TOWARDS ROBUST AND EFFICIENT EMBODIED VISUAL PERCEPTION

ANUPAM SOBTI

AMAR NATH AND SHASHI KHOSLA
SCHOOL OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI

2022

Towards Robust and Efficient Embodied Visual Perception

towards the degree of

Doctor of Philosophy

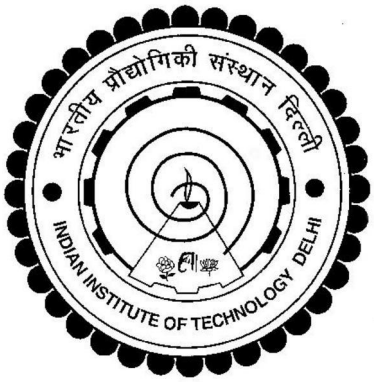
Anupam Sobti

2015ANZ8497

under the guidance of

Prof. M. Balakrishnan

Dr. Chetan Arora



Amar Nath and Shashi Khosla

School of Information Technology

Indian Institute of Technology Delhi

To my parents

and teachers

Certificate

This is to certify that the thesis titled "**Towards Robust and Efficient Embodied Visual Perception**" being submitted by **Anupam Sobti** for the award of **Doctor of Philosophy** is a record of bonafide work carried out by him under my guidance and supervision in the Amar Nath and Shashi Khosla School of Information Technology, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly.

Certain works included in this thesis involved collaboration with other researchers, which has been explicitly specified/acknowledged in the corresponding chapters and the part done by those collaborators appeared in their respective reports/theses.

M. Balakrishnan
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Chetan Arora
Associate Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgement

I express my sincere gratitude to my supervisor, **Prof M. Balakrishnan** who took the time to understand my motivation, always had the right thing to say at the right time, and provided valuable insights into building life as a researcher, a teacher and a citizen of the world. It was an extremely humbling experience to see the world through his eyes and wisdom, and I would carry this experience all through my life. **Prof. Chetan Arora**, my co-supervisor, has been a constant pillar in pulling me through the tough times, focussing me when I was scattered in my thoughts and helping me see brilliance in simple ideas. He has been instrumental towards the development of my rigor and ambition in research. I would like to thank the Ministry of Electronics, Government of India for awarding me with the Visvesvaraya Fellowship which supported me financially during the PhD.

I would like to thank all the B.Tech./M.Tech. students, interns and project staff with whom I worked on MAVI, as a teaching assistant or on research projects. It has been an enriching experience to work with, mentor and learn from them in various ways. I would also like to thank my colleagues who enriched the atmosphere with thought provoking discussions and inspired me through their disciplined work ethic.

I am ever grateful to my family for supporting my decisions throughout my career and their constant encouragement and understanding through the journey.

Last but not the least, I would like to thank all the administrative and lab staff of the Department of Computer Science and Amar Nath and Shashi Khosla School of Information Technology for their continuous support with various activities.

Abstract

Computer vision has had tremendous success recently in perception tasks, especially, in cloud applications. An embodied system brings the additional objectives of memory, compute and energy optimization in addition to system accuracy. As opposed to processing in the cloud, an embodied agent has to perform in the wild without access to additional infrastructure. In this thesis, we address four major limitations of visual perception in embodied systems.

From Images to Videos: When dealing with image streams, there is often redundancy in the temporal neighborhood since objects are often repeated. Thus, a frame-level analysis leans towards the objects present in a larger number of frames rather than a fair evaluation for all objects in a video. To mitigate this, we propose an evaluation metric that clubs together different appearances of each object from multiple frames. In our work, we show how this results in metrics that are sensitive to bias against any objects in the videos.

Resource Constrained Object Detection When object detectors run at different speeds while processing an image stream, a number of frames have to be skipped to maintain real-time operation. Thus, a faster but less accurate detector may be able to process more frames than a slower detector which may be more accurate at the frame level. It is unclear, however, which of these object detectors would be able to detect more objects. This shows the dependence of object detection performance on the speed at which the algorithms are run. In this work, we propose a resource-aware evaluation metric that jointly reasons about speed and accuracy of object detectors for a true evaluation of object detection performance in an embodied system.

Generalization to Unseen Environments: Owing to the large number of parameters in deep neural networks, models often over-fit to the data distribution on which the models are trained. Even as perception improves on a certain dataset, the test environment changes continuously for an embodied agent. There is often limited visibility into the test data distribution in a deployed system. Thus, one has to ensure that the perception module in an embodied agent is aware and adaptive of this shift. We propose a method to use sparse data from an alternate modality (LiDAR) in order to adapt a pre-trained depth prediction network to the current context and provide accurate depth even in unseen environments.

Energy Efficiency with Multimodal Sensing: Visual perception is an energy intensive task. Even human beings rely on their model of the environment or alternate senses for perception, e.g., using sound as a cue for visual engagement while resting. In this work, we observe a similar phenomenon in a multimodal obstacle detection system. We show how a low power ultrasonic sensor is able to provide a low energy sensing mechanism to trigger visual perception. We take a detailed look at the available runtime modes for an obstacle detection system and perform an accuracy energy tradeoff to demonstrate different possibilities.

संक्षेप

दृश्य बोध के क्षेत्र में पिछले दिनों वातावरण के अनुभूति कार्यों में काफी सफलता हासिल की गई है | एक अवतीर्ण उपकरण में सटीकता के अलावा मेमोरी, गणना और ऊर्जा संरक्षण के भी उद्देश्य होते हैं | एक अवतीर्ण उपकरण को ये सभी उद्देश्य बिना किसी क्लाउड सर्वर के कार्यान्वित करने होते हैं | इस शोध प्रबंध में हम दृश्य बोध की चार मुख्य बाधाओं का आँकलन करेंगे ।

चित्रों से वीडिओ तक: जब हम चित्रों की धारा में किसी वस्तु को देखते हैं तो उसमें बहुत सी अतिरेकता होती है । आस पास के चित्रों में एक ही वस्तु बार बार प्रस्तुत होती है । इसलिए हर चित्र पे किए गए वस्तु ढूंढने के कार्य का आँकलन सभी वस्तुओं का निष्पक्ष आँकलन न होकर उन वस्तुओं पे केंद्रित रहता है जो ज्यादा चित्रों में दिखाई देते हैं | इस बाधा का निवारण करने के लिए हुमने एक नया आँकलन माप का निजात किया है । इस कार्य में हमने अनेक प्रयोगों के माध्यम से यह दिखाया है की हमारे सुझाए माप से हम वस्तु ढूंढने के कार्य का निष्पक्ष आँकलन कर पाते हैं ।

संसाधन विवश वस्तु ढूंढने का कार्य: जब वस्तु ढूंढने के तरीके अलग अलग गति से किसी वीडिओ पर चलते हैं तो वास्तविक समय के संचालन को बनाए रखने के लिए उन्हें बीच के चित्रों को छोड़ कर आगे बढ़ना पड़ता है । इस प्रकार एक तेज मगर कम सटीक तरीका अधिक चित्रों पे चल के अधिक वस्तुओं को ढूंढ सकता है । हालांकि फिर भी यह अस्पष्ट है की इनमें से कौनसा तरीका ज्यादा वस्तुओं को ढूंढ पाएगा । यह वस्तु ढूंढने के तरीकों की गति पर निर्भरता को दर्शाता है | इस कार्य में हम एक संसाधन समावेशी माप का निजात किया है जो गति और वस्तु ढूंढने की क्षमता का साथ में अनुमान लगाता है।

उनदेखें वातावरण में समान्यकरणीय क्षमता: अधिक मापदंडों के कारण गहरे तांत्रिक संजाल से चलने वाले अनुभूति कार्य अक्सर अधिकतम फिट हो जाते हैं । इस कारण ये अनदेखे वातावरणों में कार्य नहीं कर पाते । किसी एक वातावरण में अनुभूति सीखने के बावजूद एक अवतीर्ण उपकरण का वातावरण लगातार बदलता रहता है । अक्सर सिस्टम के तैनात होने के बाद भी परीक्षण का माहौल बदलता रहता है । इसलिए यह महत्वपूर्ण है की अनुभूति मापांक इस बदलाव के प्रति जागरूक और अनुरूप है । इस कार्य में हम एक वैकल्पिक साधन (लाइडार) का प्रयोग करके एक पूर्व-प्रशिक्षित संजाल को वर्तमान संदर्भ के प्रति अनुरूपित करने का तरीका प्रस्तुत कर रहे हैं ।

बहुविध संवेदन द्वारा ऊर्जा दक्षिता: दृश्य बोध एक ऊर्जा गहन कार्य है । मनुष्य भी अपनी अनुभूति के लिए अपनी वतरणीय समझ और वैकल्पिक इंद्रियों पर निर्भर रहता है, जैसे की आराम करते हुए आँखों से ज्यादा ध्वनि पर निर्भर रहना । इस कार्य में हम बहुविध सवेधन में इसी तरह के तथ्य का अनुभव करते है । हम एक ऊर्जा संरखक्षणीय अल्ट्रसानिक संवेदक के माध्यम से दृश्य बोध को संचालित करते है जिससे ऊर्जा की बचत की जाती है । हम विस्तृत रूप से इस उपकरण की संचालन विधि का परीक्षण करते है ।

List of Figures

List of Tables

Contents