16th CONFERENCE ON COMPUTER AND ROBOT VISION

MAY 29 – 31, 2019
QUEEN'S UNIVERSITY
KINGSTON, CANADA

CO-CHAIRS

Anqi Xu, Element AI

Michael S. Brown, York University & Samsung

❖SPONSORS❖

Tuesday May 28th - Evening *University Club*

19:00 Joint Reception for AI/GI/CRV 2019

Wednesday May 29th - Morning *Stirling Hall – Room C*

8:30 Joint Welcoming Session for AI/GI/CRV 2019 (*Stirling Hall – Auditorium*)

Symposium 1
Chair: Michael Brown

9:00 Brian Funt, Simon Fraser University
Intrinsic Colour Problems

9:30 James Elder, York University
Explainable 3D Shape from Contour

Oral Session 1: Machine Learning Fundamentals 1
Chair: James Elder

10:00 Decoupling Spatial Pattern and Its Movement via Complex Factorization over Orthogonal Filter Pairs
Yanyan Mu, Roussos Dimitrakopoulos, Frank Ferrie

10:15 Intriguing Properties of Randomly Weighted Networks: Generalizing while Learning Next to Nothing
Amir Rosenfeld, John K. Tsotsos

10:30 Coffee Break *(Chernoff Hall - Lobby)*

Opening Keynote
Introduction: Michael Brown

11:00 David Hsu, National University of Singapore
See to Act and Act to See

Oral Session 2: Machine Learning Fundamentals II
Chair: Michael Brown

12:00 Resource-Aware Multicriterial Optimization of DNNs for Low-Cost Embedded Applications
Alexander Frickenstein, Christian Unger, Walter Stechele

12:15 Direct Fitting of Gaussian Mixture Models
Leonid Keselman, Martial Hebert

12:30 Lunch (*Leonard Hall – Cafeteria*)

Wednesday May 29th - Afternoon ***Stirling Hall – Room C***

PhD Dissertation Award Winners
Chair: Anqi Xu

14:00 Analyzing Cancers in Digitized Histopathology Images
Aïcha BenTaieb, Simon Fraser University

14:30 Learning Geometric and Lighting Priors from Natural Images
Yannick Hold-Geoffroy, Laval University

Symposium 2
Chair: Anqi Xu

15:00 Duckietown and the AI Driving Olympics
Liam Paull, University of Montreal

Oral Session 3: Applied Machine Learning I
Chair: Liam Paull

15:30 Rectification of Camera-Captured Document Images with Mixed Contents and Varied Layouts
Alexander Burden, Melissa Cote, Alexandra Branzan Albu

15:45 On Building Classification from Remote Sensor Imagery Using Deep Neural Networks and the Relation Between Classification and Reconstruction Accuracy
Bodhiswatta Chatterjee, Charalambos Poullis

16:00 Industry Spotlight / AI-GI Poster Session / Optional CRV Poster Session (*Biosciences Complex – Atrium*)

19:00 CRV 2019 VIP Dinner (*invite only*)

Thursday May 30th - Morning **Stirling Hall – Room C**

Oral Session 4: Human Action Prediction
Chair: Jean-François Lalonde

8:30 STAR-Net: Action Recognition using Spatio-Temporal Activation Reprojection
William McNally, Alexander Wong, John McPhee

8:45 Human Motion Prediction via Pattern Completion in Latent Representation Space
Yi Tian Xu, Yaqiao Li, David Meger

Symposium 3
Chair: Anqi Xu

9:00 Paul Kry, McGill University
Artistic Aerial Robots

9:30 Jean-Francois Lalonde, Laval University
Learning to Estimate Lighting

Oral Session 5: Adversarial Machine Learning
Chair: Jim Little

10:00 Adversarially Learned Abnormal Trajectory Classifier
Pankaj Raj Roy, Guillaume-Alexandre Bilodeau

10:15 Hierarchically-fused Generative Adversarial Network for Text to Realistic Image Synthesis
Xin Huang, Mingjie Wang, Minglun Gong

10:30 Coffee Break *(Chernoff Hall - Lobby)*

11:00 CRV Poster Session (See list of posters below)
**recommended poster size: A0 landscape (47" x 33" / 1.2m x 0.8m)*

12:30 Lunch (*Leonard Hall – Cafeteria*)

Thursday May 30th - Afternoon ***Stirling Hall – Room C***

Symposium 4

Chair: Michael Brown

14:00 Kwang Moo Yi, University of Victoria
Multi-view Geometry with Deep Learning

14:30 Jim Little, University of British Columbia
Challenges in Automating Sports Broadcasting

Oral Session 6: Applied Machine Learning II

Chair: François Michaud

15:00 Active Vision in the Era of Convolutional Neural Networks
Dimitri Gallos, Frank Ferrie

15:15 Apparent Age Estimation with Relational Networks
Eu Wern Teh, Graham W. Taylor

15:30 Coffee Break *(Chernoff Hall - Lobby)*

Symposium 5

Chair: Michael Brown

16:00 Olga Veksler, University of Waterloo
Efficient Graph Cut Optimization for Full CRFs with Quantized Edges

16:30 François Michaud, University of Sherbrooke
RTAB-Map and ODAS – Two Open Source Libraries for Robots in the Real World

17:00 CIPPRS AGM (tentative) (*Stirling Hall – Room C*)

18:00 Awards Banquet *(Portsmouth Olympic Harbour – off-campus)*

Friday May 31th - Morning **Stirling Hall – Room C**

Oral Session 7: Mapping

Chair: David Meger

8:30 Mapless Online Detection of Dynamic Objects in 3D Lidar
David J. Yoon, Tim Y. Tang, Timothy D. Barfoot

8:45 Network Uncertainty Informed Semantic Feature Selection for Visual SLAM
Pranav Ganti, Steven L. Waslander

Symposium 6

Chair: Anqi Xu

9:00 Florian Shkurti, University of Toronto
Collaborative Human-Robot Environmental Monitoring

9:30 David Meger, McGill University
Deep Reinforcement Learning that Matters

Oral Session 8: Localization

Chair: Yang Wang

10:00 Towards Direct Localization for Visual Teach and Repeat
Mona Gridseth, Timothy D. Barfoot

10:15 Point Me In The Right Direction: Improving Visual Localization on UAVs with Active Gimballed Camera Pointing
Bhavit Patel, Michael Warren, Angela P. Schoellig

10:30 Coffee Break *(Chernoff Hall - Lobby)*

Symposium 7

Chair: Anqi Xu

11:00 Deep Learning Models for Video Summarization
Yang Wang, University of Manitoba

Closing Keynote

Introduction: Anqi Xu

11:30 Chris Pal, École Polytechnique de Montréal
From Vision and Learning to Action and Understanding

12:30 CRV 2019 Closing Remarks
AI-GI-CRV 2019 Organization Review Meeting

17:00 AI/GI/CRV 2019 Joint Closing Remarks

❖KEYNOTES❖

David Hsu, National University of Singapore
See to Act and Act to See

In a robot system, perception and action are two interlocked essential elements. The purpose of perception is to act. The purpose of action is, at least, sometimes to achieve improved perception. Historically, the venerable modularity principle of system design led us to decompose a robot system and develop separate perception and decision modules that communicate through a "narrow" information interface. This separation became a major technical barrier and confined robots to well-controlled environments for decades. Capturing perceptual uncertainty and connecting perception with robot decision-making are key to robust robot performance. In this talk, we will look at ideas for tackling this challenge through planning, learning, and more interestingly, integrating planning and learning, in the context of autonomous driving among pedestrians and human-robot interaction.

***Dr. David Hsu** is a professor of computer science at the National University of Singapore (NUS) and a member of NUS Graduate School for Integrative Sciences & Engineering. He received PhD in computer science from Stanford University . At NUS, he co-founded NUS Advanced Robotics Center and has since been serving as the Deputy Director. He is an IEEE Fellow. His research spans robotics, AI, and computational structural biology. In recent years, he has been working on robot planning and learning under uncertainty and human-robot collaboration. He, together with colleagues and students, won the Humanitarian Robotics and Automation Technology Challenge Award at International Conference on Robotics & Automation (ICRA) 2015, the RoboCup Best Paper Award at International Conference on Intelligent Robots & Systems (IROS) 2015, and the Best Systems Paper Award at Robotics: Science & Systems (RSS), 2017. He has chaired or co-chaired several major international robotics conferences, including International Workshop on the Algorithmic Foundation of Robotics (WAFR) 2004 and 2010, ICRA 2016, and RSS 2015. He was an associate editor of IEEE Transactions on Robotics. He is currently an editorial board member of Journal of Artificial Intelligence Research and a member of the RSS Foundation Board.*

Chris Pal, École Polytechnique de Montréal
From Vision and Learning to Action and Understanding

As computer vision techniques based on machine learning mature, their potential for driving complex control policies for robotics has captured the imagination of many researchers. However, understanding the behaviour of learning algorithms coupling vision and control as well as understanding the behaviour of control policies learned by such algorithms remain challenging open problems. I'll examine some recent work from my group and collaborators on these themes.

I'll begin with some examples of deep learning techniques that more explicitly account for the three and four dimensionality of the visual world. I'll go on to examine a recurrent neural network based visual comparison technique that allows the reward signal of a reinforcement learning (RL) algorithm to be learned. This allows us to combine imitation learning techniques with RL to control physical simulations of humanoid agents by simply watching an example of a desired motion.

Going further, as deep reinforcement learning driven by visual perception becomes more widely used, there is a growing need to better understand and probe the learned agents. I'll present a new method for synthesizing visual inputs that lead to critical or risky states in RL agents -- of the type in which a very high or a very low reward can be achieved depending on which action is taken. In our experiments we show that this method can generate insights for a variety of environments and reinforcement learning methods. I'll present some of our results on the standard Atari benchmark games as well as in an autonomous driving simulator.

***Dr. Chris Pal** is an associate professor in the department of computer and software engineering at the École Polytechnique of Montreal. Prior to arriving in Montreal, he was a professor in the department of Computer Science at the University of Rochester. He has been a research scientist with the University of Massachusetts and has also been affiliated with the Interactive Visual Media Group and the Machine Learning and Applied Statistics groups at Microsoft Research. His research at Microsoft lead to three patents on image processing, computer vision and interactive multimedia.*

He earned his M. Math and PhD from the University of Waterloo in Canada. During his masters research he developed methods for automated cartography and the analysis of high resolution digital aerial photography. He was also involved with a number of software engineering projects developing spatial databases for managing environmental information. His PhD research led to contributions applying probability models and optimization techniques to image, video and signal processing.

During his PhD studies Chris was also a research assistant at the University of Toronto in the Department of Electrical and Computer Engineering. At Toronto he collaborated closely with the Banting and Best Department of Medical Research. He preformed research on image processing and statistical methods for the analysis of large scale genomics and computational molecular biology experiments using DNA microarrays. Prior to his graduate studies Chris was with the multimedia research company Interval in Palo Alto, CA (Silicon Valley). As a result of his research at Interval he was awarded a patent on audio signal processing.

❖SYMPOSIA PRESENTATIONS❖

WEDNESDAY MAY 29th

Brian Funt, Simon Fraser University
Intrinsic Colour Problems

Colour is clearly valuable for vision, both human and machine. In the research literature – the colour constancy literature especially – one often reads statements about "intrinsic colour," but what if anything is intrinsic about colour and do we need that idea in any case? Related to the issue of "intrinsic colour," this talk will discuss metamer mismatching, colour constancy, colour in multi-illuminant scenes, what colour may or may not be good for, dichromatic vision, colour invariants, and the colorimetric (in)accuracy of colour cameras.

James Elder, York University
Explainable 3D Shape from Contour

We live in a 3D world, and perceive it as such. While stereo and motion parallax are important 3D cues, humans still perceive objects and scenes in 3D when viewed in a single image or at distances where stereopsis and motion cues are weak. Recent deep-learning algorithms show that it is possible to teach a network to estimate depth from a single image, but with millions of free parameters, these networks provide little insight into the geometric and computational principles that underpin single-view 3D perception.

Recent neuroscience has revealed localized areas of the brain specialized for distinct forms of 3D shape and scene encoding. Here I will present two specialized single-view 3D algorithms inspired by these findings. The first addresses the problem of single-view 3D reconstruction of Manhattan (rectilinear) objects, which has application in rapid 3D city modeling. The second address the problem of recovering the 3D rim of a solid object given only its silhouette, which has applications in robotics and 2D to 3D film conversion. Both algorithms are explainable and can therefore also serve as scientific models for human visual processing.

Liam Paull, University of Montreal
Duckietown and the AI Driving Olympics

Robots have become a vehicle for exploring ideas in the production of creative artifacts such as drawings and paintings. At the core of many of these endeavors are important technical challenges and computational problems that require a scientific approach to designing and evaluating these robot systems. This talk will focus on two aerial robot applications that use CrazyFlie quadrotor robots. I will describe a method for creating stippled prints and the various challenges such as stipple placement, path planning, ink modeling, and control adjustments in tethered and untethered flight. I will also present trajectory generation alternatives for creating single-stroke light paintings where we reduce the cost of a minimum snap piecewise polynomial flight trajectory passing through a set of waypoints by displacing those waypoints towards or away from the camera while preserving their projected position. Inspired by Picasso and other artists, our light painting results involve a variety of one-stroke animal illustrations and other creative examples with human participation.

THURSDAY MAY 30th

Paul Kry, McGill University
Artistic Aerial Robots

Robots have become a vehicle for exploring ideas in the production of creative artifacts such as drawings and paintings. At the core of many of these endeavors are important technical challenges and computational problems that require a scientific approach to designing and evaluating these robot systems. This talk will focus on two aerial robot applications that use CrazyFlie quadrotor robots. I will describe a method for creating stippled prints and the various challenges such as stipple placement, path planning, ink modeling, and control adjustments in tethered and untethered flight. I will also present trajectory generation alternatives for creating single-stroke light paintings where we reduce the cost of a minimum snap piecewise polynomial flight trajectory passing through a set of waypoints by displacing those waypoints towards or away from the camera while preserving their projected position. Inspired by Picasso and other artists, our light painting results involve a variety of one-stroke animal illustrations and other creative examples with human participation.

Jean-François Lalonde, Laval University
Learning to Estimate Lighting

Combining virtual and real visual elements into a single, realistic image requires the accurate estimation of the lighting conditions of the real scene. Unfortunately, doing so typically requires specialized image capture, user input, and/or simplified scene models. In this talk, I will present approaches for automatically estimating lighting from a single image. In particular, I will present two recent works that rely on deep learning models for lighting estimation: 1) an approach that robustly estimates outdoor lighting in varying weather conditions, accurately capturing challenging effects such as soft shadows; and 2) another method that estimates spatially-varying indoor lighting conditions, and which automatically adapts to local effects such as occlusions. I will demonstrate that using our illumination estimates for applications like 3D object insertion can achieve photo-realistic results on a wide variety of challenging scenarios.

Kwang Moo Yi, University of Victoria
Multi-view Geometry with Deep Learning

Estimating the camera pose from images is one of the core problems in Computer Vision. It is the core enabler for Augmented/Mixed Reality, Self-driving cars, Autonomous Drones. Interestingly, despite the recent advancements in Deep Learning, simple black-box approaches fail to solve this problem, unlike many other fields in Computer Vision.
In this talk, I will show that, by decomposing the problem into multiple stages, and by solving them one by one and integrating them together, we can leverage the vast amount of data available from the internet and bring a significant leap in performance through Deep Learning. I will first introduce how we can leverage weekly labeled data, together with a carefully designed network to learn a Deep Local Feature, and then talk about how to use them properly to get a good camera pose. I will further discuss how we can even include non-differentiable components in this Deep pipeline, that further boosts performance.

Jim Little, University of British Columbia
Challenges in Automating Sports Broadcasting

We describe mechanisms that allow a collection of cameras to assemble an informative broadcast of a team sporting event, without requiring extensive augmentation of the camera systems. Using consumer-grade cameras means that camera pose must be estimated from the image stream. The system must also identify the current stage of the game and determine a plausible future sequence of play, using only the detected position of the players, in real time. Finally, the lack of ground truth for the camera selection process means we must take advantage of unannotated broadcast streams to regress a selector that determines which camera is broadcast, consistent with typical broadcast selection.

Olga Veksler, University of Waterloo
Efficient Graph Cut Optimization for Full CRFs with Quantized Edges

Fully connected pairwise Conditional Random Fields (Full-CRF) with Gaussian edge weights can achieve superior results compared to sparsely connected CRFs. However, traditional methods for Full-CRFs are too expensive. Previous work develops efficient approximate optimization based on mean field inference, which is a local optimization method and can be far from the optimum. We propose efficient and effective optimization based on graph cuts for Full-CRFs with {\em quantized} edge weights. To quantize edge weights, we partition the image into superpixels and assume that the weight of an edge between any two pixels depends only on the superpixels these pixels belong to. Our quantized edge CRF is an approximation to the Gaussian edge CRF, and gets closer to it as superpixel size decreases. Being an approximation, our model offers an intuition about the regularization properties of the Guassian edge Full-CRF. For efficient inference, we first consider the two-label case and develop an approximate method based on transforming the original problem into a smaller domain. Then we handle multi-label CRF by showing how to implement expansion moves. In both binary and multi-label cases, our solutions have significantly lower energy compared to that of mean field inference. We also show the effectiveness of our approach on semantic segmentation task.

François Michaud, University of Sherbrooke
RTAB-Map and ODAS – Two Open Source Libraries for Robots in the Real World

This talk presents two open source libraries developed to address the challenges of navigation and artificial audition for robots operating in the real world. These libraries are being developed with real-time, limited processing and robustness requirements in mind.

FRIDAY MAY 31h

Florian Shkurti, University of Toronto
Collaborative Human-Robot Environmental Monitoring

In this talk I will focus on the problem of enabling robot videographers/documentarians that autonomously navigate in unstructured 3D environments, alongside environmental scientists, to help them record footage that they deem valuable for their work.

I will present a method to infer the expert's reward function over images, using a small number of labeled and a large number of unlabeled examples. This reward function is used to guide the robot's visual exploration and data collection in unknown scenes. I will also present vision-based algorithms for tracking and navigation that are robust to long-term loss of visual contact with the subject, by making use of the subject's learned behavior, estimated via inverse reinforcement learning. Experimental validation of these methods on multi-robot systems operating in underwater and aerial environments will be shown.

David Meger, McGill University
Deep Reinforcement Learning that Matters

In recent years, significant progress has been made in solving challenging problems across various domains using deep reinforcement learning (RL). Reproducing existing work and accurately judging the improvements offered by novel methods is vital to sustaining this progress. Unfortunately, reproducing results for state-of-the-art deep RL methods is seldom straightforward. In particular, non-determinism in standard benchmark environments, combined with variance intrinsic to the methods, can make reported results tough to interpret. Without significance metrics and tighter standardization of experimental reporting, it is difficult to determine whether improvements over the prior state-of-the-art are meaningful. In this paper, we investigate challenges posed by reproducibility, proper experimental techniques, and reporting procedures. We illustrate the variability in reported metrics and results when comparing against common baselines and suggest guidelines to make future results in deep RL more reproducible. We aim to spur discussion about how to ensure continued progress in the field by minimizing wasted effort stemming from results that are non-reproducible and easily misinterpreted.

Yang Wang, University of Manitoba
Deep Learning Models for Video Summarization

With the large amount of videos available online, video summarization has become an important topic. Given an input long video, the goal of video summarization is to produce a shorter video that summarizes the main content of the input video. In this talk, I will present our recent work on using deep learning for video summarization. First, I will present our work on fully convolutional sequence model. We formulate video summarization as a sequence labeling problem. Unlike existing approaches that use recurrent models, we propose fully convolutional sequence models to solve video summarization. We firstly establish a novel connection between semantic segmentation and video summarization, and then adapt popular semantic segmentation networks for video summarization. Second, I will present our work on learning video summarization models from unpaired data. This is motivated by the fact that collecting supervised data for video summarization is expensive. To address this limitation, we propose a novel formulation to learn video summarization from unpaired data. We present an approach that learns to generate optimal video summaries using a set of raw videos (V) and a set of summary videos (S), where there exists no correspondence between V and S. We argue that this type of data is much easier to collect.

❖ORAL PRESENTATIONS❖

WEDNESDAY MAY 29th

Decoupling Spatial Pattern and Its Movement via Complex Factorization over Orthogonal Filter Pairs

Yanyan Mu, Roussos Dimitrakopoulos, Frank Ferrie

Variations between related images (e.g. due to motions) can caused by different independent factors. A qualified representation can decouple the underlying explanatory factors rather than keeping them mixed. After decoupling, each factor lies in a lower dimension abstract space. Different computer vision tasks can be done in different abstract spaces more efficiently than in the original pixel space. For example, conducting object recognition in appearance space can result in an invariant recognition; estimating object motion in location space yields a result regardless of the object itself. In this paper, we propose an algorithm to decouple object appearance and location to amplitude and phase in static images by using complex factorization over orthogonal filter pairs. In particular, we show that, i) Orthogonal filter pairs can be learned in an unsupervised manner from multiple consecutive frames; ii) Object movement is encoded in the factorized phase gradient between frames over time. As a proof of concept, we present experiments on the application of our framework to the recovery of the optical flow. Here object movement is successfully captured by phase gradient.

Intriguing Properties of Randomly Weighted Networks: Generalizing While Learning Next to Nothing

Amir Rosenfeld, John K. Tsotsos

Training deep neural networks results in strong learned representations that show good generalization capabilities. In most cases, training involves iterative modification of all weights in the network via back-propagation. In Extreme Learning Machines, it has been suggested to set the first layer of a network to fixed random values instead of learning it. In this paper, we propose to take this approach a step further and fix almost all layers of a deep convolutional neural network, allowing only a small portion of the weights to be learned. As our experiments show, fixing even the majority of the parameters of the network often results in performance which is on par with the performance of learning all of them. The implications of this intriguing property of deep neural networks are discussed and we suggest practical ways to harness it to create more robust and compact representations.

Resource-Aware Multicriterial Optimization of DNNs for Low-Cost Embedded Applications

Alexander Frickenstein, Christian Unger, Walter Stechele

Despite their outstanding success in solving complex computer vision problems, Deep Neural Networks (DNNs) still require high-performance hardware for real-time inference. Therefore they are not applicable to low-cost embedded hardware, where memory resources, computational performance and power consumption are restricted. Furthermore, current approaches of fitting neural networks to embedded hardware are time consuming, inducing slow development cycles. To address these drawbacks and satisfy the demands of embedded hardware, this paper proposes a computationally efficient magnitude-based pruning scheme, based on a half-interval search, combined with effective weight sharing, fixed-point quantization, and lossless compression. The proposed solution can be utilized to generate an optimized model, either with respect to memory demand or execution time. For instance, the memory demand of LeNet is compressed about 385×. VGG16 is pruned by about 14.5×, whilst its computational costs are reduced by about 1.6× for a CPU-based application and 4.8× for an FPGA one.

Direct Fitting of Gaussian Mixture Models

Leonid Keselman, Martial Hebert

When fitting Gaussian Mixture Models to 3D geometry, the model is typically fit to point clouds, even when the shapes were obtained as 3D meshes. Here we present a formulation for fitting Gaussian Mixture Models (GMMs) directly to geometric objects, using the triangles of triangular mesh instead of using points sampled from its surface. We demonstrate that this modification enables fitting higher-quality GMMs under a wider range of initialization conditions. Additionally, models obtained from this fitting method are shown to produce an improvement in 3D registration for both meshes and RGB-D frames.

Rectification of Camera-Captured Document Images with Mixed Contents and Varied Layouts

Alexander Burden, Melissa Cote, Alexandra Branzan Albu

This paper focuses on the rectification of camera captured document images with varied layouts of mixed contents. Document images acquired via cameras, including smartphones, are typically plagued by perspective, geometric, and/or rotational distortion that hinders document analysis processes. In this paper, we propose an approach to camera-captured image rectification of text and non-text regions that handles perspective, geometric and rotational distortions present in planar and curled documents, extending a state-of-the-art content-based rectification method. We define surface projections via a three-tiered local transformation model, in which primary curved surface projections are formed from individual text regions, and secondary and tertiary surface projections are formed from non-text regions, resulting in a ’patchwork’ combination of surfaces spanning the document image. This transformation model allows us to process document images with varied layouts of mixed contents, including large images and graphics, that also contain some justified text. Experiments and comparisons with a state-of-the-art content-based rectification approach on the public IUPR dataset demonstrate the value of the proposed approach on two levels: 1) a significantly improved rectification performance using standard optical character recognition metrics, along with increased document readability, and 2) an improved range of applicability, i.e. ability to correct document images showing various layouts and content types.

On Building Classification from Remote Sensor Imagery Using Deep Neural Networks and the Relation Between Classification and Reconstruction Accuracy Using Border Localization as Proxy

Bodhiswatta Chatterjee, Charalambos Poullis

Convolutional neural networks have been shown to have a very high accuracy when applied to certain visual tasks and in particular semantic segmentation. In this paper we address the problem of semantic segmentation of buildings from remote sensor imagery. We present ICTNet: a novel network with the underlying architecture of a fully convolutional network, infused with feature re-calibrated Dense blocks at each layer. Uniquely, the proposed network combines the localization accuracy and use of context of the U-Net network architecture, the compact internal representations and reduced feature redundancy of the Dense blocks, and the dynamic channel-wise feature re-weighting of the Squeeze-and-Excitation(SE) blocks. The proposed network has been tested on INRIA’s benchmark dataset and is shown to outperform all other state-of-the-art by more than 1.5% on the Jaccard index. Furthermore, as the building classification is typically the first step of the reconstruction process, in the latter part of the paper we investigate the relationship of the classification accuracy to the reconstruction accuracy. A comparative quantitative analysis of reconstruction accuracies corresponding to different classification accuracies confirms the strong correlation between the two. We present the results which show a consistent and considerable reduction in the reconstruction accuracy. The source code and supplemental material is publicly available at http://www.theICTlab.org/lp/2019ICTNet/.

THURSDAY MAY 10th

STAR-Net: Action Recognition using Spatio-Temporal Activation Reprojection
William McNally, Alexander Wong, John McPhee

While depth cameras and inertial sensors have been frequently leveraged for human action recognition, these sensing modalities are impractical in many scenarios where cost or environmental constraints prohibit their use. As such, there has been recent interest on human action recognition using low-cost, readily-available RGB cameras via deep convolutional neural networks. However, many of the deep convolutional neural networks proposed for action recognition thus far have relied heavily on learning global appearance cues directly from imaging data, resulting in highly complex network architectures that are computationally expensive and difficult to train. Motivated to circumvent the challenges associated with training complex network architectures, we introduce the concept of spatio-temporal activation reprojection (STAR). More specifically, we reproject the spatio-temporal activations generated by human pose estimation layers in space and time using a stack of 3D convolutions. Experimental results on UTD-MHAD and J-HMDB demonstrate that an end-to-end architecture based on the proposed STAR framework (which we nickname STAR-Net) is proficient in single-environment and small-scale applications. On UTD-MHAD, STAR-Net outperforms several methods using richer data modalities such as depth and inertial sensors.

Human Motion Prediction via Pattern Completion in Latent Representation Space
Yi Tian Xu, Yaqiao Li, David Meger

Inspired by ideas in cognitive science, we propose a novel and general approach to solve human motion understanding via pattern completion on a learned latent representation space. Our model outperforms current state-of-the-art methods in human motion prediction across a number of tasks, with no customization. To construct a latent representation for time-series of various lengths, we propose a new and generic autoencoder based on sequence-to-sequence learning. While traditional inference strategies find a correlation between an input and an output, we use pattern completion, which views the input as a partial pattern and to predict the best corresponding complete pattern. Our results demonstrate that this approach has advantages when combined with our autoencoder in solving human motion prediction, motion generation and action classification.

Adversarially Learned Abnormal Trajectory Classifier
Pankaj Raj Roy, Guillaume-Alexandre Bilodeau

We address the problem of abnormal event detection from trajectory data. In this paper, a new adversarial approach is proposed for building a deep neural network binary classifier, trained in an unsupervised fashion, that can distinguish normal from abnormal trajectory-based events without the need for setting manual detection threshold. Inspired by the generative adversarial network (GAN) framework, our GAN version is a discriminative one in which the discriminator is trained to distinguish normal and abnormal trajectory reconstruction errors given by a deep autoencoder. With urban traffic videos and their associated trajectories, our proposed method gives the best accuracy for abnormal trajectory detection. In addition, our model can easily be generalized for abnormal trajectory-based event detection and can still yield the best behavioural detection results as demonstrated on the CAVIAR dataset.

Hierarchically-fused Generative Adversarial Network for text to realistic image synthesis

Xin Huang, Mingjie Wang, Minglun Gong

In this paper, we present a novel Hierarchically-fused Generative Adversarial Network (HfGAN) for synthesizing realistic images from text descriptions. While existing approaches on this topic have achieved impressive success, to generate 256×256 images from captions, they commonly resort to coarse-to-fine scheme and associate multiple discriminators in different stages of the networks. Such a strategy is both inefficient and prone to artifacts. Motivated by the above findings, we propose an end-to-end network that can generate 256 × 256 photo-realistic images with only one discriminator. We fully exploit the hierarchical information from different layers and directly generate the fine-scale images by adaptively fusing features from multi-hierarchical layers. We quantitatively evaluate the synthesized images with Inception Score, Visual-semantic Similarity and average training time on the CUB birds, Oxford-102 flowers, and COCO datasets. The results show that our model is more efficient and noticeably outperforms the previous state-of-the-art methods.\

Active Vision in the Era of Convolutional Neural Networks

Dimitri Gallos, Frank Ferrie

In this work, we examine the literature of active object recognition in the past and present. We note that methods in the past used a notion of recognition ambiguity in order to find a next best view policy that could disambiguate the object with the fewest camera moves. Present methods on the other hand use deep reinforcement learning to learn camera control policies from the data. We show on a public dataset, that reinforcement learning methods are not superior to a policy of adequately sampling the object view-sphere. Instead of focusing on finding the next best view, we examine a recent method of quantifying recognition uncertainty in deep learning as a potential application to active object recognition. We find that predictions with this technique are well calibrated with respect to the performance of a network on a test-set, showing that it could be useful in an active vision scenario.

Apparent Age Estimation with Relational Networks

Eu Wern Teh, Graham W. Taylor

Apparent age estimation is a newly proposed and under-studied problem of predicting the age that someone “looks” rather than their actual age. It has applications in many areas within the beauty industry[3]. Methods based on convolutional neural networks (CNNs) have proved to be state-of-the-art on the few datasets used to benchmark this task. However, such CNNs typically collapse spatial information via a Global Average Pooling operation. They do not perform any explicit treatment of spatial relationships of the higher-level features which emerge in the later stages of the network and which may correspond to facial parts or blemishes that are characteristic of age. In this paper, we consider a newly proposed CNN module called relational networks that explicitly capture spatial relationships. We hypothesize that we can estimate age better by learning such relationships in the final set of CNN feature maps where spatial information is still retained. Experiments were conducted on both ChaLearn LAP 2015 and 2016 datasets [6], [7] showing that on average, there is a 3.53% improvement on Mean Absolute Error and 3.31% improvement on _x005F_x000f_-error when compared to the baseline. A test was also calculated to show that the improvement is statistically significant.

FRIDAY MAY 31th

Mapless Online Detection of Dynamic Objects in 3D Lidar

David J. Yoon, Tim Y. Tang, Timothy D. Barfoot

This paper presents a model-free, setting-independent method for online detection of dynamic objects in 3D lidar data. We explicitly compensate for the moving-while-scanning operation (motion distortion) of present-day 3D spinning lidar sensors. Our detection method uses a motion-compensated freespace querying algorithm and classifies between dynamic (currently moving) and static (currently stationary) labels at the point level. For a quantitative analysis, we establish a benchmark with motion-distorted lidar data using CARLA, an open-source simulator for autonomous driving research. We also provide a qualitative analysis with real data using a Velodyne HDL-64E in driving scenarios. Compared to existing 3D lidar methods that are model-free, our method is unique because of its setting independence and compensation for pointcloud motion distortion.

Network Uncertainty Informed Semantic Feature Selection for Visual SLAM

Pranav Ganti, Steven L. Waslander

In order to facilitate long-term localization using a visual simultaneous localization and mapping (SLAM) algorithm, careful feature selection can help ensure that reference points persist over long durations and the runtime and storage complexity of the algorithm remain consistent. We present SIVO (Semantically Informed Visual Odometry and Mapping), a novel information-theoretic feature selection method for visual SLAM which incorporates semantic segmentation and neural network uncertainty into the feature selection pipeline. Our algorithm selects points which provide the highest reduction in Shannon entropy between the entropy of the current state and the joint entropy of the state, given the addition of the new feature with the classification entropy of the feature from a Bayesian neural network. Each selected feature significantly reduces the uncertainty of the vehicle state and has been detected to be a static object (building, traffic sign, etc.) repeatedly with a high confidence. This selection strategy generates a sparse map which can facilitate longterm localization. The KITTI odometry dataset is used to evaluate our method, and we also compare our results against ORB SLAM2. Overall, SIVO performs comparably to the baseline method while reducing the map size by almost 70%.

Towards Direct Localization for Visual Teach and Repeat

Mona Gridseth, Timothy D. Barfoot

Vision-based path following allows robots to autonomously repeat manually taught paths. Stereo Visual Teach and Repeat (VT&R) [1] accomplishes robust long-range path following in unstructured outdoor environments. VT&R uses sparse features to match images for visual odometry (VO) and localization. This paper describes our first implementation of direct localization for VT&R. Instead of using sparse visual features for image matching, we minimize a photometric residual cost over the whole image. We compare the performance of feature-based and direct localization using challenging offroad driving datasets. The results show that direct localization consistently achieves more accurate pose estimation under nominal conditions, but further work is required to increase robustness to large lighting change between the teach and repeat images.

Point Me In The Right Direction: Improving Visual Localization on UAVs with Active Gimballed Camera Pointing

Bhavit Patel, Michael Warren, Angela P. Schoellig

Robust autonomous navigation of multirotor UAVs in GPS-denied environments is critical to enable their safe operation in many applications such as surveillance and reconnaissance, inspection, and delivery services. In this paper, we use a gimballed stereo camera for localization and demonstrate how the localization performance and robustness can be improved by actively controlling the camera's viewpoint. For an autonomous route-following task based on a recorded map, multiple gimbal pointing strategies are compared: offthe-shelf passive stabilization, active stabilization, minimization of viewpoint orientation error, and pointing the camera optical center at the centroid of previously observed landmarks. We demonstrate improved localization performance using an active gimbal-stabilized camera in multiple outdoor flight experiments on routes up to 315 m, and with 6-25 m altitude variations. Scenarios are shown where a static camera frequently fails to localize while a gimballed camera attenuates perspective errors to retain localization. We demonstrate that our orientation matching and centroid pointing strategies provide the best performance; enabling localization despite increasing velocity discrepancies between the map-generation flight and the live flight from 3-9 m/s, and 8 m path offsets.

❖POSTER PRESENTATIONS❖

Generative Adversarial Networks Using Adaptive Convolution
Nhat M. Nguyen, Nilanjan Ray
Most existing GAN architectures that generate images use transposed convolution or resize-convolution as their upsampling algorithm from lower to higher resolution feature maps in the generator. We propose a novel adaptive convolution method that learns the upsampling algorithm based on the local context at each location to address this problem. We modify a baseline GAN architecture by replacing normal convolutions with adaptive convolutions in the generator. Our method is orthogonal to others that seek to improve GAN by incorporating high level information. Experiments on CIFAR10 dataset show that our modified models improve the baseline model by a large margin on visually diverse datasets.

Instance Segmentation based Semantic Matting for Compositing Applications
Guanqing Hu, James J. Clark
Image compositing is a key step in film making and image editing that aims to segment a foreground object and combine it with a new background. Automatic image compositing can be done easily in a studio using chromakeying when the background is pure blue or green. However, image compositing in natural scenes with complex backgrounds remains a tedious task, requiring experienced artists to hand-segment. In order to achieve automatic compositing in natural scenes, we propose a fully automated method that integrates instance segmentation and image matting processes to generate high-quality semantic mattes that can be used for image editing task. Our approach can be seen both as a refinement of existing instance segmentation algorithms and as a fully automated semantic image matting method. It extends automatic image compositing techniques such as chroma-keying to scenes with complex natural backgrounds without the need for any kind of user interaction. The output of our approach can be considered as both refined instance segmentations and alpha mattes with semantic meanings. We provide experimental results which show improved performance results as compared to existing approaches.

Robust Facial Alignment with Internal Denoising Auto-Encoder
Decky Aspandi, Oriol Martinez, Federico Sukno, Xavier Binefa
The development of facial alignment models is growing rapidly thanks to the availability of large facial landmarked datasets and powerful deep learning models. However, important challenges still remain for facial alignment models to work on images under extreme conditions, such as severe occlusions or large variations in pose and illumination. Current attempts to overcome this limitation have mainly focused on building robust feature extractors with the assumption that the model will be able to discard the noise and select only the meaningful features. However, such an assumption ignores the importance of understanding the noise that characterizes unconstrained images, which has been shown to benefit computer vision models if used appropriately on the learning strategy. Thus, in this paper we investigate the introduction of specialized modules for noise detection and removal, in combination with our state-of-the-art facial alignment module and show that this leads to improved robustness both to synthesized noise and in-the-wild conditions. The proposed model is built by combining two major subnetworks: internal image denoiser (based on the Auto-Encoder architecture) and facial landmark localiser (based on the inception-resnet architecture). Our results on the 300-W and Menpo datasets show that our model can effectively handle different types of synthetic noise, which also leads to enhanced robustness in real-world unconstrained settings, reaching top state-of-the-art accuracy.

HandSeg: An Automatically Labeled Dataset for Hand Segmentation from Depth Images

Abhishake Kumar Bojja, Franziska Mueller, Sri Raghu Malireddi, Markus Oberweger, Vincent Lepetit, Christian Theobalt, Kwang Moo Yi, Andrea Tagliasacchi

We propose an automatic method for generating high-quality annotations for depth-based hand segmentation, and introduce a large-scale hand segmentation dataset. Existing datasets are typically limited to a single hand. By exploiting the visual cues given by an RGBD sensor and a pair of colored gloves, we automatically generate dense annotations for two hand segmentation. This lowers the cost/complexity of creating high quality datasets, and makes it easy to expand the dataset in the future. We further show that existing datasets, even with data augmentation, are not sufficient to train a hand segmentation algorithm that can distinguish two hands. Source and datasets are publicly available at the project page.

Automated Acquisition of Anisotropic Friction

Keno Dreßel, Kenny Erleben, Paul G. Kry, Sheldon Andrews

Automated acquisition of friction data is an interesting approach to more successfully bridge the reality gap in simulation than conventional mathematical models. To advance this area of research, we present a novel inexpensive computer vision platform as a solution for collecting and processing friction data, and we make available the open source software and data sets collected with our vision robotic approach. This paper is focused on gathering data on anisotropic static friction behavior as this is ideal for inexpensive vision approach we propose. The data set and experimental setup provide a solid foundation for a wider robotics simulation community to conduct their own experiments.

Traffic Risk Assessment: A Two-Stream Approach Using Dynamic-Attention

Gary-Patrick Corcoran, James Clark

The problem being addressed in this research is performing traffic risk assessment on visual scenes captured via outward-facing dashcam videos. To perform risk assessment, a two-stream dynamic-attention recurrent convolutional neural architecture is used to provide a categorical risk level for each frame in a given input video sequence. The two-stream approach consists of a spatial stream, which analyzes individual video frames and computes high-level appearance features and a temporal stream, which analyzes optical flow between adjacent frames and computes high-level motion features. Both spatial and temporal streams are then fed into their respective recurrent neural networks (RNNs) that explicitly models the sequence of features in time. A dynamic-attention mechanism which allows the network to learn to focus on relevant objects in the visual scene is added. These objects are detected by a state-of-the-art object detector and correspond to vehicles, pedestrians, traffic signs, etc. The dynamic-attention mechanism not only improves classification performance, but also provides a method to visualize what the network “sees” when predicting a risk level. This mechanism allows the network to implicitly learn to focus on hazardous objects in the visual scene. Additionally, this research introduces an offline and online model that differ slightly in their implementations. The offline model analyzes the complete video sequence and scores a classification accuracy of 84.89%. The online model deals with an infinite stream of data and produces results in near real-time (7 frames-per-second); however, it suffers from a slight decrease in classification accuracy (79.90%).

Commodifying Pointing in HRI: Simple and Fast Pointing Gesture Detection from RGB-D Images

Bita Azari, Angelica Lim, Richard T. Vaughan

We present and characterize a simple and reliable method for detecting pointing gestures suitable for human-robot interaction applications using a commodity RGB-D camera. We exploit an existing Deep CNN model to robustly find hands and faces in RGB images, then examine the corresponding depth channel pixels to obtain full 3D pointing vectors. We test several methods of estimating the hand end-point of the pointing vector. The system runs at better than 30Hz on commodity hardware: exceeding the frame rate of typical RGB-D sensors. An estimate of the absolute pointing accuracy is found empirically by comparison with ground-truth data from a VICON motion-capture system, and the useful interaction volume established. Finally, we show an end-to-end test where a robot estimates where the pointing vector intersects the ground plane, and report the accuracy obtained. We provide source code as a ROS node, with the intention of contributing a commodity implementation of this common component in HRI systems.

Two-stream Action Recognition in Ice Hockey using Player Pose Sequences and Optical Flows

Kanav Vats, Helmut Neher, David A. Clausi, John Zelek

Current action recognition algorithms in ice hockey do not fully exploit the temporal cues available in video. To solve this challenge, we introduce a two-stream network utilizing player pose sequences and optical flow features for recognizing hockey actions. Player pose sequences are compact representations consisting of frame by frame human and stick joint locations and angles between joints. The optical flow features are obtained by a state-of-the-art optical flow algorithm. The player pose sequences are processed by a two-layered Long short-term memory (LSTM) network. The LSTM output is fused with optical flow features processed by a convolutional neural network (CNN). Experimental results demonstrate the efficacy of the method by achieving 90.48% test accuracy on the HARPET (Hockey Action Recognition Pose Estimation, Temporal) dataset thus surpassing current benchmark by 5%. The network performs better than the current benchmark in segregating similar classes like passing and shooting. It achieves a 90% reduction in parameters and 90% reduction in floating point operations per second (FLOPs) than the benchmark on the HARPET dataset, thus furthering the effectiveness of the network.

Automatic Temporally Coherent Video Colorization

Harrish Thasarathan, Kamyar Nazeri, Mehran Ebrahimi

Greyscale image colorization for applications in image restoration has seen significant improvements in recent years. Many of these techniques that use learning-based methods struggle to effectively colorize sparse inputs. With the consistent growth of the anime industry, the ability to colorize sparse input such as line art can reduce significant cost and redundant work for production studios by eliminating the in-between frame colorization process. Simply using existing methods yields inconsistent colors between related frames resulting in a flicker effect in the final video. In order to successfully automate key areas of large-scale anime production, the colorization of line arts must be temporally consistent between frames. This paper proposes a method to colorize line art frames in an adversarial setting, to create temporally coherent video of large anime by improving existing image to image translation methods. We show that by adding an extra condition to the generator and discriminator, we can effectively create temporally consistent video sequences from anime line arts.

Investigating Trust Factors in Human-Robot Shared Control: Implicit Gender Bias Around Robot Voice

Alexander Wong, Anqi Xu, Gregory Dudek

This paper explores the impact of warnings, audio feedback, and gender on human-robot trust in the context of autonomous driving and specifically shared robot control. We use pre-existing methods for the estimation and assessment of human-robot trust where trust was found to vary as a function of the quality of behavior of an autonomous driving controller. We extend these models and empirical methods to examine the impact of audio cues on trust, specifically studying the impacts of gender-specific audio cues on the elicitation of trust. Our study compares agents with and without human-voiced indicators of uncertainty and evaluates differences in trust with inferred and introspective methods. We find that a person's trust in a robot can be influenced by verbal feedback from the robot agent. Specifically, people tend to lend more trust to agents whose voice is of the same gender as their own.

Pose-projected Action Recognition Hourglass Network (PARHN) in Soccer

Mehrnaz Fani, Kanav Vats, Chris Dulhanty, David A. Clausi, John Zelek

Current research on soccer action recognition does not focus on player-level actions. We introduce the pose-projected action recognition hourglass network (PARHN) for performing player-level action recognition in soccer. This network is inspired by ARHN, a network originally introduced for hockey action recognition. PARHN has two main novelties in its structure. First, it includes an embedded pose projection component that regularizes the numerical range of the player's pose vector, by applying two separate zero-phase component analysis (ZCA)-whitening. The projected pose vector is effectively learned by few succeeding layers, and significantly improves the performance of the network and its generalization ability. Second, PARHN incorporates the temporal information by having a parallel structure for extracting projected pose vectors from al frames of an input sequence and also by using Long short-term memory (LSTM) layers to integrate the pose vectors across the input frames. Also, a new dataset, named SAR4 (standing for, Soccer Action Recognition for 4 action types), is generated. It includes 1292 video sequences, in which soccer players are tracked and labeled for performing four types of action (i.e., goalkeeping diving, player shooting, receiving pass and giving pass). Introduced network achieves the overall F1-score of 88.1% on the test data, which is 20.1% better than the result of the baseline network, ARHN.

aUToTrack: A Lightweight Object Detection and Tracking System for the SAE AutoDrive Challenge

Keenan Burnett, Sepehr Samavi, Steven L. Waslander, Timothy D. Barfoot, Angela P. Schoellig

The University of Toronto is one of eight teams competing in the SAE AutoDrive Challenge – a competition to develop a self-driving car by 2020. After placing first at the Year 1 challenge [1], we are headed to MCity in June 2019 for the second challenge. There, we will interact with pedestrians, cyclists, and cars. For safe operation, it is critical to have an accurate estimate of the position of all objects surrounding the vehicle. The contributions of this work are twofold: First, we present a new object detection and tracking dataset (UofTPed50), which uses GPS to ground truth the position and velocity of a pedestrian. To our knowledge, a dataset of this type for pedestrians has not been shown in the literature before. Second, we present a lightweight object detection and tracking system (aUToTrack) that uses vision, LIDAR, and GPS/IMU positioning to achieve state-of-the-art performance on the KITTI Object Tracking benchmark. We show that aUToTrack accurately estimates the position and velocity of pedestrians, in real-time, using CPUs only. aUToTrack has been tested in closed-loop experiments on a real self-driving car (seen in Figure 1), and we demonstrate its performance on our dataset.

CIPPRS Dissertation Awards
Committee: Michael Brown, Greg Mori, Anqi Xu

Best Computer Vision Paper Award
Sponsored by Sportlogiq
Committee: Michael Brown, Jim Little, Anqi Xu

Best Robot Vision Paper Award
Committee: Michael Brown, Jean-François Lalonde, Anqi Xu

Best Oral Presentation Award
Jury: Michael Brown, James Clark, Anqi Xu

Best Poster Presentation Award
Jury: Michael Brown, Amir Rosenfeld, Anqi Xu

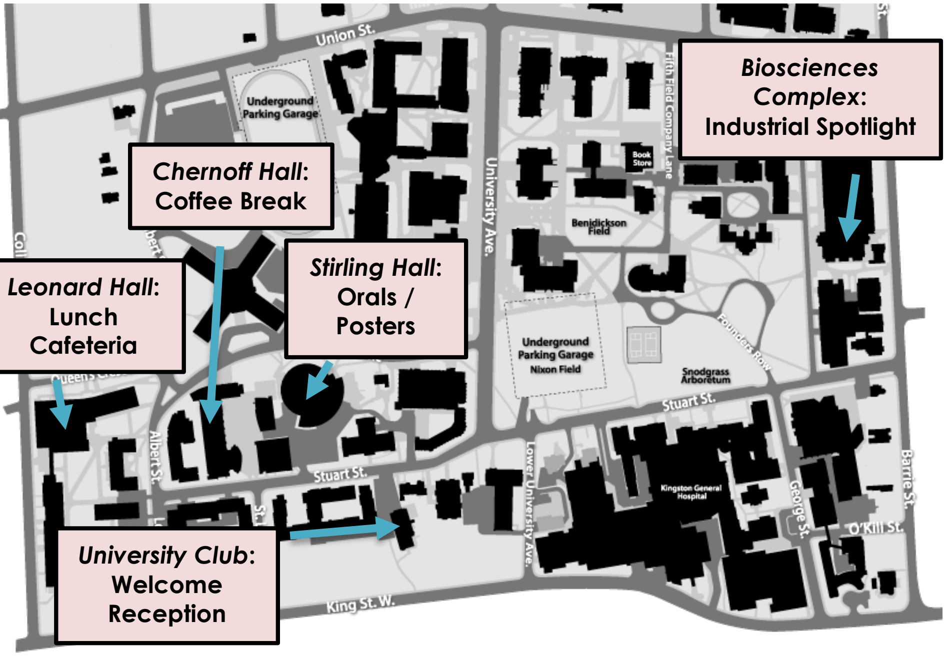
CIPPRS Lifetime Achievement Awards Committee
Michael Brown, Anqi Xu

CIPPRS Executives
Michael Jenkin (President), Steven L. Waslander (Treasurer), Jim Little (Secretary)

Program Committee

Abdelrahman Abdelhamed
Robert Allison
Amjad Almahairi
Alexander Andreopoulos
Daniel Asmar
Ralf Bachmayer
Christine Bakhous
John L. Barron
Robert Bergevin
Guillaume-Alexandre Bilodeau
Michael S. Brown
Daniel Buckstein
Cunjian Chen
Jianhui Chen
Jianning Chi
James Clark
David A. Clausi
Michaël Clément
Jack Collier
Eduardo R. Corral-Soto
Elham Dolatabadi
Marc-Antoine Drouin
James Elder
Mark Eramian
Adel Fakih
Paul Fieguth
Martin Gerdzhev
Philippe Giguère
Minglun Gong
Eric Granger
Michael Greenspan
Mohamed A. Helala
Andrew Hogue
Yannick Hold-Geoffroy
Md Jahidul Islam
Michael Jenkin
Jean-François Lalonde
Jochen Lang
Michael Langer
Jim Little
Fahim Mannan
Mathieu Marquis-Bolduc
David Meger
Gregor Miller
Vida Movahedi
J. Krishna Murthy
Hui-Lee Ooi
Boris Oreshkin
Liam Paull
Marco Pedersoli
Hughes Perreault
Pedro Oliveira Pinheiro
François Pomerleau
Abhijith Punnappurath
Faisal Z. Qureshi
Dhanesh Ramachandram
Mehdi Rezagholizadeh
Morteza Rezanejad
Mrigank Rochan
Junaed Sattar
Amir H. Shabani
Florian Shkurti
Kaleem Siddiqi
Gunho Sohn
Tabish Syed
Graham Taylor
Matthew Toews
Bryan P. Tripp
Alvaro Uribe Quevedo
Carlos Vazquez
David Vazquez
Kedar Vilankar
Ruisheng Wang
Yang Wang
Steven L. Waslander
Alexander Wong
Yijun Xiao
Herb Yang
Linwei Ye
Xin Yi
John Zelek
Maryam Ziaeefard

❖CAMPUS MAP❖



❖CRV 2019 POSTERS❖

Authors Present: **11AM – 12:30PM Thu. May 30th**
(*Stirling Hall – Lobby*)

Optional Viewing: **4PM – 7PM Wed. May 29th**
(*Biosciences Complex – Atrium*)
(limited space; see email)