17th Conference on Computer and Robot Vision - CRV 2020
Program Schedule

[Link to the CRV2020 Slack]

Schedule Overview

Time - Canada ET	Day 1 - Wednesday - May 13	Day 2 - Thursday - May 14	Day 3 - Friday - May 15
11:15-11:30am	Welcome to CRV		
11:30-12:00pm	AI Keynote: Pascal Poupart	CRV Keynote #1: Jon How	CRV Keynote #2: Simon Lucey
12:00-12:30pm			
12:30-1:00pm	Break		
1:00-1:30pm	Session 1 - Q/A (Slack)	Break	Break
1:30-2:00pm		AI Keynote: G. Carenini	AI Keynote:C. Szepesvári
2:00-2:30pm	Break		
2:30-3:00pm	Session 2 - Q/A (Slack)	Break	Break
3:00-3:30pm		Session 3 - Q/A (Slack)	Session 5 - Q/A (Slack)
3:30-4:00pm	Break		
4:00-4:30pm	CIPPRS:Annual General Meeting	Break	Break
4:30-5:00pm		Session 4 - Q/A (Slack)	Session 6 - Q/A (Slack)
5:00-5:30pm			
5:30-6:00pm			CVR Closing Remarks
6:00-6:30pm			
6:30-7:00pm			
7:00-7:30pm		Online Social	
7:30-8:00pm		CRV - Awards	
8:00-8:30pm		Social cont’d	
8:30-9:00pm		Social cont’d	

Color Scheme

CRV Conference Items	🟦	🟧	Canada AI Conference Items	🟨

[Link to Canada AI 2020 Program]

Day 1 (Wednesday) - May 13, 2020

11:15 - 11:30am

CRV Welcome **(via Youtube streaming [link] [slides] [general chair welcome])**
Liam Paull and Michael S. Brown
CRV 2020 Program Chairs

11:30am - 12:30pm

AI Keynote #1 - Pascal Poupart **(via Youtube streaming [link])**

1:00pm - 2:00pm

Session 1 (via Slack [link])
Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions.

Symposium Talks
Gregor Miller, Google (USA)
Negar Rostamzadeh, Google Brain

Conference Papers

12	A Non-contact Method for Extracting Heart and Respiration Rates
27	TimeConvNets: A Deep Time Windowed Convolution Neural Network Design for Real-time Video Facial Expression Recognition
3	Simultaneous Demosaicing and Chromatic Aberration Correction through Spectral Reconstruction
13	Tree bark re-identification using a deep-learning feature descriptor
16	Image classification by Distortion-Free Graph Embedding and KNN-Random forest

2:00pm - 2:30pm

Coffee Break (via Slack [link])

2:30-pm - 3:30pm

Session 2 (via Slack [link])
Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions.

Symposium Talks
Francois Pomerleau, Laval
Igor Gilitschenski, MIT

Conference Papers

15	Towards Richer 3D Reference Maps in Urban Scenes
34	Geometry-Guided Adaptation for Road Segmentation
37	Unsupervised depth prediction from monocular sequences: Improving performances through instance segmentation
10	Recognizing and Tracking High-Level, Human-Meaningful Navigation Features of Occupancy Grid Maps
56	Depth Prediction for Monocular Direct Visual Odometry

4:00pm - 5:00pm

Canadian Image Processing and Pattern Recognition Society (CIPPRS)
Annual General Meeting
(Michael S. Brown, Liam Paull, and Steve Waslander Moderators)
Open to all CRV participants [Zoom link] (for pwd see the #cipprs-agm slack channel description)
[Minute of meeting]

Day 2 (Thursday) - May 14, 2020

11:30am - 12:45pm **CRV Keynote #1 - Prof. Jon How (MIT)** **(via Youtube streaming [link])**
Chair - Liam Paull

1:30pm - 2:30pm **AI Keynote #2 - Giuseppe Carenini (via Youtube streaming [link])**

3:00pm - 4:00pm **Session 3 (via Slack [link])**

Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions.

Symposium Talks

Minglun Gong, University of Guelph

Conference Papers

31 Pre-trained CNNs as Visual Feature Extractors: A Broad Evaluation
21 Domain Generalization via Universal Non-volume Preserving Approach
32 Automatic Classification of Woodcuts and Copperplate Engravings
22 Histological Image Classification: Deep Features or Transfer Learning?
6 MASC-Net: Multi-scale Anisotropic Sparse Convolutional Network \\for Sparse Depth Densification
28 Domain Generalization via Optical Flow: Training a CNN in a Low-Quality Simulation to Detect Obstacles in the Real World

4:00pm - 4:30pm **Coffee Break (via Slack [link])**

4:30-pm - 5:30pm **Session 4 (via Slack [link])**

Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions.

Symposium Talks

Animesh Garg - University of Toronto
Matthew Walter, Toyota Technological Institute at Chicago

Conference Papers

55 PVF-NET: Point & Voxel Fusion 3D Object Detection Framework for Point Cloud
47 Leveraging Temporal Data for Automatic Labelling of Static Vehicles
38 In-Time 3D Reconstruction and Instance Segmentation from Monocular Sensor Data
4 Domain Adaptation in Crowd Counting
57 Gas Prices of America: The Machine-Augmented Crowd-Sourcing Era

7:00 - 9:00pm **Online Conference Banquet (via Zoom + Breakout Rooms [link])**
(for pwd see #conference-banquet slack channel description)
7:30pm - Short Awards Ceremony **(will be streamed to Youtube [link])**

- CIPPRS Dissertation Awards
- CRV Best Robotics and Vision Paper Awards

Day 3 (Friday) - May 15, 2020

11:30am - 12:45pm **CRV Computer Vision Keynote - Prof. Simon Lucey (CMU)** **(via Youtube streaming [link])**
Chair - Michael S. Brown

1:30pm - 2:30pm **AI Keynote #3 - Csaba Szepesvari** **(via Youtube streaming [link])**

3:00pm - 4:00pm **Session 5** **(via Slack [link])**

Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions.

Symposium Talks
Sajad Saeedi - Ryerson
Xiaoming Liu, Michigan State University

Conference Papers

33	Gradient-Based Auto-Exposure Control Applied to a Self-Driving Car
18	CVNodes: A Visual Programming Paradigm for Developing Computer Vision Algorithms
24	Real-time Motion Planning for Robotic Teleoperation Using Dynamic-goal Deep Reinforcement Learning
26	Towards End-to-end Learning of Visual Inertial Odometry with an EKF
30	Evaluation of Skid-Steering Kinematic Models for Subartic Environments

4:00pm - 4:30pm **Coffee Break** **(via Slack [link])**

4:30-pm - 5:30pm **Session 6** **(via Slack [link])**

Videos talks and poster PDF available online before the conference. This session will allow an interactive Slack forum to ask the authors questions about their paper. At least one author per paper will be available on the Slack channel to answer questions..

Symposium Talks
Helge Rhodin, UBC
Ismail Benayed, ETS Montreal

Conference Papers

29	Single-Stage End-to-End Temporal Activity Detection in Untrimmed Videos
20	Depth from Defocus on a Transmissive Diffraction Mask-based Sensor
35	Differentiable Mask for Pruning Convolutional and Recurrent Networks
36	SpotNet: Self-Attention Multi-Task Network for Object Detection
42	It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity

5:30pm - 5:45pm **CRV Closing Comments** **(via Youtube streaming [link])**
Liam Paull and Michael S. Brown

Welcome Message
17th Conference on Computer and Robot Vision - CRV 2020

Welcome virtually to Ottawa, Ontario, and the Seventeenth Conference on Computer and Robot Vision (CRV 2020)! This conference series provides a high-quality forum for the international and Canadian computer and robot vision communities to share their work. Like many meetings in 2020, the Covid-19 situation has forced up to hold the meeting virtually.

Our conference is sponsored by the Canadian Image Processing and Pattern Recognition Society / Association Canadienne de Traitement d’Images et de Reconnaissance des Formes (CIPPRS/ACTIRF). CIPPRS/ACTIRF is a special interest group of the Canadian Information Processing Society (CIPS) and is the official Canadian member of the governing board of the International Association for Pattern Recognition (IAPR). The goal of CIPPRS/ACTIRF is to promote research and development activities in Computer Vision, Robot Vision, Image Processing, Medical Imaging and Pattern Recognition.
The papers here have each been peer-reviewed by at least three reviewers drawn from a 68 member Program Committee composed of internationally recognized computer and robot vision researchers. We wish to thank the Program Committee for the careful and professional reviews they provided, despite a short reviewing period.

This year we received a total of 61 submissions, from which 31 papers were accepted. Given the virtual nature of the conference this year, all papers were presented via a pre-recorded presentation and a poster. A Q/A session was held for each paper session where authors could engage with participants to discuss the papers. The CRV conference proceedings are published by the IEEE Computer Society Conference Publishing Services (CPS) and will be available online via IEEE Xplore. The proceedings are additionally indexed through the Inspec indexing service.

We are delighted to welcome two outstanding keynote speakers this year: Dr. Jon How from the Massachusetts Institute of Technology and Dr. Simon Lucey from Carnegie Mellon University. Also, we are continuing the invited Symposium Speaker Series this year, featuring 12 leading computer vision and robotics researchers from Canada and the United States.
CRV recognizes stellar research and contributions through multiple awards, including the CIPPRS Lifetime Achievement Award - Research, the CIPPRS Lifetime Achievement Award - Community, the Best Paper Award - Computer Vision, and the Best Paper Award - Robot Vision. Winners will be announced at the virtual conference awards social.

Many people collaborated to organize this meeting. We gratefully acknowledge and thank Marina Sokolova from the University of Ottawa and Chris Drummond at NRC who acted as general chairs for the entire AI-CRV conference. We also thank Michael Jenkin of York University for serving on the Awards Committee.

We thank CIPPRS Treasurer Steven Waslander of the University of Toronto, CIPPRS Secretary Jim Little of the University of British Columbia, and CIPPRS President Michael Jenkin of York University for their ongoing guidance and support.
Finally, and most importantly, many thanks to the authors who contributed exciting research to CRV 2020. We thank the attendees of CRV for their patience and willingness to participate in a virtual meeting. We hope to meet everyone in person next year!

Michael S. Brown, *York University*
Liam Paull, *University of Montreal*
CRV 2020 Program Co-Chairs

CRV Keynote #1

Speaker: Jonathan How (Massachusetts Institute of Technology)
Talk Title: Navigation and Mapping for Robot Teams in Uncertain Environments
[youtube stream link]

Abstract
Many robotic tasks require robot teams to autonomously operate in challenging, partially observable, dynamic environments with limited field-of-view sensors. In such scenarios, individual robots need to be able to plan/execute safe paths on short timescales to avoid imminent collisions. Robots can leverage high-level semantic descriptions of the environment to plan beyond their immediate sensing horizon. For mapping on longer timescales, the agents must also be able to align and fuse imperfect and partial observations to construct a consistent and unified representation of the environment. Furthermore, these tasks must be done autonomously onboard, which typically adds significant complexity to the system. This talk will highlight three recently developed solutions to these challenges that have been implemented to (1) robustly plan paths and demonstrate high-speed agile flight of a quadrotor in unknown, cluttered environments; and (2) plan beyond the line-of-sight by utilizing the learned context within the local vicinity, with applications in last-mile delivery. We further present a multi-way data association algorithm to correctly synchronize partial and noisy representations and fuse maps acquired by (single or multiple) robots, showcased on a simultaneous localization and mapping (SLAM) application.
Bio
Jonathan P. How is the Richard C. Maclaurin Professor of Aeronautics and Astronautics at the Massachusetts Institute of Technology. He received a B.A.Sc. from the University of Toronto in 1987, and his S.M. and Ph.D. in Aeronautics and Astronautics from MIT in 1990 and 1993, respectively. Prior to joining MIT in 2000, he was an assistant professor in the Department of Aeronautics and Astronautics at Stanford University. He was the editor-in-chief of the IEEE Control Systems Magazine (2015-19) and was elected to the Board of Governors of the IEEE Control System Society (CSS) in 2019. His research focuses on robust planning and learning under uncertainty with an emphasis on multiagent systems. His work has been recognized with multiple awards, including the 2020 AIAA Intelligent Systems Award, the 2002 Institute of Navigation Burka Award, the 2011 IFAC Automatica award for best applications paper, the 2015 AeroLion Technologies Outstanding Paper Award for Unmanned Systems, the 2015 IEEE Control Systems Society Video Clip Contest, the IROS Best Paper Award on Cognitive Robotics (2017 and 2019) and three AIAA Best Paper in Conference Awards (2011-2013). He was awarded the Air Force Commander's Public Service Award (2017). He is a Fellow of IEEE and AIAA.

CRV Keynote #2
Speaker: Simon Lucey (Carnegie Mellon University)
Talk Title: Geometric reasoning in machine vision with using only 2D supervision
[youtube stream link]

Abstract
Machine vision has made tremendous progress over the last decade with respect to perception. Much of this progress can be attributed to two factors: the ability of deep neural networks (DNNs) to reliably learn a direct relationship between images and labels; and access to a plentiful number of images with corresponding labels. We often refer to these labels as supervision – because they directly supervise what we want the vision algorithm to predict when presented with an input image. 2D labels are relatively easy for the computer vision community to come by: human annotators are hired to draw – literally through mouse clicks on a computer screen – boxes, points, or regions for a few cents per image. But how to obtain 3D labels is an open problem for the robotics and vision community. Rendering through computer generated imagery (CGI) is problematic, since the synthetic images seldom match the appearance and geometry of the objects we encounter in the real world. Hand annotation by humans is preferable, but current strategies rely on the tedious process of associating the natural images with a corresponding external 3D shape – something we refer to as “3D supervision”. In this talk, I will discuss recent efforts my group has been taking to train a geometric reasoning system using solely 2D supervision. By inferring the 3D shape solely from 2D labels we can ensure that all geometric variation in the training images is learned. This innovation sets the ground work for training the next generation of reliable geometric reasoning AIs needed to solve emerging needs in: autonomous transport, disaster relief, and endangered species preservation.

Bio
Simon Lucey (Ph.D.) is an associate research professor within the Robotics Institute at Carnegie Mellon University, where he is part of the Computer Vision Group, and leader of the CI2CV Laboratory. Since 2017, he is also a principal scientist at Argo AI. Before this he was an Australian Research Council Future Fellow at the CSIRO (Australia's premiere government science organization) for 5 years. Simon’s research interests span computer vision, robotics, and machine learning. He enjoys drawing inspiration from vision researchers of the past to attempt to unlock computational and mathematic models that underly the processes of visual perception.

Symposium Speakers

Invited speaker: Gregor Miller, Google (USA)
Talk: OpenVL, a developer-friendly abstraction of computer vision
[video]

Abstract: Computer vision is a complicated topic and the fruits of our efforts often gets included in libraries such as OpenCV or as open source projects released by university labs or by companies. Here, the presentation of our work is often aimed at other researchers or those who are well versed in computer vision. However, to encourage widespread or faster adoption of these technologies, it is important that they be accessible to those not necessarily expert in the field. This talk is about the principles that underly our OpenVL framework, and that guided us to create a computer vision platform that was usable by mainstream developers with no specific expertise in computer vision. My hope is that this will be inspirational and potentially guide others in how to present their research or products more effectively to their targeted users.

Invited speaker: Negar Rostamzadeh, Google Brain
Talk Title: On Label Efficient Machine Perception
[video]

Abstract: Deep learning methods often require large amounts of labelled data which can be impractical or expensive to acquire. My talk will cover four categories of work in minimizing the required labeling effort without sacrificing performance: (i) algorithms requiring less annotations per instance (point-level annotation for object counting and instance level semantic segmentation) (ii) using active learning to label the most informative part of the data (iii) domain adaptation when source domain annotation is cheaper and easier to acquire and (iv) multi-modal learning for the task of few-shot classification. I will briefly touch on the first 3 categories while discussing multi-modal learning in depth.

Invited speaker: Francois Pomerleau, Laval University
Talk Title: From subterranean to subarctic autonomous exploration
[video]

Abstract: Autonomous navigation algorithms are now pushed to the extreme with robotic deployment happening in harsh experimental sites. Gaining robustness against dynamic and unstructured environments along with managing unforeseen robot dynamics is mandatory to reach a larger spectrum of autonomous navigation applications. This talk will give lessons learned from such difficult environments, namely a subterranean urban circuit and subarctic forest. The first part of the presentation will present results from our latest participation to the DARPA Subterranean (SubT) Challenge, for which our lab was the only Canadian participant. In February 2020, we supported the deployment of nine robots in a disaffected nuclear power plant in Elma, Washington. Our team, named CTU-CRAS-NORLAB, finished third against leading research laboratories in the world. The second part of the presentation will present our latest research results on lidar-based mapping in winter conditions.

Invited Speaker: Igor Gilitschenski, MIT
Talk Title: Robust Perception for Autonomous Systems
[video]

Abstract: In recent years we have seen an exploding interest in the real-world deployment of autonomous systems, such as autonomous drones or autonomous ground vehicles. This interest was sparked by major advancements in robot perception, planning, and control. However, robust operation in the “wild” remains a challenging goal. Correct consideration of the broad variety of real-world conditions cannot be achieved by merely optimizing algorithms that have been shown to work well in controlled environments. In this talk, I will focus on robust perception for autonomous systems. First, I will discuss the challenges involved in handling dynamic and changing environments in Visual-SLAM. Second, I will discuss autonomous vehicle navigation concepts that do not rely on highly detailed maps or sufficiently good weather for localization. Finally, I will discuss the role of perception in interactive autonomy. Particularly, I will focus on the use of simulators for edge-case generation, learning, and real-world transfer of deep driving policies.

Symposium Speakers cont'

Invited Speaker: Minglun Gong, University of Guelph
Talk Title: Novel network architectures for arbitrary image style transfer
[video]

Abstract: Style transfer has been an important topic in both Computer Vision and Graphics. Since the pioneer work of Gatys et al. demonstrated the power of stylization through optimization in deep feature space, a number of approaches have been developed for real-time arbitrary style transfer. However, even the state-of-the-art approaches may generate insufficiently stylized results under challenging cases. Two novel network architectures are discussed in this talk for addressing the issues and delivering better performances. We first observe that only considering features in the input style image for the global deep feature statistic matching or local patch swap may not always ensure a satisfactory style transfer. Hence, we propose a novel transfer framework that aims to jointly analyze and better align exchangeable features extracted from the content and style image pair. This allows the style features used for transfer to be more compatible with content information in the content image, leading to more structured stylization results. Another observation is that existing methods try to generate stylized result in a single shot, making it difficult to satisfy constraints on semantic structures in the content images and style patterns in the style images. Inspired by the works on error-correction, we propose a self-correcting model to predict what is wrong with the current stylization and refine it accordingly in an iterative manner. For each refinement, we transit the error features across both the spatial and scale domains and invert the processed features into a residual image.

Invited Speaker: Animesh Garg, University of Toronto
Talk Title: Generalizable Autonomy for Robot Manipulation
[video]

Abstract: Data-driven methods in Robotics circumvent hand-tuned feature engineering, albeit lack guarantees and often incur a massive computational expense. My research aims to bridge this gap and enable generalizable imitation for robot autonomy. We need to build systems that can capture semantic task structures that promote sample efficiency and can generalize to new task instances across visual, dynamical or semantic variations. And this involves designing algorithms that unify learning with perception, control and planning. In this talk, I will how inductive biases and priors help with Generalizable Autonomy. First I will talk about choice of action representations in RL and imitation from ensembles of suboptimal supervisors. Then I will talk about latent variable models in self-supervised learning. Finally I will talk about meta-learning for multi-task learning and data gather in robotics.

Invited Speaker: Matthew Walter, Toyota Technological Institute at Chicago
Talk Title: Natural Language Learning for Human-Robot Collaboration
[video]

Abstract:
Natural language promises an efficient and flexible means for humans to communicate with robots, whether they are assisting the physically or cognitively impaired, or performing disaster mitigation tasks as our surrogates. Recent advancements have given rise to robots that are able to interpret natural language commands that direct object manipulation and spatial navigation. However, most methods require prior knowledge of the metric and semantic properties of the objects and places that comprise the robot's environment.

In this talk, I will present our work that enables robots to successfully follow natural language navigation instructions within novel, unknown environments. I will first describe a method that treats language as a sensor, exploiting information implicit and

explicit in the user's command to learn distributions over the latent spatial and semantic properties of the environment and over the robot's intended behavior. The method then learns a belief space policy that reasons over these distributions to identify suitable navigation actions. In the second part of the talk, I will present an alternative formulation that represents language understanding as a multi-view sequence-to-sequence learning problem. I will introduce an alignment-based neural encoder-decoder architecture that translates free-form instructions to action sequences based on images of the observable world. Unlike previous methods, this architecture uses no specialized linguistic resources and can be trained in a weakly supervised, end-to-end fashion, which allows for generalization to new domains. Time permitting, I will then describe how we can effectively invert this model to enable robots to generate natural language utterances. I will evaluate the efficacy of these methods on a combination of benchmark navigation datasets and through demonstrations on a voice-commandable wheelchair.

Symposium Speakers cont'

Invited Speaker: Sajad Saeedi, Ryerson
Talk Title: Bringing Computer Vision to Robotics
[video]

Abstract: The technological advancements in machine learning and robotics have moved the presence of robots from exclusively in manufacturing facilities, into households where they are executing simple tasks such as vacuuming and lawn mowing. To extend the capabilities of these systems, robots need to advance beyond just reporting ‘what’ is ‘where’ in an image to developing spatial AI systems, necessary to interact usefully with their unstructured and dynamic environment. Therefore, there is an urgent need for novel perception and control systems that can deal with many real-world constraints such as limited resources, dynamic objects, and uncertain information. In this talk, several recent projects related to robotics and machine perception are presented. Recent developments such as accelerated inference on focal-plane sensor-processor arrays are introduced. These developments have significant economic and scientific impacts on our society and will open up new possibilities for real-time and reliable utilization of AI and robotics in real-world and dynamic environments. At the end of the talk, future research directions will be outlined. The main goal for future research will be developing reliable, high-speed, and low-power perception systems that can be deployed in real-world applications. It has been hypothesized that while machine learning algorithms will give us the required reliability, data processing in the focal plane will help us to achieve the desired energy consumption and run-time speed limits. Reliable, fast, and low-power computation for scene understanding and spatial awareness will be of great interest not only to the robotics community, but also other fields, such as the Internet of Things (IoT), Industry 4.0, privacy-aware devices, and networked-visual devices. These research directions will help entrepreneurs and academic researchers to identify new opportunities in machine learning and its application in robotics in real-world and dynamic environments.

Invited Speaker: Xiaoming Liu, Michigan State University
Talk Title: Monocular Vision-based 3D Perception for Autonomous Driving
[video]

Abstract: Perception in the 3D world is an essential requirement for autonomous driving. Most existing algorithms rely on depth sensors such as LiDAR for 3D perception. In this talk, we will present our recent efforts on 3D perception based solely on monocular RGB images. First, we describe a unified 3D-RPN for 3D detection of vehicles, pedestrians, and bicycles. Secondly, a novel inverse graphics framework is designed to model the 3D shape and albedo for generic objects, while fitting these models to an image leads to 3D reconstruction of objects. Finally, we will also briefly present the low-level, and high-level computer vision efforts for autonomous driving at MSU, including LiDAR and RGB fusion, depth estimation, and semantic segmentation forecasting.

Invited Speaker: Helge Rhodin, UBC
Talk Title: Computer Vision for Interactive Computer Graphics
[video]

Abstract: I work at the intersection of computer graphics and machine learning-based computer vision. I will be talking about my works on human and animal motion capture and their impact on gesture-driven character animation, VR telepresence, and automation in neuroscience. Moreover, I will outline my ongoing work on replacing hand-crafted CG and CV models with learned ones using self-supervision through multi-view and other geometric and physical constraints, including gravity.

Invited Speaker: Ismail Benayed, ETS Montreal
Talk Title: Constrained Deep Networks
[video]

Abstract: Embedding constraints on the outputs of deep networks has wide applicability in learning, vision and medical imaging. For instance, in weakly supervised learning, constraints can mitigate the lack of full and laborious annotations, leveraging unlabeled data and guiding training with domain-specific knowledge. Also, adversarial robustness, which currently attracts substantial interest in the field, amounts to imposing constraints on network outputs. In this talk, I will discuss some recent developments in those research directions, emphasize how more attention should be paid to optimization methods, and include various illustrations, applications and experimental results.

SESSION 1 PAPERS

(Paper 12)
Title: A Non-contact Method for Extracting Heart and Respiration Rates
Christian Hessler, Mohamed Abouelenien, and Mihai Burzo (University of Michigan)
[paper][poster][video][slack channel]

Abstract
Physiological signals provide a reliable method to identify the physical and mental state of a person at any given point in time. Accordingly, there are a myriad of techniques used to extract physiological signals from the human body. However, these techniques often require direct contact with the body. This demands the cooperation of the individual as well as the human effort required to connect devices and collect measurements.

In this paper, we propose reliable, non-contact based methods for extracting respiration rate and heart rate from thermal images using a large dataset of human thermal recordings. These methods leverage a combination of image and signal processing techniques in order to extract and filter physiological signals from the thermal domain. Our results evidently show that features extracted from thermal images highly correlate with the ground truth measurements as well as indicate the feasibility of developing non-contact based methods to extract physiological signals.

(Paper 27)
Title: TimeConvNets: A Deep Time Windowed Convolution Neural Network Design for Real-time Video Facial Expression Recognition
James Lee, Alexander Wong (University of Waterloo)
[paper][poster][video][slack channel]

Abstract
A core challenge faced by the majority of individuals with Autism Spectrum Disorder (ASD) is an impaired ability to infer other people's emotions based on their facial expressions. With significant recent advances in machine learning, one potential approach to leveraging technology to assist such individuals to better recognize facial expressions and reduce the risk of possible loneliness and depression due to social isolation is the design of computer vision-driven facial expression recognition systems. Motivated by this social need as well as the low latency requirement of such systems, this study explores a novel deep time windowed convolutional neural network design (TimeConvNets) for the purpose of real-time video facial expression recognition. More specifically, we explore an efficient convolutional deep neural network design for spatiotemporal encoding of time windowed video frame sub-sequences and study the respective balance between speed and accuracy. Furthermore, to evaluate the proposed TimeConvNet design, we introduce a more difficult dataset called BigFaceX, composed of a modified aggregation of the extended Cohn-Kanade (CK+), BAUM-1, and the eNTERFACE public datasets. Different variants of the proposed TimeConvNet design with different backbone network architectures were evaluated using BigFaceX alongside other network designs for capturing spatiotemporal information, and experimental results demonstrate that TimeConvNets can better capture the transient nuances of facial expressions and boost classification accuracy while maintaining a low inference time.

(Paper 3)
Title: Simultaneous Demosaicing and Chromatic Aberration Correction through Spectral Reconstruction
Bernard Llanos, Herb Yang (University of Alberta)
[paper][poster][video][slack channel]

Abstract
We present an algorithm for simultaneously demosaicing digital images, and correcting chromatic aberration, that operates in terms of spectral bands. Chromatic aberration depends on both the camera's optical system, and on the spectral characteristics of the light entering the camera. Previous works on calibrating chromatic aberration produce models of chromatic aberration that assume fixed relationships between image channels, an assumption that is only valid when the image channels capture narrow regions of the electromagnetic spectrum. When the camera has wideband channels, as is the case for conventional trichromatic (RGB) cameras, the aberration observed both within and between channels can only be accurately predicted given the spectral irradiance of the theoretical, aberration-free image. For an RGB camera, we use bandpass-filtered light to calibrate its chromatic aberration in terms of image position and light wavelength. Inspired by literature on reconstructing spectral images from RGB images, we then correct images for chromatic aberration by estimating

aberration-free, spectral images. As we model within-channel chromatic aberration, our reconstructed images are sharper than those obtained by calibrated warping of color channels, yet we avoid artifacts commonly produced by explicit deblurring algorithms.

SESSION 1 papers cont'

(Paper 13)
Title: Tree Bark Re-identification Using a Deep-Learning Feature Descriptor
Martin Robert, Philippe Giguère, Patrick Dallaire (Laval University)
[paper][poster][video][slack channel]

Abstract

The ability to visually re-identify objects is a fundamental capability in vision systems. Oftentimes, it relies on collections of visual signatures based on descriptors, such as SIFT or SURF. However, these traditional descriptors were designed for a certain domain of surface appearances and geometries (limited relief). Consequently, highly-textured surfaces such as tree bark pose a challenge to them. In turns, this makes it more difficult to use trees as identifiable landmarks for navigational purposes (robotics) or to track felled lumber along a supply chain (logistics). We thus propose to use data-driven descriptors trained on bark images for tree surface re-identification. To this effect, we collected a large dataset containing 2,400 bark images with strong illumination changes, annotated by surface and with the ability to pixel-align them. We used this dataset to sample from more than 2 million 64 x 64 pixel patches to train our novel local descriptors DeepBark and SqueezeBark. Our DeepBark method has shown a clear advantage against the hand-crafted descriptors SIFT and SURF. For instance, we demonstrated that DeepBark can reach a mAP of 87.2% in a database of 7,900 images with only 11 relevant images. Our work thus suggests that re-identifying tree surfaces in challenging illuminations context is possible. We also make public our dataset, which can be used to benchmark surface re-identification techniques.

(Paper 16)
Title: Image Classification by Distortion-Free Graph Embedding and KNN-Random Forest
Askhat Temir, Kamalkhan Artykbayev, Fatih M. Demirci (Nazarbayev University)
[paper][poster][video][slack channel]

Abstract

Image classification algorithms play an important role in various computer vision problems such as object tracking, image labeling, and object segmentation. A number of methodologies have been proposed to tackle this problem. One of the possible approaches employed extensively in the literature is to represent an image as a graph based on its hand- crafted features. However, recent advancements in deep neural networks have shown their ability to learn more discriminative and representative features. Therefore, the deep features have become considerable alternatives of hand-crafted ones. In this paper, we propose a novel framework based on distortion-free graph embedding using deep features and KNN-Random forest. Our method outperforms the state-of-the-art graph embedding-based image classification approach for the task of image classification. Particularly, the proposed framework obtains 97.5% top - 1 image classification accuracy for the ImageNet dataset for 5 classes and 93.3% for 10 classes.

SESSION 2 PAPERS

(Paper 15)

Title: Towards Richer 3D Reference Maps in Urban Scenes

Maya Antoun, Daniel Asmar, Rema Daher (American University of Beirut)

[paper][poster][video][slack channel]

Abstract

Meaningful 3D maps have a lot to offer to the design of safe and intelligent transportation systems. To doso, street content such as cars, traffic lights and signs need to be segmented in their 3D form, and accurately localized in a map. This paper proposes a first step towards producing 3Drich urban maps in which cars are segmented from a point cloud, and rendered in their true form on a metric map.Our system is based on the integration of stereo SLAM for point cloud extraction, Fast-RCNN for car detection, shape completion, meshing, and optimization of camera pose based on the detected cars. We test our system on the KITTI dataset and produce very realistic maps. As a second contribution, we assess the effect of including the detected cars as objects in a semanticSimultaneous Localization and Mapping (SLAM) pipeline, and demonstrate the potential for improved localization.

(Paper 34)

Title: Geometry-Guided Adaptation for Road Segmentation

Gong Cheng (York University), Yue Wang (Dalian University of Technology), Yiming Qian (York University), James Elder (York University)

[paper][poster][video][slack channel]

Abstract

We propose a novel adaptation method for generalizing road segmentation to novel weather, lighting or viewing geometries. The method assumes a source domain consisting of an ensemble of labeled training datasets and an unlabeled target test dataset that deviates substantially from the training ensemble. The training dataset is used to compile a geometry-anchored prior over the road pixel locations and to train a fully-convolutional network road segmentation system. At inference, a probabilistic Houghing method is used to detect line intersections in the test image and thereby estimate the vanishing point of the road, thus anchoring the learned geometric prior. This prior is then used to extract high confidence road and background regions which serve as surrogate ground truth to adapt the network to the target domain. Leave-one-out evaluation across five diverse road segmentation datasets demonstrates substantial improvement in generalization across changes in viewing geometry and weather conditions, yielding results that are on average comparable and in some cases superior to a more complex GAN-based domain adaptation approach. These results demonstrate the potential for classical computer vision methods to guide adaptation of supervised machine learning algorithms, leading to improved generalization across domains.

(Paper 37)

Title: Unsupervised depth prediction from monocular sequences: Improving Performances Through Instance Segmentation

Ambroise Moreau, Thierry Dutoit, Matei Mancas (University of Mons)

[paper][poster][video][slack channel]

Abstract

Depth is a valuable piece of information for robots and autonomous vehicles. Indeed, it enables them to move in space and avoid obstacles. Nevertheless, depth alone is not enough to let them interact with their surroundings. They also need to locate the different objects that are present in their environment. In this paper, we propose a deep learning model that solves unsupervised monocular depth estimation and supervised instance segmentation at the same time with a common architecture. The first task is solved through novel view synthesis while the second is solved by minimising an embedding loss function. Our approach is motivated by the idea that knowing where objects are in the scene could improve the depth estimation of unsupervised monocular depth models. We tested our architecture on two datasets, Kitti and Cityscapes and reached state-of-the-art depth estimation results while solving a second task.

Session 2 cont'

(Paper 10)

Title: Recognizing and Tracking High-Level, Human-Meaningful Navigation Features of Occupancy Grid Maps

Payam Nikdel, Mo Chen, Richard Vaughan (Simon Fraser University)

[paper][poster][video][slack channel]

Abstract

This paper describes a system whereby a robot detects and track human-meaningful navigational cues as it navigates in an indoor environment. It is intended as the sensor front-end for a mobile robot system that can communicate its navigational context with human users. From simulated LiDAR scan data we construct a set of 2D occupancy grid bitmaps, then hand-label these with human-scale navigational features such as closed doors, open corridors and intersections. We train a Convolutional Neural Network (CNN) to recognize these features on input bitmaps. In our demonstration system, these features are detected at every time step then passed to a tracking module that does frame-to-frame data association to improve detection accuracy and identify stable unique features. We evaluate the system in both simulation and the real world. We compare the performance of using input occupancy grids obtained directly from LiDAR data, or incrementally constructed with SLAM, and their combination.

(Paper 56)

Title: Depth Prediction for Monocular Direct Visual Odometry

Ran Cheng (McGill University), Christopher Agia (University of Toronto), David Meger (McGill University), Gregory Dudek (McGill University)

[paper][poster][video][slack channel]

Abstract

Depth prediction from monocular images with deep CNNs is a topic of increasing interest to the community. Advances have lead to models capable of predicting disparity maps with consistent scale, which are an acceptable prior for gradient-based direct methods. With this in consideration, we exploit depth prediction as a candidate prior for the course initialization, tracking, and marginalization steps of the direct visual odometry system, enabling the second-order optimizer to converge faster into a precise global minimum. In addition, the given depth prior supports large baseline stereo scenarios, maintaining robust pose estimations against challenging motion states like in-place rotation. We further refine our pose estimation with semi-online loop closure. The experiments on KITTI demonstrate that our proposed method achieves state-of-the-art performance compared to both traditional direct visual odometry and learning-based counterparts.

SESSION 3 PAPERS

(Paper 31)

Title: Pre-trained CNNs as Visual Feature Extractors: A Broad Evaluation

Andrew Holliday (McGill University)

[paper][poster][video][slack channel]

Abstract

In this work, we perform a wide-ranging evaluation of Convolutional Neural Networks (CNNs) as feature extractors for matching visual features under large changes in appearance, perspective, and visual scale. Our evaluation covers 82 different layers from twelve different CNN architectures belonging to four families: AlexNets, VGG Nets, ResNets, and DenseNets. To our knowledge, this is the most comprehensive analysis of its kind in the literature. We find that the intermediate layers of DenseNets serve as the best feature extractors overall, providing the best overall trade-off of robustness to feature size. Moreover, we find that for each network, the later-intermediate layers provide the best performance, regardless of the total number of layers in the network.

(Paper 21)

Title: Domain Generalization via Universal Non-volume Preserving Approach

Thanh-Dat Truong (University of Arkansas) , Chi Nhan Duong (Concordia University), Khoa Luu (University of Arkansas), Minh-Triet Tran (University of Science, VNU-HCM) , Ngan Le (University of Arkansas)

[paper][poster][video][slack channel]

Abstract

Recognition across domains has recently become an active topic in the research community. However, it has been largely overlooked in the problem of recognition in new unseen domains. Under this condition, the delivered deep network models are unable to be updated, adapted, or fine-tuned. Therefore, recent deep learning techniques, such as domain adaptation, feature transferring, and fine-tuning, cannot be applied. This paper presents a novel approach to the problem of domain generalization in the context of deep learning. The proposed method is evaluated on different datasets in various problems, i.e. (i) digit recognition on MNIST, SVHN, and MNIST-M, (ii) face recognition on Extended Yale-B, CMU-PIE and CMU-MPIE, and (iii) pedestrian recognition on RGB and Thermal image datasets. The experimental results show that our proposed method consistently improves performance accuracy. It can also be easily incorporated with any other CNN frameworks within an end-to-end deep network design for object detection and recognition problems to improve their performance.

(Paper 32)

Title: Automatic Classification of Woodcuts and Copperplate Engravings

Michael J. Cormier, Sarah Park, Lauren Beck (Mount Allison University)

[paper][poster][video][slack channel]

Abstract

In this paper we introduce two related Bayesian approaches to distinguishing woodcut and copperplate engraving illustrations. Our approach uses simple methods grounded in the characteristics of the techniques to achieve high performance and excellent interpretability. Specifically, we use the distributions of image gradient magnitudes in illustrations to capture differences in the use of lines in each artistic technique, and the output of a probability for each class aides in interpretability. Our experiments demonstrate high accuracy and robustness with respect to parameters for both algorithms.

(Paper 22)
Title: Histological Image Classification using Deep Features and Transfer Learning
Sadiq Alinsaif, Jochen Lang (University of Ottawa)
[paper][poster][video][slack channel]

Abstract
A major challenge in the automatic classification of histopathological images is the limited amount of data available. Supervised learning techniques can not be applied without some adjustment. We compare two common techniques to deal with limited domain data: using deep features and fine-tuning convolutional neural networks (CNN). We examine the following state-of-art CNN models: SqueezeNet-v1.1, MobileNet-v2, ResNet-18, and DenseNet-201. We demonstrate that using feature vectors that are extracted from one of the four CNN models with a classical support vector machine (SVM) for training and testing can lead to higher accuracy on publicly available datasets: Warwick-QU, Epistroma, BreaKHis, multi-class Kather, than previously published results. Similar results can be obtained with fully fine-tuning the aforementioned CNN models. We also study the effectiveness of block-wise fine-tuning of two models (i.e., SqueezeNet-v1.1 and ResNet-18) and show that it is not necessary to fully fine-tune leading to savings in time and space.

(Paper 6)
Title: MASC-Net: Multi-scale Anisotropic Sparse Convolutional Network for Sparse Depth Densification
Seungchul Ryu, Ji-Ho Cho, Neeth Kunnath (Airy 3D Inc)
[paper][poster][video][slack channel]

Abstract
Irregular sparse depth densification has attracted a significant amount of recent interests in computer vision, robotics, autonomous driving and augmented reality applications. Low-cost 3D sensors available on the market today produce sparse and irregular depth data resulting in sometimes inaccurate or inconsistent results. This paper proposes a novel approach for sparse depth densification called the Multi-scale Anisotropic Sparse Convolutional Network (MASC-Net). Conventional sparse convolutional approaches face two main challenges in deep neural networks: 1) blurry depth and fattening artifacts near object boundaries due to isotropic validity masks, and 2) the validity mask mismatch problem across different layers. To address these problems, we propose anisotropic sparse convolutional layers based on spatially varying validity masks with guidance features. We also propose Validity-Aware Modules (VAMs) to resolve the validity mask mismatch problem, which enables many modern deep learning components to be applied to sparse data. Further, a multi-scale completion module is proposed to utilize multiple scales of context to fill the missing information. Experimental results show that MASC-Net outperforms the state-of-the-art depth densification methods on the KITTI depth completion benchmark in both quantitative and qualitative measures.

(Paper 28)
Title: Domain Generalization via Optical Flow: Training a CNN in a Low-Quality Simulation to Detect Obstacles in the Real World
Moritz Sperling (Darmstadt University of Applied Sciences), Giovanni Beltrame (Polytechnique de Montreal)
[paper][poster][video][slack channel]

Abstract
Many applications in robotics and autonomous sys-tems benefit from machine learning applied to computer vision,but often the acquisition and preparation of data for training is complex and time-consuming. Simulation can significantly reduce the effort and potential risk of data collection, thereby allowing faster prototyping. However, the ability of a data-driven system to generalize from simulated data to the real world is far from obvious and often leading to inconsistent real-world results. This paper demonstrates that some properties of optical flow can be exploited to address this generalization problem. In this work, we train a neural network to detect collisions with simulated optical flow data. Our network, FlowDroNet, is able to correctly predict up to 89% of the collisions of a real-world dataset and easily achieves a higher detection accuracy when compared to a network trained on a similar dataset of real-world collisions. We release our code, models and a real-world dataset for collision avoidance as open-source. We also explore the relationship between the complexity of the input information and the ability to generalize to unseen environments, and show that in some situations, optical flow is an interesting tool to bridge the reality gap.

SESSION 4 PAPERS

(Paper 55)
Title: PVF-NET: Point & Voxel Fusion 3D Object Detection Framework for Point Cloud
Zhihao Cui, Zhenhua Zhang (University Technology of Sydney)
[paper][poster][video][slack channel]

Abstract

In this paper, we present a novel 3D object detection framework for locating 3D bounding boxes of the target in autonomous driving scenes. Our proposed framework consists of two novel modules, which are twofold proposal fusion module and the RoI deep fusion module. In the former module, we utilized the 3D voxel Sparse Convolution Neural Network (CNN) and PointNet-like network to coarse generate the voxel-based and point-based proposals, where these proposals contain voxel-dense and pointwise features under the raw point cloud. Twofold proposal fusion module integrated those proposals and extended the proposal generalization, thereby dramatically improve the proposals' recall rate in the first stage for further utilizing in the proposals refinement stage. Given the coarse integrated 3D proposals produced by the twofold proposal module, the RoI deep fusion module is proposed to abstract and aggregate the multi-scale voxel-based feature and the point-wise feature through voxel-aware pooling layer and point-aware pooling layer, respectively. Follow by that, the specific features on the different proposals are integrated via the proposals-aware fusion layer to further enrich the feature dimensionality and utilize the high-quality proposals for the proposals refinement stage to reinforce the prediction of the target bounding boxes. We conduct the experiments on KITTI dataset and evaluate our method on 3D object detection task. Our method achieved 76.79 mAP in moderate difficulty and outperformed many influential object detection models on the KITTI benchmark leaderboard.

(Paper 47)
Title: Leveraging Temporal Data for Automatic Labelling of Static Vehicles
Sean Walsh, Steven L. Waslander, Jason L. Ku, Alex D. Pon (University of Toronto)
[paper][poster][video][slack channel]

Abstract

The development of advanced 3D object detection algorithms for autonomous driving requires a variety of environments to be captured in labelled datasets. While a number of such datasets exist, new ones will continue to be needed to adapt to new domains, sensors and conditions. In this paper, we propose a method to ease the workload of annotators by automatically proposing high-recall labels for static vehicles. We make use of an object detection network pre-trained on an existing dataset to propose detections within a sequence. By determining the location of each frame in a common reference frame, all detections of a static vehicle will share the same location. By averaging these overlapping detections and extending the prediction to all reasonable frames, we generate identical labels for the same object throughout the sequence. We show how our method sequentially refines predictions in order to improve static object recall by over 20% and precision by 7% over an initial set of network proposals.

(Paper 38)
Title: In-Time 3D Reconstruction and Instance Segmentation from Monocular Sensor Data
Stefan Schurischuster (Technical University - Wien), Jorge Mario Loaiciga (COZYO), Andrija Kurtic (COZYO), Robert Sablatnig (Technical University - Wien)
[paper][poster][video][slack channel]

Abstract

In most implementations of 3D reconstruction, depth information is provided by RGB-D sensors recording RGB and depth for each pixel. However, these sensors are still considerably expensive, characterized by high consumption of resources on the hardware itself, and have failed to reach mass markets of everyday mobile devices. With this paper, we aim to demonstrate the possibility of leveraging devices lacking 3D sensory data, in combination with external processing, to enrich the AR experience. We propose a client-server architecture that, in combination with a mobile client, allows for scanning of indoor environments as 3D semantically labeled sceneries while providing the user with instant feedback about the scanning results. At its core, the system is composed of the following expendable and exchangeable modules: (1) Depth Prediction from 2D images (2) Semantic Instance Segmentation of 2D images and (3) 3D Projection and Reconstruction. The result is a continuously updated mesh of the scenery with instance level segmentations in 3D. In comparison to the state-of-the-art, we are not only independent of the RGB-D input, but offer an architecture that complements and enhances the current AR client frameworks while demanding little extra computation from the mobile device.

Session 4 cont'

(Paper 4)
Title: Domain Adaptation in Crowd Counting
Mohammad Hossain (Huawei Canada), Mahesh Kumar Krishna Reddy (University of Manitoba), Kevin Cannons (Huawei Canada), Zhan Xu (Huawei Canada), Yang Wang (University of Manitoba)
[paper][poster][video][slack channel]

Abstract
We consider the problem of domain adaptation in crowd counting. Given an input image of a crowd scene, our goal is to estimate the count of people in the image. Previous work in crowd counting usually assumes that training and test images are captured by the same camera. We argue that this is not realistic in real-world applications of crowd counting. In this paper, we consider a domain adaptation setting in crowd counting where we have a source domain and a target domain. For example, these two domains might correspond to cameras at two different locations (i.e., with differing viewpoints, illumination conditions, environment objects, crowd densities, etc.). We have enough labeled training data from the source domain, but we only have either unlabeled data or a small number of labeled data in the target domain. Our goal is to train a crowd counting system that performs well in the target domain. We believe this setting is closer to real-world deployment of crowd counting systems. Due to the domain shift, a model trained from the source domain is unlikely to perform well in the target domain. In this paper, we propose several domain adaptation techniques for this problem. Our experimental results demonstrate the superior performance of our proposed approach on several benchmark datasets.

(Paper 57)
Title: Gas Prices of America: The Machine-Augmented Crowd-Sourcing Era
Kevin Dick, François Charih, Jimmy Woo, James R. Green (Carleton University)
[paper][poster][video][slack channel]

Abstract
Google Street View (GSV) comprises the largest collection of vehicle-based imagery of the natural environment. With high spatial resolution, GSV has been widely adopted to study the natural environment despite its relatively low temporal resolution (i.e. little time-series imagery available at a given location). However, vehicular-based imagery is poised to grow dramatically with the prophesied circulation of fleets of highly instrumented autonomous vehicles (AVs), producing high spatio-temporal resolution imagery of urban environments. As with GSV, leveraging these data presents the opportunity to extract information about the lived environment, while their high temporal resolution enables the study and annotation of time-varying phenomena. For example, circulating AVs will often capture location-coded images of gas stations. With a suitable computer vision system, one could extract the advertised numerical gas prices and automatically update the crowd-sourced GasBuddy application. To this end, we assemble and release the Gas Prices of America (GPA) dataset, a large-scale, benchmark dataset of advertised gas prices from GSV imagery across the forty-nine mainland United States of America. Comprising 2,048 high quality annotated images, the GPA dataset enables the development and evaluation of computer vision models for gas price extraction from complex urban scenes. More generally, this dataset provides a challenging benchmark against which computer vision models can be evaluated for multi-number, multi-digit recognition tasks of metrics in the wild. Highly accurate models, when integrated with AV platforms, will represent the first opportunity to automatically update the traditionally human crowd-sourced GasBuddy dataset, heralding an era of machine-augmented crowd-sourcing. The dataset is available online at cu-bic.ca/gpa and at doi.org/10.5683/SP2/KQ6VNG.

SESSION 5 PAPERS

(Paper 33)

Title: Gradient-Based Auto-Exposure Control Applied to a Self-Driving Car

Ishaan Mehta, Mingliang Tang, Tim Barfoot (University of Toronto)

[paper][poster][video][slack channel]

Abstract

As vision plays a central role in the operation of autonomous cars, one key challenge is that the limited dynamic range of camera sensors can only capture a certain portion of the scene radiance. This can lead to loss of information from images, which affects the performance of autonomous cars. To address this, we present an implementation of an exposure compensation method from the literature to auto-adjust camera exposure for the cameras mounted on a self-driving car. Furthermore, we extend this algorithm to incorporate gain compensation. The algorithm dynamically changes camera exposure time and gain settings with the intent to maximize image gradient information. The algorithm was evaluated in both indoor and outdoor environments, and experimental results demonstrate the effectiveness of our implementation. An open-source implementation of our technique is provided.

(Paper 18)

Title: CVNodes: A Visual Programming Paradigm for Developing Computer Vision Algorithms

Andrew Hogue, Junfeng Wang (University of Ontario Institute of Technology)

[paper][poster][video][slack channel]

Abstract

Advances in machine learning has led to a rapid pace of innovation in Computer vision and deep learning classification algorithms. Deep learning classification models are often limited in flexibility due to their fixed pre-processing steps embedded into the algorithm, and lack ways to easily iterate, debug, and analyze developed algorithms without programming knowledge. The lack of high level tools for developing vision algorithms leads to longer development times that require significant knowledge of underlying algorithms. What about individuals without this deep knowledge of machine learning and vision yet wish to develop algorithms to prototype ideas? What about non- programmers such as designers and artists that wish to utilize the state-of-the-art in computer vision in their work? To address this under-served community, we propose a visual-programming solution akin to those found in modern game engines geared towards computer vision algorithm development. This results in a new prototyping tool to empower researchers and non- programmers to easily iterate algorithm development, use pre-trained classification models, and provide statistical post analysis tools.

(Paper 24)

Title: Real-time Motion Planning for Robotic Teleoperation Using Dynamic-goal Deep Reinforcement Learning

Kaveh Kamali(École de technologie supérieure - Montreal)

[paper][poster][video][slack channel]

Abstract

We propose Dynamic-goal Deep Reinforcement Learning (DGDRL) method to address the problem of robot arm motion planning in telemanipulation applications. This method intuitively maps human hand motions to a robot arm in real-time, while avoiding collisions, joint limits and singularities. We further propose a novel hardware setup, based on the HTC VIVE VR system, that enables users to smoothly control the robot tool position and orientation with hand motions, while monitoring its movements in a 3D virtual reality environment. A VIVE controller captures 6D hand movements and gives it as reference trajectory to a deep neural policy network for controlling the robot's joint movements. Our DGDRL method leverages the state-of-art Proximal Policy Optimization (PPO) algorithm for deep reinforcement learning to train the policy network with the robot joint values and reference trajectory observed at each iteration. Since training the network on a real robot is time-consuming and unsafe, we developed a simulation environment called RobotPath which provides kinematic modeling, collision analysis and a 3D VR graphical simulation of industrial robots. The deep neural network trained using RobotPath is then deployed on a physical robot (ABB IRB 120) to evaluate its performance. We show that the policies trained in the simulation environment can be successfully used for trajectory planning on a real robot.

Session 5 cont'

(Paper 26)

Title: Towards End-to-end Learning of Visual Inertial Odometry with an EKF
Chunshang Li (University of Toronto)
[paper][poster][video][slack channel]

Abstract

Classical visual-inertial fusion relies heavily on manually crafted image processing pipelines, which are prone to failure in situations with rapid motion and texture-less scenes. While end-to-end learning methods show promising results in addressing these limitations, embedding domain knowledge in the form of classical estimation processes within the end-to-end learning architecture has the potential of combining the best of both worlds. In this paper, we propose the first end-to-end trainable visual-inertial odometry (VIO) algorithm that leverages a robo-centric Extended Kalman Filter (EKF). The EKF propagates states through a known inertial measurement unit (IMU) kinematics model and accepts relative pose measurements and uncertainties from a deep network as updates. The system is fully differentiable and can be trained end-to-end through backpropagation. Our method achieves a translation error of 1.27\% on the KITTI odometry dataset, which is competitive among classical and learning VIO methods. The implementation is publicly available on GitHub.

(Paper 30)

Title Evaluation of Skid-Steering Kinematic Models for Subartic Environments
Dominic Baril; Vincent Grondin, Simon-Pierre Deschênes, Johann Laconte, Maxime Vaidis, Vladimir Kubelka, André Gallant, Philippe Giguère, François Pomerleau (University of Laval)
[paper][poster][video][slack channel]

Abstract

In subarctic and arctic areas, large and heavy skid-steered robots are preferred for their robustness and ability to operate on difficult terrain. State estimation, motion control and path planning for these robots rely on accurate odometry models based on wheels velocities. However, the state-of-the-art odometry models for skid-steer mobile robots (SSMRs) have usually been tested on relatively lightweight platforms. In this paper, we focus on how these models perform when deployed on a large and heavy (590kg) SSMR. We collected more than 2km of data on both snow and concrete. We compare the ideal differential-drive, extended differential-drive, radius-of-curvature-based, and full linear kinematic models commonly deployed for SSMRs. Each of the models is fine-tuned by searching their optimal parameters on both snow and concrete. We then discuss the relationship between the parameters, the model tuning, and the final accuracy of the models.

SESSION 6 PAPERS

(Paper 29)
Title: Single-Stage End-to-End Temporal Activity Detection in Untrimmed Videos
Md Atiqur Rahman, Robert Laganiere (University of Ottawa)
[paper][poster][video][slack channel]

Abstract

In this paper, we address the problem of human activity detection in temporally untrimmed long video sequences, where the goal is to classify and temporally localize each activity instance in the input video. Inspired by the recent success of the single-stage object detection methods (e.g., SSD [1]), we propose an end-to-end trainable framework that learns task-specific spatio-temporal feature representations of a video sequence using a 3D convolutional neural network followed by a multi-scale detection architecture for direct classification and localization of the activities at varying temporal scales. Our proposed approach sets new state-of-the-art on the highly challenging THUMOS’14 temporal activity detection benchmark – up from 44.2% mAP to 49.0% mAP (an absolute 4.8% improvement) when the tIoU threshold is set to 0.5 during evaluation.

(Paper 20)
Title: Depth from Defocus on a Transmissive Diffraction Mask-based Sensor
Neeth Kunnath (Airy3D Inc), Ji-Ho Cho (Airy3D Inc), Michael Langer (McGill University)
[paper][poster][video][slack channel]

Abstract

In traditional depth from defocus (DFD) models, the blur kernel is a symmetric function whose width is proportional to the absolute distance in diopters between the scene point and the focal plane. A symmetric blur kernel implies a two-fold front-back ambiguity in the depth estimates, however. To resolve this ambiguity using only a single image of a scene, one typically introduces an asymmetry into the optics. Here we propose a fast and simple solution which uses a Transmissive Diffraction Mask (TDM), namely a transmissive grating placed directly in front the sensors. The grating is vertically oriented with a period of two pixels, and yields two interleaved images which both have asymmetric blur kernels. The sum of the two kernels behaves like a traditional symmetric blur kernel, and so we can apply a classical single-image edge-based DFD method to the sum of the two TDM images to estimate depth up to a sign ambiguity at each point. We then show how to use the difference of the two TDM images to resolve this sign ambiguity. The result is a sparse depth map which one can interpolate to a dense depth map using standard techniques.

(Paper 35)
Title: Differentiable Mask for Pruning Convolutional and Recurrent Networks
Ramchalam Ramakrishnan, Eyyüb Sari, Vahid Partovi Nia (Huawei Noah's Ark Lab)
[paper][poster][video][slack channel]

Abstract

Pruning is one of the well-known model reduction techniques. Deep networks require massive computation and such models need to be compressed to bring them on edge devices. Most existing pruning techniques are focused on convolutional networks, while text based models are still evolving. The emergence of multi-modal multi-task learning calls for a general method that works on vision and text architecture simultaneously. We introduce a differentiable mask, that induces sparsity on various granularity to fill this gap. We apply our method successfully to prune weights, filters, subnetwork of a convolutional architecture, as well as nodes of a recurrent network.

Session 6 cont'

(Paper 36)

Title: SpotNet: Self-Attention Multi-Task Network for Object Detection

Hughes Perreault (Polytechnique Montréal), Guillaume-Alexandre Bilodeau (Polytechnique Montréal), Nicolas Saunier (Polytechnique Montréal), Maguelonne Heritier (Genetec)

[paper][poster][video][slack channel]

Abstract

Humans are very good at directing their visual attention toward relevant areas when they search for different types of objects. For instance, when we search for cars, we will look at the streets, not at the top of buildings. The motivation of this paper is to train a network to do the same via a multi-task learning approach. To train visual attention, we produce foreground/background segmentation labels in a semi-supervised way, using background subtraction or optical flow. Using these labels, we train an object detection model to produce foreground/background segmentation maps as well as bounding boxes while sharing most model parameters. We use those segmentation maps inside the network as a self-attention mechanism to weight the feature map used to produce the bounding boxes, decreasing the signal of non-relevant areas. We show that by using this method, we obtain a significant mAP improvement on two traffic surveillance datasets, with state-of-the-art results on both UA-DETRAC and UAVDT.

(Paper 42)

Title: It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity

Rahul K. Dass (University of Miami) , Nick Petersen (University of Miami) , Ubbo Visser (University of Miami) , Marisa Omori (University of Missouri–St. Louis)

[paper][poster][video][slack channel]

Abstract

Analyses of existing public face datasets have shown that deep learning models (DLMs) grapple with racial and gender biases, raising concerns about algorithmic fairness in facial processing technologies (FPTs). Because these datasets are often comprised of celebrities, politicians, and mainly white faces, increased reliance on more diverse face databases has been proposed. However, techniques for generating more representative datasets are underdeveloped. To address this gap, we use the case of defendant mugshots from Miami–Dade County's (Florida, U.S.) criminal justice system to develop a novel technique for generating multidimensional race–ethnicity classifications for four groups: Black Hispanic, White Hispanic, Black non-Hispanic, and White non-Hispanic. We perform a series of experiments by fine-tuning seven DLMs using a full sample of mugshots (194,393) with race-ethnicity annotations from court records and a random subsample of mugshots (13,927) annotated by a group of research assistants. When evaluating DLMs' accuracies for a small subsample of 1,000 mugshots per race category, we find comparably high performance when predicting a Black and White binary race classification. However, when classifying mugshots based on four racial-ethnic subgroups, we find greater disparities in accuracy rates, where student annotated data outperform court records by 12.51% to 22.15%. Our methodology considers race as a multidimensional feature particularly for a more diverse face dataset and uses an averaged (consensus-based) approach to achieve a 74.09% accuracy rate based on annotated data representing only 2% of the full dataset. Our approach can be used to make DLM based FPTs to be more inclusive of the various multidimensional subcategories of race and ethnicity as they are being increasingly adopted by various organizations including the criminal justice system.