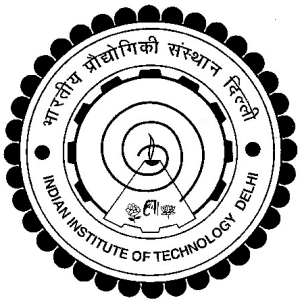
PEDESTRIAN DETECTION IN THE INDIAN ROAD SCENARIO

ABHISHEK GAGNEJA



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
FEBRUARY 2024

PEDESTRIAN DETECTION IN THE INDIAN ROAD SCENARIO

by

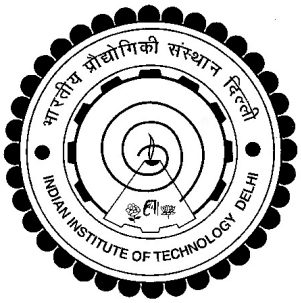
ABHISHEK GAGNEJA

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

FEBRUARY 2024

To my Guru and my family

Certificate

This is to certify that the thesis entitled "**PEDESTRIAN DETECTION IN THE INDIAN ROAD SCENARIO**" being submitted by **Mr. ABHISHEK GAGNEJA** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** is the record of the bonafide research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted either in part or in full to any other university or institute for the award of any degree or diploma.

Prof. Brejesh Lall
Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110016, India.

Date: 21-02-2024
Place: New Delhi

Acknowledgments

I would like to express my sincere gratitude to my advisor **Prof. Brejesh Lall**. The genesis of this thesis would not have been possible without his guidance, critical review and encouragement. His intelligent ideas and comprehensive understanding helped me to establish the overall direction of the research, got me past many significant challenges and always motivated me to strive for excellence.

I thank the members of my thesis committee: Prof. Ranjan Bose, Prof. S.D. Joshi and Prof. Monika Aggarwal for their insightful suggestions which encouraged me to widen my research from various perspectives. I sincerely admire the contribution of all my peers and friends for their unstinting support and cooperative attitude during the course of this study.

I thank my parents for being my support during the happy times and the trying ones. I thank my wife and my son for their unwavering belief in me and for their unconditional love and support.

Most importantly, I thank Navajyothi Shree Karunakara Guru, Santhigiri Ashram, for showing me the path of hope, positivity and faith, without which I would have been lost.

ABHISHEK GAGNEJA

Abstract

Object detection is one of the key problems in core computer vision. With wide-scale applications, the area has been a mainstay of the vision research for over a decade. As a significantly challenging subset of the general problem, pedestrian (or person) detection has been, and to date remains, one of the most challenging vision related tasks. The modelling of any object detector relies on three tasks - dataset, feature extraction backbone and detection head. Each of these tasks are well researched problem statements, but there are still gaps which we attempt to address.

Many datasets have been made available for pedestrian detection over the years, but none with the chaotic complexity of the city roads of a developing nation like India. Hence, in our first work, we created the first video based Indian Pedestrian Detection Dataset, where we recorded pedestrian instances using a dashboard camera on different roads of New Delhi. For ease of annotation, we also developed an in-house video annotation tool, LabelVDOS, which help annotate pedestrian instances across a large number of sequential frames with ease using interpolation and minimal adjustments.

Addressing our second objective, we identified a gap in the training of the popular Feature Pyramid Network (FPN) [1] based detection heads. Multi scale proposals generated by FPN [1] are dependent on a set of various hyper parameters, which critically affect the quality of detection of any model that uses it. In our second

work, therefore, we show the effectiveness of choosing FPN [1] hyper parameters based on the dataset statistics of application when we show an improvement in the detection ability of the ResNet-50 [2] based RetinaNet [3] model when ablated over a large range variations of the hyper parameters in comparison to its vanilla setting on the Caltech [4] as well as the Indian Pedestrian Detection Datasets.

While looking into feature extraction backbones, we found that over time, the research in object detection has evolved from using simply CNN based backbones to the use of models which were earlier designed for completely different tasks. One such crossover that was highly successful was the introduction of transformers [5] in vision applications. We identified the limitation the available models pose in terms of the immense computational capability as well as training time they require and, therefore, employed ConvMixer (CM) [6], a much simpler yet powerful feature extraction backbone which performs at par with Vision Transformers (ViT) [7] on image classification. We integrated the CM model to a Faster RCNN based detection mechanism and created a novel object detector, CM-Det. Upon comparison, we found that CM-Det, with anchor parameters optimized on basis of our second work, outperformed ViTDet [8] on the Indian dataset in terms of mean Average Precision (mAP).

We also developed a socially relevant use case application of pedestrian detection with a Social Distancing Monitoring System. We used a pre-trained YOLO V3 [9] and SORT [10] tracker to generate detection and assign them unique IDs for the duration of their visibility. We then use a combination of homographic projection and Euclidean distance measurement to record whether a given pair of IDs are violating the social distancing norms of distance and duration of violation, hence incorporating both guidelines issued by the WHO regarding social distancing.

To summarise, we present incremental work in prominent research aspects of pedestrian detection as well as introduce a much needed baseline for application of pedestrian detection in the Indian road scenario.

सार

कोर कंप्यूटर विज़न में ऑब्जेक्ट डिटेक्शन प्रमुख समस्याओं में से एक है। व्यापक पैमाने पर अनुप्रयोगों के साथ, यह क्षेत्र एक दशक से अधिक समय से दृष्टि अनुसंधान का मुख्य आधार रहा है। सामान्य समस्या के एक महत्वपूर्ण रूप से चुनौतीपूर्ण उपसमुच्चय के रूप में, पैदल यात्री (या व्यक्ति) का पता लगाना दृष्टि संबंधी सबसे चुनौतीपूर्ण कार्यों में से एक रहा है और आज भी बना हुआ है। किसी भी ऑब्जेक्ट डिटेक्टर का मॉडलिंग तीन कार्यों पर निर्भर करता है - डेटासेट, फीचर एक्सट्रैक्शन बैकबोन और डिटेक्शन हेड। इनमें से प्रत्येक कार्य अच्छी तरह से शोध किए गए समस्या कथन हैं, लेकिन अभी भी कमियां हैं जिन्हें हम संबोधित करने का प्रयास करते हैं।

पिछले कुछ वर्षों में पैदल यात्रियों का पता लगाने के लिए कई डेटासेट उपलब्ध कराए गए हैं, लेकिन भारत जैसे विकासशील देश की शहरी सड़कों की अराजक जटिलता के बारे में कोई जानकारी उपलब्ध नहीं है। इसलिए, अपने पहले काम में, हमने पहला वीडियो आधारित भारतीय पैदल यात्री डिटेक्शन डेटासेट बनाया, जहां हमने नई दिल्ली की विभिन्न सड़कों पर डैशबोर्ड कैमरे का उपयोग करके पैदल चलने वालों की घटनाओं को रिकॉर्ड किया। एनोटेशन में आसानी के लिए, हमने एक इन-हाउस वीडियो एनोटेशन टूल, लेबलवीडीओएस भी विकसित किया है, जो इंटरपोलेशन और न्यूनतम समायोजन का उपयोग करके आसानी से कई अनुक्रमिक फ़्रेमों में पैदल यात्री उदाहरणों को एनोटेट करने में मदद करता है।

अपने दूसरे उद्देश्य को संबोधित करते हुए, हमने लोकप्रिय फ़ीचर पिरामिड नेटवर्क (एफपीएन) आधारित डिटेक्शन हेड्स के प्रशिक्षण में एक अंतर की पहचान की। एफपीएन द्वारा उत्पन्न बहु-स्तरीय प्रस्ताव विभिन्न हाइपरपैरामीटरों के एक सेट पर निर्भर होते हैं, जो इसका उपयोग करने वाले किसी भी मॉडल की पहचान की गुणवत्ता को गंभीर रूप से प्रभावित करते हैं। इसलिए, हमारे दूसरे काम में, हम एप्लिकेशन के डेटासेट आँकड़ों के आधार पर एफपीएन हाइपरपैरामीटर चुनने की प्रभावशीलता दिखाते हैं, जब हम हाइपरपैरामीटर की एक बड़ी रेंज भिन्नता पर पृथक होने पर रेसनेट-50 आधारित रेटिनानेट मॉडल की पहचान क्षमता में सुधार दिखाते हैं। कैलटेक के साथ-साथ भारतीय पैदल यात्री जांच डेटासेट पर इसकी वेनिला सेटिंग की तुलना ।

फीचर निष्कर्षण बैकबोन पर गौर करते हुए, हमने पाया कि समय के साथ, ऑब्जेक्ट डिटेक्शन में अनुसंधान केवल सीएनएन आधारित बैकबोन का उपयोग करने से लेकर उन मॉडलों के उपयोग तक विकसित हुआ है जो पहले पूरी तरह से अलग कार्यों के लिए डिज़ाइन किए गए थे। ऐसा ही एक क्रॉसओवर जो अत्यधिक सफल रहा, वह था दृष्टि अनुप्रयोगों में ट्रांसफार्मर की शुरूआत। हमने उपलब्ध मॉडलों की विशाल कम्प्यूटेशनल क्षमता के साथ-साथ उनके लिए आवश्यक प्रशिक्षण समय के संदर्भ में उत्पन्न होने वाली सीमा की पहचान की और इसलिए, कन्वमिक्सर (सीएम) को नियोजित किया, जो एक बहुत ही सरल लेकिन शक्तिशाली फीचर निष्कर्षण रीढ़ है जो विज़न ट्रांसफॉर्मर (वीआईटी) के बराबर प्रदर्शन करता है। छवि वर्गीकरण पर. हमने सीएम मॉडल को तेज़ आरसीएनएन आधारित डिटेक्शन तंत्र में एकीकृत किया और एक नया ऑब्जेक्ट डिटेक्टर, सीएम-डेट बनाया। तुलना करने पर, हमने पाया कि CM-Det, हमारे दूसरे काम के आधार पर अनुकूलित एंकर मापदंडों के साथ, औसत औसत परिशुद्धता (एमएपी) के मामले में भारतीय डेटासेट पर ViTDet से बेहतर प्रदर्शन करता है।

हमने सामाजिक दूरी निगरानी प्रणाली के साथ पैदल यात्रियों का पता लगाने का एक सामाजिक रूप से प्रासंगिक उपयोग का अनुप्रयोग भी विकसित किया है। हमने पहचान उत्पन्न करने और उनकी दृश्यता की अवधि के लिए उन्हें अद्वितीय आईडी निर्दिष्ट करने के लिए पूर्व-प्रशिक्षित YOLO V3 और SORT ट्रैकर का उपयोग किया। फिर हम यह रिकॉर्ड करने के लिए होमोग्राफिक प्रोजेक्शन और यूक्लिडियन दूरी माप के संयोजन का उपयोग करते हैं कि क्या आईडी की एक जोड़ी दूरी और उल्लंघन की अवधि के सामाजिक दूरी के मानदंडों का उल्लंघन कर रही है, इसलिए सामाजिक दूरी के संबंध में डब्ल्यूएचओ द्वारा जारी किए गए दोनों दिशानिर्देशों को शामिल किया गया है।

संक्षेप में, हम पैदल यात्री पहचान के प्रमुख अनुसंधान पहलुओं में वृद्धिशील कार्य प्रस्तुत करते हैं और साथ ही भारतीय सड़क परिदृश्य में पैदल यात्री पहचान के अनुप्रयोग के लिए एक बहुत जरूरी आधार रेखा पेश करते हैं।

Contents

List of Figures

List of Tables

List of Abbreviations

ADAS Advanced Driving Assistance Systems. 4

AI Artificial Intelligence. 73, 76

BB Bounding Box. 37, 38, 40

CM ConvMixer. iv, 8, 9, 12

CM-Det ConvMixer Detector. iv, 58, 63, 64, 75, 77

CNN Convolutional Neural Network. iv, 17, 21, 22, 56, 57, 59, 60

DETR Detection Transformer. 59

ECP Euro City Persons. 5, 15

FFN Feed Forward Network. 59

FPN Feature Pyramid Network. iii, 19, 20, 35, 36, 37, 39, 40, 42, 54, 55, 59, 75

fps frame per second. 26, 68, 69

ft feet. 6, 65, 67

GPU Graphical Processing Unit. 7, 53, 63, 77, 78

H:W Height to Width. ix, 41, 62

HOG Histogram of Oriented Gradients. 16

HRnet High Resolution Network. 6, 17

IDD Indian Driving Dataset. 5

IOU Intersection Over Union. 47

LabelVDOS Label Very Dense Object Sequences. iii, 9, 11, 28, 34, 74

mAP mean Average Precision. iv, 48, 63

MLP Multi-Layer Perceptron. 59

MOT Multi Object Tracking. 16

MS-COCO Microsoft Common Objects in Context. 14, 21, 42, 46, 63

NCR National Capital Region. 7, 9, 74

NMS Non-Maximum Suppression. 47, 53

OHEM Online Hard Example Mining. 20

PascalVOC Pascal Visual Object Classes. 14

px pixels. 27, 28, 37, 38, 62, 67, 68

RCNN Region-based Convolutional Neural Network. iv, x, xi, 3, 5, 8, 9, 18, 19, 20, 35, 36, 37, 51, 52, 54, 55, 58, 62, 74, 75

ResNet Residual Network. iv, 16, 17, 20, 39, 42, 43, 47

ROI Region of Interest. 18, 19, 62

RPN Region Proposal Network. 9, 19, 35, 52, 53, 58, 61, 62, 64

sec second(s). 6, 65, 68

SGD Stochastic Gradient Descent. 52

SORT Simple Online and Realtime Tracker. iv, 8, 12, 22, 23, 68, 73, 76

SOTA State Of The Art. 21, 60

SPP Spatial Pooling Pyramid. 19

SSD Single Shot Detector. 20

TIMM Torch Image Models. 58, 62

VGG Visual Geometry Group. 16, 17, 18, 52

ViT Vision Transformer. iv, 9, 12, 17, 57, 58, 59, 60, 61, 63, 75, 77

ViTDet Vision Transformer Detector. iv, 8, 9, 12, 58, 63, 75

W:H Width to Height. 51, 62

WHO World Health Organization. iv, 6, 9, 65, 76

YOLO You Only Look Once. iv, 8, 20, 67, 76