo vykdomi mokymai kompiuterių klasteryje, sudarytame iš AMD 3900X ir „Intel i7" 8-osios kartos procesorių. Be to, buvo naudojami grafiniai greitintuvai, tokie kaip NVIDIA RTX2080Ti ir NVIDIA Geforce GTX1080Ti.

Darbo mokslinis naujumas

Išsprendus disertacijos uždavinius, buvo viešai paskelbtas duomenų rinkinys, kuriame yra 122 000 anotacijų, iš kurių daugiau nei 79 000 surinktų per šlapdribą ar lietų. Likusios anotacijos buvo surinktos esant šalnoms ir debesuotomis oro sąlygomis. Šis duomenų rinkinys unikalus ir svarbus kuriant ir tiriant pėsčiųjų aptikimo algoritmus, nes iki rinkinio paskelbimo, analogiško duomenų rinkinio pasaulyje nebuvo. Be to, naujai paskelbtame duomenų rinkinyje pateikti automobilio CAN magistralės duomenys, kurie gali būti naudojami kuriant šiuolaikines ADAS sistemas kartu su tolimosios infraraudonosios spinduliuotes vaizdo aptiktuvais. Disertacijoje atlikti eksperimentiniai tyrimai parodė, kad naujajame rinkinyje surinktų TIS vaizdų išplėstas dinaminis diapazonas iki 16 bitų pagerina pėsčiųjų aptiktuvo, grįsto GNT mokymą ir aptikimo tikslumą. Pasiūlytas mokymo ir GNT aptiktuvo struktūros mažinimo būdas parodo, kaip pašalinti nereikalingus GNT elementus, kad būtų padidintas apdorojimo greitis ir neprarandamas pėsčiųjų aptikimo tikslumas. Be to, naudojant vaizdų papildymą su atvirkščiai veikiančiu triukšmą šalinančiu GNT tinklu, galima dirbtinai sukurti naujus vaizdus, imituojančius vaizdo iškraipymus dėl blogų oro sąlygų, leidžiančias pagerinti aptiktuvo mokymą ir tikslumą.

Darbo rezultatų praktinė reikšmė

Disertacijos rengimo metu pasiūlyta nauja metodika GNT mokymui, atlikti eksperimentiniai tyrimai, pasiūlyti GNT architektūros mokymo ir optimizavimo būdai, surinktas ir paskelbtas atvirai prieigai duomenų rinkinys, suteikiantis galimybę tobulinti ir tęsti vairuotojų pagalbinių sistemų tyrimus ir jų plėtrą. Pasiūlytas aptiktuvų mokymo ir struktūros optimizavimo būdas padeda kelis kartus pagerinti aptiktuvo veikimo greitį, neprarandant vidutinio tikslumo. Pasiūlytas TIS vaizdų rinkinio didinimo būdas padeda išplėsti neuroniniais tinklais grįsto aptiktuvo mokymo galimybes naudojant mažos apimties ir įvairovės duomenų rinkinius.

Ginamieji teiginiai

1. Nuo 8 bitų iki 16 bitų išplėstas tolimosios infraraudonosios spinduliuotės vaizdų dinaminis diapazonas pagerina TinyV3 struktūros pėsčiųjų aptiktuvo vidutinį tikslumą 11,2 %, išlaikant tą patį atpažinimo greitį.

2. Mokymo pavyzdžių pasirinkimas pagal mažos aptiktuvo aptikimo tikimybės pasiskirstymą pagerina ResNext50 struktūros pėsčiųjų aptiktuvo vidutinį tikslumą 6,24 %, išlaikant tą patį atpažinimo greitį.

3. ResNext50 struktūrą galima optimizuoti sumažinant skaičiavimus 4,83 karto, pagreitinant vaizdo apdorojimą 11,9 kadrų per sekundę ir pagerinant vidutinį pėsčiųjų aptikimo tikslumą 8,38 % su AGX kompiuteriu.

4. Duomenų rinkinio padidinimas imituojant blogo oro įtaką TIS vaizdams pagerina ResNext50 struktūros pėsčiųjų aptiktuvo vidutinį tikslumą 9,38 %, išlaikant tą patį atpažinimo greitį.

Disertacijos struktūra

Disertaciją sudaro: įvadas, trys skyriai, bendrosios išvados, literatūros sąrašas su atskirai pateiktomis autoriaus publikacijomis ir trys priedai. Darbo apimtis yra 102 puslapiai, kuriuose yra pateikta: 4 formulės, 43 paveikslai ir 37 lentelės.Disertacijoje remtasi 126 kitų autorių literatūros šaltinių.

1. Šiluminės spinduliuotės vaizdų pėsčiųjų aptiktuvų apžvalga

Pėsčiųjų aptikimas visada buvo aktuali tema vaizdų atpažinimo taikymo srityje, ypač infraraudonųjų spindulių spektro srityje dėl dviejų pagrindinių priežasčių: mažos šiluminės kameros skiriamosios gebos, kuri suteikia mažai informacijos apie pėsčiųjų išvaizdą, aprangą, kūno formą, ir didelio masto pėsčiųjų infraraudonųjų spindulių duomenų rinkinio trūkumo, kad būtų užtikrintas sėkmingas giliųjų neuronų tinklų mokymasis ir tikslus pėsčiųjų aptikimas. Taip pat esant skirtingai aplinkos temperatūrai, šiluminės kameros užfiksuotas vaizdas skiriasi, o tai apsunkina sukurti universalų pėsčiųjų aptiktuvo modelį, kuris visada veiktų tiksliai, nepriklausomai nuo sezoniškumo ar paros laiko. Be to, pėsčiųjų ap-

tikimas taip pat yra ir sudėtinga užduotis, nes egzistuoja daugybė skirtingų pėsčiųjų vaizdų variacijų, tokių kaip:

- skirtingos kūno pozos,
- pėsčiųjų kūno persidengimai,
- dalinai pėsčiuosius užstojantys objektai,
- vizualiai panašūs objektai, kaip kelio ženklai ar stulpeliai,
- saulės atokaitoje įšilę medžiai ar konstrukcinės detalės.

Vienas iš pirmųjų šio skyriaus tikslų yra apžvelgti skirtingus infraraudonųjų spindulių spektro juostų pločius, siekiant nustatyti, kurios spektro dalys pateikia geriausią informaciją apie pėsčiųjų aptikimą infraraudonųjų spindulių spektro srityje. Toliau analizuojama, kur trūksta pėsčiųjų aptikimo taikymo tyrimų. Vėliau šiame skyriuje apžvelgiami esami pramonės sprendimai ir žinomi pažangiausi aptikimo metodai.

TIS skiriasi nuo kitų infraraudonųjų spindulių spektro bangos ilgio juostų, nes suteikia detalų žmogaus kūno vaizdą (Kim, Kim 2018). Taip pat pastebima, kad esant blogam orui, TIS veikia geriau nei kitos kameros, nes tolimosios infraraudonosios spinduliuotės spinduliai yra mažiau jautrūs drėgmei nei kitų bangų juostų spinduliai (Saito *et al.* 2008). Skirtingai nei artimosios infraraudonosios spinduliuotės ar matomo spektro kamerų jutikliai, TIS kameros jutikliai nėra jautrūs trikdančiai šviesai, pavyzdžiui, artėjantiems iš priekio automobilio žibintams. Galimi TIS trūkumai yra tie, kad dieną ir ypač vasarą pėsčiųjų ir fone esančių objektų (pvz., pastatų, kelių) temperatūrų skirtumas yra labai mažas, todėl yra sunkiai pastebimi pėstieji (Jeong *et al.* 2017). Be to, TIS kameros yra brangesnės (objektyvai ir vaizdo jutikliai yra ypač brangūs) nei artimosios infraraudonosios spinduliuotės kameros ar matomo spektro kameros. Panašius teiginius taip pat suformalavo Gonzalez Alzate *et al.* 2016, kur autoriai bandė naudoti dvi kameras (matomo spektro ir TIS). Pasak autoriaus, „aptiktuvas veikė geriausiai, derindamas matomo spektro kamerą su TIS vaizdais."

Vienas iš pirmųjų bandymų sukurti pėsčiųjų aptikimo aptiktuvą buvo paskelbtas 2005 metais (Dalal, Triggs 2005). Aptiktuvo įėjime buvo naudojamas ne vaizdas, o apskaičiuoti histogramomis orientuoti gradientai (HOG). Taip pat buvo paskelbtas sudėtingas duomenų rinkinys, kuriame buvo daugiau nei 1800 anotuotų žmogaus vaizdų. Tačiau, laikui bėgant atsirado pažangesnių būdų aptikti objektus vaizduose ir vienas iš jų – Gilieji Neuronų Tinklai (GNT) (Krizhevsky *et al.* 2012). AlexNet buvo tai pirmasis GNT tinklas objektams aptikti, turintis penkis sluoksnius, naudojantis grafikos spartintuvo procesorių ir paskelbimo metu aplenkęs tikslumu visus tuomet buvusius aptiktuvus.

Patys moderniausi GNT tipo aptiktuvai yra klasifikuojami į dvi kategorijas – vienos pakopos ir dviejų pakopų aptiktuvai. Vienos pakopos aptiktuvai pranašūs tuo, kad su šiuo metu esančia aparatine įranga jau galima pasiekti realiojo laiko vaizdų apdorojimą objektams aptikti, tačiau jie nėra tikslūs. Dviejų pakopų aptiktuvai geba beveik be klaidų aptikti vaizduose objektus, tačiau su pačia galingiausia aparatine įranga pasiekia vos kelių kadrų per sekundę apdorojimą (Dai *et al.* 2016; Kaarmukilan *et al.* 2020; Le *et al.* 2016; Ren *et al.* 2015a,b). Šiuo metu vyraujantys vienos pakopos aptiktuvai YOLO (Bochkovskiy *et al.* 2020a; Redmon *et al.* 2016; Redmon, Farhadi 2017, 2018a), SSD (Liu *et al.* 2016), ResNet50 ir ResNext50, kurių tikslumas varijuoja nuo 58,5 % VT iki 97,5 % VT.

2. Pėsčiųjų aptiktuvo prototipo tyrimai

Tradicinis HOG (Histograma orientuotų gradientų) vaizdo objektų požymių skaičiavimo būdas grindžiamas analizės lango stumdymu per vaizdo kadrą. Jis lemia pastebimą vaizdo apdorojimo užlaikymą todėl buvo pasiūlytas pagreitinimas, kuris pagrįstas objektų atskyrimu nuo fono. S2.1 lentelėje yra palyginti pirminiai keturių aptiktuvų rezultatai, kur HOG aptiktuvas sugebėjo aptikti pėsčiuosius 0,5 Hz greičiu ir buvo lėčiausias lyginant su kitais aptiktuvais.

Atlikus aptiktuvo modifikacijas, buvo palyginta trys objektų atskyrimo nuo fono būdai – du dinaminiai ir vienas parinktas eksperimentiškai. S2.2 lentelėje yra pateikti aptikimo rezultatai, kurioje matomas algoritmo pagreitėjimas 12 kartų, slenkančio-lango sprendimą pakeitus į fono pašalinimą pagal slenkstį. Be to pastebima, jog automatiniai slenksčio parinkimo būdai pasiekia iki 80 % aptikimo greičio, kurį galima pasiekti eksperimentiškai parenkant slenkstines vertes. Padaryta išvada, kad fono pašalinimu grįstas metodas priklauso nuo esančių objektų skaičiaus vaizde ir netinka aptikti pėsčiųjų realiuoju laiku.

Po pirmojo eksperimento, buvo pradėtos duomenų rinkinio paieškos pėsčiųjų aptiktuvo mokymui. Remiantis literatūroje skelbiamais tyrimais (Sze *et al.* 2017; Zhao *et al.* 2019), aptiktuvo tikslumas priklauso nuo duomenų rinkinio pavyzdžių įvairovės, naudojamos aptiktuvo įvesties / tipo ir įgyvendinimo detalių. Dabartiniai GNT praktiniai tyrimai parodė, kad norint pasiekti tikslų aptiktuvo veikimą reikia turėti tūkstančius įvairių vaizdų. Matomo spektro duomenų rinkinių skirtų pėstiesiems aptikti yra labai daug, pvz., PASCAL VOC 2012 (Everingham *et al.* 2011), GM-ATCI (Silberstein *et al.* 2014), NICTA (Ove-

S2.1 lentelė. Nemodifikuotų aptiktuvų vidutinis apdorojimo greitis

Aptiktuvas	KPS	SD
HOG	0,5	0,17
Faster R-CNN	1,1	0,08
YOLOv2	2,3	0,21
Tiny YOLO	7,0	0,05

S2.2 lentelė. Modifikuotų aptiktuvų vidutinis apdorojimo greitis

Aptiktuvas	Otsu (KPS)	Gauso (KPS)	Eksperimentinis (KPS)
4 objektai + HOG	4,0	3,2	**6,0**
4 objektai + Faster RCNN	1,3	1,2	1,6
Tik 4 objektai	10,0	8,0	13,0
6 objektai + HOG	3,0	2,7	5,1
6 objektai + Faster RCNN	0,7	0,9	1,3
Tik 6 objektai	8,0	7,1	11,0
8 objektai + HOG	2,1	1,9	3,9
8 objektai + Faster RCNN	0,5	0,6	0,9
Tik 8 objektai	5,1	5,3	8,8
10 objektai + HOG	3,0	2,7	5,1
10 objektai + Faster RCNN	0,4	0,3	0,7
Tik 10 objektai	2,4	2,1	3,5
Be objektų	12,0	9,0	15,0

rett *et al.* 2008), INRIA ir kt. Tačiau esant poreikiui mokyti GNT, naudojant tolimuosios infraraudonosios spinduliuotės vaizdus, galima rasti tik iki dešimties laisvai prieinamų duomenų rinkinių:

- CVC-09 (Socarras *et al.* 2013),
- CVC-14 (Gonzalez Alzate *et al.* 2016),
- FLIR-ADAS (FLIR Systems Inc 2018),
- KAIST (Choi *et al.* 2018),
- KMU (Jegham, Ben Khalifa 2017),
- LSIFIR (Khellal *et al.* 2015),
- OTCBVS (Davis, Keck 2005),
- RISWIR (Miron 2014),
- Terravic Motion IR (Davis, Keck 2005),
- SCUT (Xu *et al.* 2019).

KAIST ir SCUT yra reprezentatyviausi duomenų rinkiniai, skirti naudoti vairuotojų pagalbinėse sistemose. Tačiau kiti duomenų rinkiniai netinka pagalbinėms sistemoms kurti, nes vaizdai yra paimti iš kamerų, primontuotų prie pastato sienų, trikojų ir dronų (Kim, Kim 2018). SCUT duomenų rinkinys, skirtingai nuo kitų duomenų rinkinių, turi vaizdų, kurie užfiksuoti, vairuojant automobilį miesto centre, priemiesčiuose, miesteliuose ir greitkeliuose. Šiuo metu šis duomenų rinkinys yra didžiausias pagal kadrų ir anotacijų skaičių (apimantis 216 000 kadrų ir 448 000 anotacijų). SCUT pateikti vaizdai, užfiksuoti 384×288 skyros jutikliu ir interpoliuoja į 720×576 skyrą. Vaizdai SCUT rinkinyje paruošti griežtai laikantis iš anksto nustatyto anotavimo protokolo. Šiame rinkyje yra pateikta šešios skirtingos klases, apibūdinančios pėsčiuosius skirtingose situacijose. KAIST yra antras pagal dydį duomenų rinkinys, kuriame yra daugiaspektriniai vaizdai, užfiksuoti matomame ir tolimosios infraraudonosios spinduliuotės spektruose. Vaizo įrašai buvo daromi keliose vietovėse, tokiose kaip miesteliai, miestai ir miesto pakraščiai. KAIST rinkinio vaizdai buvo anotuoti rankiniu būdu, išskiriant tris klases. Duomenų rinkinyje iš viso yra 103 128 anotacijos ir 95 000 kadrų.

Tačiau abu duomenų rinkiniai turi trūkumų. Pavyzdžiui, KAIST duomenų rinkinyje yra blogai sužymėtų arba visai nesužymėtų situacijų, o SCUT rinkinyje nemažai anotacijų, kai keli pėstieji susilietę rankomis yra pažymėti viena anotacija. Taip pat nei vienas rinkinys nepateikia tiek neapdorotų (RAW) vaizdų, tiek vaizdų, esant blogoms oro sąlygoms. Be to, nepateikiama informacijos apie automobilio greitį, lauko temperatūrą ir t. t. Dėl šių priežasčių buvo surinktas naujas duomenų rinkinys, pavadintas ZUT-FIR-ADAS (ZUT).

S2.3 lentelėje yra palyginti KAIST, SCUT ir ZUT duomenų rinkiniai, kur ZUT duomenų rinkinys turi antrą pagal dydį anotuotų vaizdų skaičių, užfiksuotą keturiose Europos Sąjungos šalyse esant blogoms oro sąlygoms. Be to, pateikiami duomenys, nuskaityti iš automobilio prietaisų duomenų magistralės (CAN), įskaitant važiavimo greitį, stabdžių pedalo būseną ir lauko temperatūrą. Taip pat ZUT duomenų rinkinys išsiskiria, kaip turintis didžiausią kelio scenų įvairovę, klasių skaičių bei platesniu skiltiškumu.

S2.3 lentelė. KAIST, SCUT ir ZUT duomenų rinkinių palyginimas

Parametras	KAIST	SCUT	ZUT
kadrai	95 000	216 000	110 000
klasės	4	6	9
anotacijos	103 000	448 000	122 000
kelio scenos	3	4	10
fiksuojamas atstumas	2,4–61 m	4,6–132 m	10–100 m
skiltiškumas	8 bitai	8 bitai	16 bitai
temperatūra	nepateikta	nepateikta	-0,5 to 12 °C
didžiausias greitis	nepateikta	80 km/h	180 km/h
didžiausias greitis	7,5–13,0 µm	8,0–14,0 µm	7,5–13,0 µm

S2.4 lentelėje, yra parodyta, kaip anotacijų kiekis ZUT duomenų rinkinyje pasiskirsto pagal klases, kur didžioji dauguma ZUT duomenų rinkinio anotacijų sudaro pėstieji ir dviratininkai. S2.5 lentelėje yra pateikta ne su pėsčiaisiais susijusių anotacijų kiekis ZUT duomenų rinkinyje, kur didžioji dalis anotacijų yra sužymėta žmogaus kūno dalys, kai persidengia su priešais esančiais objektais. S2.6 lentelėje yra pateikta informacija apie ZUT duomenų rinkinio fiksuotas oro sąlygas, šalį (kurioje filmuota) ir mokymo bei testavimo duomenų rinkinius padalinimą. Svarbu paminėti, kad dauguma pažymėtų anotacijų yra nufilmuota esant dulksnai arba lietuj.

Surinkus duomenų rinkinį buvo pradėti mokyti YOLOv3 ir TinyV3 aptiktuvai, lyginant 8 ir 16 bitų skiltiškumo vaizdus. S2.7 lentelėje pateikiama informacija apie eksperimento konfigūraciją ir gautus rezultatus. Aptiktuvų testavimo metu naudoti du vidutinio

S2.4 lentelė. Anotacijų pasiskirstymas ZUT duomenų rinkinyje

	Pėstysis	Persidengę	Dviratininkai	Motociklininkai	Paspirtukininkai
Mokymo	59 649	4008	7908	173	94
Testavimo	21 083	1112	2355	49	0
Viso	**80 732**	5120	**10 263**	**222**	**94**

S2.5 lentelė. Kitų anotacijų pasiskirstymas ZUT duomenų rinkinyje

	Kūno dalys	Neaiškūs	Vežimėliai	Gyvūnai
Mokymo	16 611	14	27	140
Testavimo	9091	1	10	107
Viso	25 702	15	37	**247**

S2.6 lentelė. Anotacijų pasiskirstymas pagal oro sąlygas

Šalis	Rinkinys	Dulksna	Šalna	Lietus	Debesuota	Rūkas	Giedra
Danija	Mokymo	**20 886**	0	**37 064**	3051	0	0
Danija	Testavimo	**16 291**	0	0	0	0	0
Vokietija	Mokymo	0	10 206	13	0	1535	0
Lenkija	Mokymo	212	0	0	15 657	0	0
Lenkija	Testavimo	0	0	1153	6441	0	752
Lietuva	Testavimo	3687	0	25	5459	0	0

S2.7 lentelė. ZUT duomenų rinkinio mokymo rezultatai

Skyra	Versija	Skiltiškumas, bitai	Netektis	VT50	VT25	Žingsniai
416×416	YOLOv3	16	0,2448	80,5	91,5	251 000
416×416	TinyV3	16	0,2954	66,3	86,0	243 000
640×480	YOLOv3	16	0,2514	85,4	92,5	220 000
640×480	TinyV3	16	0,2681	79,1	92,3	250 000
640×480	YOLOv3 + TI	16	**0,1514**	**89,1**	**95,4**	383 000
640×480	TinyV3 + TI	16	**0,1681**	**82,3**	**94,2**	420 000
640×480	YOLOv3 + TI	8	0,1914	**79,6**	92,3	123 000
640×480	TinyV3 + TI	8	0,2414	**71,1**	89,1	78 000

S2.8 lentelė. YOLOv3 ir TinyV3 aptiktuvų mokymo rezultatai (8 bitų skiltiškumo)

Aptiktuvas	Testavimo rinkinys	VT	Persidengimas
YOLOv3	ZUT mokymo	**37,7**	50
TINYv3	ZUT mokymo	**32,4**	50
YOLOv3	ZUT mokymo	55,0	25
TINYv3	ZUT mokymo	45,8	25
YOLOv3	ZUT testavimo	**27,7**	50
TINYv3	ZUT testavimo	**25,4**	50
YOLOv3	ZUT testavimo	37,8	25
TINYv3	ZUT testavimo	33,9	25

tikslumo vertinimo kriterijai (esant skirtingam persidengimo kriterijui (angl. *IoU*): 50 ir 25), netekties funkcija ir iteracijų skaičius. Iš pradžių geriausias rezultatas buvo gautas su YOLOv3 aptiktuvu – pasiektas 80,5 % VT tikslumas. Po 243 000 mokymo žingsnių TinyV3 pasiekė tik 66,3 % VT, o tai rodo, kad aptiktuvas negali išmokti naujų situacijų. Po to aptiktuvo įvestis buvo padidinta iki 640×480 skyros ir YOLOv3 tikslumas pagerėjo iki 85,4 % VT, o TinyV3 iki 79,1 % VT. Taip pat, vaizdai buvo apdoroti naudojant intensyvumo slenkstį (angl. *()*TI), kuris reguliuoja maksimalų taškų intensyvumą kadre pagal automobilio temperatūros jutiklį. Naudojant didesnę tinklo įvestį ir apdorotus vaizdus, YOLOv3 tikslumas pagerėjo iki 89,1 % VT ir TinyV3 iki 82,3 % VT.

Eksperimento pabaigoje buvo apmokyti YOLOv3 ir TinyV3 aptiktuvai naudojant SCUT duomenų rinkinį. S2.8 lentelėje pateikti mokymo rezultatai rodo, jog YOLOv3 pasiekė 86,4% VT ir TinyV3 79,3% VT. Pagal šiuos rezultatus palyginti YOLOv3 ir TinyV3 aptiktuvai, apmokymui naudojant 8 bitų skiltiškumo ZUT duomenų rinkinio vaizdus ir testuojant SCUT duomenų rinkinyje ir atvirkščiai. Tyrimas parodė, kad abu aptiktuvai prastai aptinka pėsčiuosius tiek ZUT, tiek SCUT duomenų rinkiniuose. Todėl buvo padaryta išvada, kad tolimesniuose tyrimuose yra tikslinga abu duomenų rinkinius sujungti į vieną bei tęsti aptiktuvų mokymą ir testavimą naudojant apjungtą duomenų rinkinį.

3. Pėsčiųjų aptiktuvo gerinimas ir eksperimentiniai tyrimai

Eksperimentinėje dalyje buvo tęsiamas aptiktuvų tyrimas apjungiant SCUT ir ZUT duomenų rinkinius. Lentelėje S3.1 yra pateikta galutiniai skaičiai, kurie parodo jog naujas

S3.1 lentelė. Duomenų rinkinys apjungus ZUT ir SCUT

Rinkinys	Mokymo kadrai	Mokymo anotacijos	Mokymo kadrai	Testavimo kadrai
ZUT	69 455	88 624	40 103	33 808
SCUT	78 942	118 377	76 381	122 537
Viso	**148 397**	**207 001**	**116 484**	**156 345**

S3.2 lentelė. Mokymo rezultatai, apjungus ZUT ir SCUT rinkinius bei permokymui naudojant mažiausiai „įsitikinęs" strategiją

Aptiktuvas	KPS	VT apjungtu	VT permokius
TinyV3	**55,57**	73,25	78,77
TinyL3	43,10	80,14	80,13
YOLOv3	17,88	80,48	84,41
YOLOv4	15,97	**86,05**	**86,69**
ResNet50	**19,82**	**81,00**	82,84
ResNext50	17,70	77,07	**83,31**

duomenų rinkinys turi daugiau nei 207 001 anotacijų GNT mokymui ir 156 345 anotacijų patikrai. Taip pat ekperimentų rezultatų patikimumo vertinimui buvo mokomi 6 aptiktuvai: TinyV3 (Redmon, Farhadi 2018c); a TinyV3 su papildoma aptikimo etapu pavadinta TinyL3 (Bochkovskiy 2019c); YOLOv3; YOLOv4; ResNet50 (He *et al.* 2016b; Wong 2020) ir ResNeXt50 (Bochkovskiy 2019b; Wang *et al.* 2020; Xie *et al.* 2017a).

S3.2 lentelėje yra pateikti pirminiai duomenys, gauti naudojant apjungtą duomenų rinkinį. Tiksliausias aptiktuvas buvo YOLOv4, pasiekęs 86,05 % VT ir 15,97 kadrų per sekundę greitį. Mažiausiai tikslus buvo TinyV3 aptiktuvas, pasiekęs vos 73,25 % VT, tačiau jis sugebėjo apdoroti vaizdą 55,57 kadrų per sekundę greičiu.

Siekiant pagerinti tikslumą buvo padaryta prielaida, kad aptiktuvas išmoksta bendrai pasikartojančias situacijas, tačiau neišmoksta mokymo metu retai pasitaikančių pavyzdžių. Todėl buvo nuspręsta atrinkti situacijas kur aptiktuvas yra mažiausiai „įsitikinęs", kad anotacijoje yra pažymėtas žmogus ir toliau jį apmokyti tik šiose situacijose. S3.2 lentelėje parodyti rezultatai, kuriuos pavyko gauti naudojant tokią strategiją. Pasirodo, jog pasitelkus šią strategiją galima pasiekti net 6,24 % VT prieaugį, neprarandant apdorojimo greičio.

Vėliau, buvo panaudoti vienos plokštės mažos galios NVIDIA Jetson TX2 (TX2) id NVIDIA AGX Xavier (AGX) kompiuteriai ir pamatuotas apmokytų aptiktuvų vaizdo apdorojimo greitis. S3.3 lentelėje parodyti matavimo rezultatai, kur geriausiai pasirodė

S3.3 lentelė. Kadrų apdorojimo greičio matavimai naudojant aptiktuvus su TX2 ir AGX kompiuteriais

Aptiktuvas	BFLOPS	TX2	AGX
TinyV3	9,67	**10,1**	**42,8**
TinyL3	12,60	8,9	35,0
YOLOv3	115,93	2,2	5,7
YOLOv4	105,73	2,1	5,3
ResNet50	86,52	3,1	6,5
ResNext50	82,59	2,3	5,1

S3.4 lentelė. Aptiktuvo optimizavimo antros iteracijos rezultatai

Aptiktuvas	TX2	AGX	RTX	VT50	BFLOP
TinyV3	**12,3**	**87,9**	229,0	78,10	2,45
TinyL3	10,8	49,5	**251,6**	80,27	3,85
YOLOv3	4,3	13,0	35,8	**85,33**	34,37
YOLOv4	5,8	14,3	77,2	70,27	18,60
ResNet50	11,5	21,8	112,8	78,58	10,72
ResNext50	6,3	17,0	**78,6**	**85,45**	17,09

S3.5 lentelė. Papildomo mokymo rezultatai

Aptiktuvas	VT (papildžius)	VT (su žemėlapiu)
TinyV3	78,22	78,74
TinyL3	81,98	82,73
YOLOv3	83,87	**84,91**
YOLOv4	**87,02**	87,20
ResNet50	82,72	83,44
ResNext50	**86,45**	86,67

AGX kompiuteris gebantis apdoroti 42,8 kadrus per sekundę, TX2 nusileido sparta, bet vis tiek pasiekė 10,1 kadro per sekundę apdorojimo greitį, naudojant TinyV3 aptiktuvą. TinyL3 buvo antras greičiausias aptiktuvas pasiekęs 35,0 kadrus per sekundę su AGX kompiuteriu, o TX2 pasiekė 8,9 kadro per sekundę apdorojimo greitį.

Toliau sekė paieška, kaip būtų galima supaprastinti aptiktuvų struktūrą, pritaikant aptiktuvus krašto kompiuterijai. Pasirinkta Liu *et al.* 2017b pasiūlyta strategija, kuria buvo stebimos neaktyvios aptiktuvo filtrų zonos bei naudojama prieš tai disertacijoe pasiūlyta mokymo strategija. Pasirodo, detektoriaus struktūrą galima sumažinti iki keturių kartų neprarandant tikslumo, nes aptiktuvų struktūros yra bendrosios paskirties (nespecializuotos) ir sprendžia bendrą objektų aptikimo problemą. S3.4 lentelėje yra pateikti geriausi mokymo rezultatai, gauti panaudojus abi strategijas bei matyti, kad YOLOv3 ir ResNext50 aptiktuvai pagerino pirminį rezultatą neprarandant tikslumo bei pagerinant kadrų per sekundę apdorojimo greitį.

Kitas uždavinys buvo ištirti duomenų rinkinio padidinimo būdus, siekiant pagerinti pėsčiųjų aptikimą esant blogoms oro sąlygoms. Naujiems vaizdams sugeneruoti panaudotas sąsūkomis grįstas neuronų specializuotas GNT (Xu *et al.* 2015), kuriuo dažniausiai yra šalinamas triukšmas. Toliau, pašalinus „baltojo triukšmo" generatorių ir pakeitus tikslo funkciją, specializuotas GNT leido išmokti blogų oro sąlygų bruožus vaizduose ir juos perkelti į SCUT duomenų rinkinio vaizdus. S3.5 lentelėje yra pateikta papildomai apmokytų aptiktuvų rezultatai, iš kurių matyti, kad geriausius rezultatus pasiekė YOLOv4 aptiktuvas. Didžiausią tikslumo prieaugį pasiekė ResNext50 aptiktuvas (9,38 % VT). Taip pat buvo pasiūlyas pėsčiųjų aptikimo vietos vaizde tikimybės žemėlapis, kuriuo pavyko pagerinti aptiktuvo tikslumą. S3.5 lentelėje yra pateikti aptiktuvų rezultatai, kur geriausias fiksuotas prieaugis yra YOLOv3 aptiktuvo ir siekia 1,02 % VT.

Bendrosios išvados

1. Klasikiniai pėsčiųjų aptikimo metodai paremti histograma orientuotų gradientų bruožais leidžia pagreitinti slenkančio lango žingsnį:
   1.1. Taikant vaizdo fono pašalinimą vietoje analizės slenkančiu langu, galima iki 12 kartų paspartinti HOG požymius naudojančio pėsčiųjų aptiktuvo tikslumą.
   1.2. Fono pašalinimu grįsto pėsčiųjų aptikimo algoritmo veikimo greitis priklauso nuo vaizde aptiktų objektų skaičiaus ir nėra tinkamas veikti realiuoju laiku.
   1.3. Automatinių slenksčio parinkimo būdų naudojimas pėsčiųjų aptikimo algoritme sulėtina algoritmo veikimą iki 20 %, lyginant su fiksuotą slenkstį turinčiais algoritmais.
2. Naujai pristatytas tolimosios infraraudonosios spinduliuotės duomenų rinkinys leidžia sukurti tikslesnius ir tvaresnius aptiktuvus:
   2.1. 16 bitų dinaminio diapazono TIS vaizdai leidžia padidinti pėsčiųjų aptiktuvų vidutinį tikslumą 9,5 % jei naudojame YOLOv3 bei 11,2 % jei naudojame TinyV3 struktūras.
   2.2. Temperatūra pagrįstas maksimalios taškų intensyvumo ribos keitimas, leidžia sumažinti „karštų" objektų intensyvumą vaizde ir padidina YOLOv3 aptiktuvo vidutinį tikslumą 3,2 %, o TinyV3 aptiktuvo vidutinį tikslumą padidina 4,3 % ZUT duomenų rinkinyje.
   2.3. Aplinkos sąlygos vaizdų surinkimo ZUT rinkiniui metu ir TIS jutiklio dinaminis diapazonas keičia vaizdo savybes, svarbias pėsčiųjų aptikimui ir sumažina ant SCUT duomenų rinkinio apmokyto YOLOv3 aptiktuvo vidutinį tikslumą 58,7 %, o TinyV3 aptiktuvo vidutinį tikslumą 53,7 %.
3. Galima sėkmingai patobulinti pėsčiųjų aptiktuvus sujungus kelis duomenų rinkinius į vieną ir papildomai pagreitinti naudojant aptiktuvo architektūros mažinimo procedūrą:
   3.1. Mokymo pavyzdžių pasirinkimas pagal mažos aptiktuvo aptikimo tikimybės pasiskirstymą pagerina ResNext50 struktūros pėsčiųjų aptiktuvo vidutinį tikslumą 6,24 %, išlaikant tą patį atpažinimo greitį.
   3.2. ResNext50 struktūrą galima optimizuoti sumažinant skaičiavimus 4,83 karto, pagreitinant vaizdo apdorojimą 11,9 kadro per sekundę ir pasiekiant 8,38 % aptiktuvo vidutinio tikslumo prieaugį su AGX kompiuteriu.
4. Naudojant giliojo mokymo tinklais grįstą rinkinio plėtimo metodiką galima sukurti tikslesnį aptiktuvą:
   4.1. Naudojant DcDNN mokymo rinkinio plėtimui, galima padidinti ResNext50 struktūros aptiktuvo vidutinį tikslumą 9,38 %.
   4.2. Naudojant šilumos žemėlapius galima padidinti GNT grįsto aptiktuvo vidutinį tikslumą 0,18–1,02 %.

Annexes[1]

Annex A. Declaration of Academic Integrity
Annex B. The Co-authors' Agreements to Present Publications Material in the Dissertation
Annex C. The Copies of Scientific Publications by the Author on the Topic of the Dissertation

[1] The annexes are supplied in the enclosed compact disc.

Paulius TUMAS

IMPROVEMENT OF INTELLIGENT METHODS FOR PEDESTRIAN DETECTION IN FAR-INFRARED RADIATION IMAGES

Doctoral Dissertation

Technological Sciences,
Electrical and Electronic Engineering (T 001)

INTELEKTUALIŲJŲ METODŲ PĖSTIESIEMS APTIKTI TOLIMOSIOS INFRARAUDONOSIOS SPINDULIUOTĖS VAIZDUOSE TOBULINIMAS

Daktaro disertacija

Technologijos mokslai,
elektros ir elektronikos inžinerija (T 001)

2021 06 07. 10 sp. l. Tiražas 20 egz.
Leidinio el. versija *https://doi.org/10.20334/2021-030-M*
Vilniaus Gedimino technikos universitetas
Saulėtekio al. 11, 10223 Vilnius,
Spausdino BĮ UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius