Mobile Scene Perception via Convolutional Neural Networks

Silvia Ferrari*
Co-PIs: Mark Campbell *, and Kilian Weinberger#
*Sibley School of Mechanical and Aerospace Engineering
#Department of Computer Science
Cornell University
Research Goals

ONR BRC grant N00014-17-1-2175

- **Decentralized perception**: control a team of autonomous agents providing video coverage and situational awareness.

- **Data parsing**: extract agent-level task-relevant data for high-level reasoning.

- **Contested communications**: reason about the scene using asynchronous decentralized video data obtained from different viewpoints and environmental conditions.

- **Active planning**: plan and coordinate agent actions to actively obtain video that is task-relevant and improves scene perception and interpretation.

**Decentralized Video Surveillance via Mobile Camera Network:**
Research Goals

- **Decentralized perception**: control a team of autonomous agents providing video coverage and situational awareness.

- **Data parsing**: extract agent-level task-relevant data for high-level reasoning.

- **Contested communications**: reason about the scene using asynchronous decentralized video data obtained from different viewpoints and environmental conditions.

- **Active planning**: plan and coordinate agent actions to actively obtain video that is task-relevant and improves scene perception and interpretation.

**Decentralized Video Surveillance via Mobile Camera Network:**

- Crowded
- Action Understanding
- Low Visibility
- Unfavorable Illumination
Research Goals

- **Decentralized perception**: control a team of autonomous agents providing video coverage and situational awareness.

- **Data parsing**: extract agent-level task-relevant data for high-level reasoning.

- **Contested communications**: reason about the scene using asynchronous decentralized video data obtained from different viewpoints and environmental conditions.

- **Active planning**: plan and coordinate agent actions to actively obtain video that is task-relevant and improves scene perception and interpretation.

Decentralized Video Surveillance via Mobile Camera Network:

- **Scene Perception**
  - Actors
  - Environment
  - Actions
Research Goals

- Decentralized perception: control a team of autonomous agents providing video coverage and situational awareness.
- Data parsing: extract agent-level task-relevant data for high-level reasoning.
- Contested communications: reason about the scene using asynchronous decentralized video data obtained from different viewpoints and environmental conditions.
- Active planning: plan and coordinate agent actions to actively obtain video that is task-relevant and improves scene perception and interpretation.

Decentralized Video Surveillance via Mobile Camera Network:

- Control
  - Agents
  - Plan
  - Adapt
Background on Computer Vision
Object, Action Recognition Methods

Handcrafted Features

(Marszalek et. al. 2009; Dollár et al. 2005; Reddy and Shah 2013)

- Detect and extract sparse features and statistics of image patches
- Easy implementation; intuitive; scale, translation, rotation, or illumination invariant
- Feature effectiveness is problem-dependent; feature class must be chosen by user

Statistical Models

(Ning et. al. 2008; Natarajan and Nevatia 2008; Zhang and Gong 2010)

- User-crafted models of human actions; parameters learned from data
- Interpretability; compact model of relationships; generate inference and predictions
- Automatic feature selection; feature diversity and richness; unsurpassed performance

Deep Learning

(Simonyan and Zisserman, 2014; Ji et. al. 2013; Singh et. al. 2016; Zhu et. al. 2016)

- Learn important features and object/action classification from data
- Automatic feature selection; feature diversity and richness; unsurpassed performance
- Require large datasets; lack of feature interpretation; output large feature vectors
Scene Perception and Interpretation

Feature Level

- Relies on interest points and surrounding pixels
- Identifies task-relevant invariant image patches
- Lacks semantic and spatial information; little or no predictive ability

Object Level

- Relies on semantic segmentation and object-scene co-occurrence
- Identifies semantic information; segments (parses) image frames
- Lacks depth and texture information; including temporal information requires reconciling pixel and spatial coordinates and shapes

Global Level

- Relies on scene features and entire image/frame
- Identifies spatial and texture information; provides contextual information
- Lacks predictive, generative models

(Derpanis et al. 2012; Fei-fei and Perona 2005; Bosch et al. 2008)


(Greene and Oliva 2009; Oliva and Torralba 2002; Lipson et al. 1997)
Approach: Multi-level Perception
**Subordinate:** object-level detection of task-relevant elements \(\leftrightarrow\) feature level

**Basic:** categorical representation of similar components and their relationships \(\leftrightarrow\) object level

**Superordinate:** highest level of abstraction of the scene environment \(\leftrightarrow\) global level
Closed-loop Virtual Experiments in UE4™

Real-time simulation and control of actors within the environment

- Logical behavior tree
- C++ syntax
- Built-in UE4 functions and classes
Closed-loop Virtual Experiments in UE4™

Real-time simulation and control of actors within the environment

- Logical behavior tree
- C++ syntax
- Built-in UE4 functions and classes

Visual Scripting: Blueprints
Convolutional Feature-level Perception
Resource Aware Re-identification

- Deep Anytime Person Re-ID network [Co-PI Weinberger]
- Combine features across multiple layers using skip connections
- Allows early stop and gives results instead of propagating through the network if a running budget is reached.
Actor Re-ID, Association, and Tracking

- Traditional tracking uses motion info for data association, e.g., position, speed
  - Suffer from difficulties in data association and performance degrades in crowded environments
- Deep CNN Re-ID: integrate convolutional features to improve data association
  - Applicable to other tracking frameworks [Co-PI Campbell]

\[
p(\text{association}) \quad p(\text{location}) \quad p(\text{appearance})
\]

\[
p(a \mid z_{\text{pos}}, z_{\text{re-id}}) \propto p(z_{\text{pos}} \mid a) \times p(z_{\text{re-id}} \mid a) \times p(a)
\]

- Likelihood based on Re-ID embedding

\[
p(z_{\text{re-id}} \mid a) \propto \frac{\exp^{-c(z_{\text{re-id}}, z_a)}}{\sum_j \exp^{-c(z_{\text{re-id}}, z_j)}}
\]

- \(Z_j\) Embedding of identity \(j\)
- \(Z_{re-id}\) Embedding of current identity
Multiple Hypothesis Tracker with Deep CNN Re-ID

- Substantial increase in robustness during crossings, close walking, occlusions
Basic-level Perception and Modeling
GP Modeling of Optical Flow

- Optical flow: input to action recognition classifier or deep CNN
- Predict optical flow in short predictive horizons from input video frames.
- Obtain *Gaussian Process* (GP) model of pixel-based optical flow.
- Predictive distribution of optical flow \( p^* = (p_{x}(T + \Delta T), p_{y}(T + \Delta T)) \) for pixel \( q^* (x, y, T + \Delta T) \)

\[
\begin{align*}
\mathbf{p}^* &\sim \mathcal{N}(\mathbf{\bar{p}}, \Sigma_p + \sigma_n \mathbf{I}) \\
\mathbf{\bar{p}} &= K(q^*, \mathbf{Q})K(\mathbf{Q}, \mathbf{Q})^{-1} \mathbf{P} \\
\Sigma_p &= K(q^*, q^*) - K(q^*, \mathbf{Q})K(\mathbf{Q}, \mathbf{Q})^{-1}K(\mathbf{Q}, q^*)
\end{align*}
\]

 Kernel function

\[
\mathbf{P} = \{(p_{x}(t), p_{y}(t))\}_{x \in X, y \in Y, t=0:T}
\]

where

\[
\mathbf{Q} = \{q(x, y, t)\}_{x \in X, y \in Y, t=0:T}
\]
Ego-motion Subtraction

- Extract motion model of target represented as the optical flow (OF) \( \mathbf{p}_T = [p_x, p_y] \) using a moving camera.
- Subtract induced optical flow caused by camera motion \( \mathbf{p}_T = \mathbf{p} - \mathbf{p}_A \).

Compute the camera motion-induced optical flow \( \mathbf{p}_A \) given the camera focal length \( \lambda \), velocity \( \mathbf{V}_c = [\dot{X}, \dot{Y}, \dot{Z}] \), rotation speed \( [\dot{\psi}, \dot{\theta}, \dot{\phi}] \), and the distance from the camera focus \( z \):

\[
\mathbf{p}_A = H \begin{bmatrix} R_{\psi} & R_{\theta} & R_{\phi} & 0 \\ 0 & 0 & 0 & w_T \end{bmatrix} \begin{bmatrix} -\mathbf{V}_c \\ \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix}
\]

where

\[
H = \begin{bmatrix} \frac{\lambda}{z} & 0 & -\frac{q_x}{z} & \frac{q_x q_y}{\lambda} & -\frac{q_x^2 + \lambda^2}{\lambda} & q_y \\ 0 & \frac{\lambda}{z} & -\frac{q_y}{z} & \frac{q_x q_y}{\lambda} & \frac{q_y^2 + \lambda^2}{\lambda} & -q_x \\ \end{bmatrix}
\]

\[
w_T = \begin{bmatrix} 1 & 0 & -\sin(\theta) \\ 0 & \cos(\phi) & \sin(\phi) \cos(\theta) \\ 0 & -\sin(\psi) & \cos(\phi) \cos(\theta) \end{bmatrix}
\]

\[
R_{\psi} = \begin{bmatrix} \cos(\psi) & \sin(\psi) & 0 \\ -\sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad R_{\theta} = \begin{bmatrix} \cos(\theta) & 0 & -\sin(\theta) \\ 0 & 1 & 0 \\ \sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \quad R_{\phi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\phi) & \sin(\phi) \\ 0 & -\sin(\phi) & \cos(\phi) \end{bmatrix}
\]
Optical Flow Prediction

Actual Optical Flow  GP Model Predicted Flow
Active Mobile Perception
Goal: obtain high-quality frames of task-relevant actor (person) via mobile camera

Unicycle (Segway) robot kinematics:

\[
\dot{q}(t) = \begin{bmatrix}
\cos \theta(t) & 0 \\
\sin \theta(t) & 0 \\
0 & 1
\end{bmatrix} \begin{bmatrix}
v(t) \\
\omega(t)
\end{bmatrix} = G(q(t))u(t)
\]

CNN detects and IDs actor → Labeled bounding box

Control law, \( u(t) \), drives CNN bounding box to match desired actor bounding box in camera pixel coordinates

Desired Target
Position in Image

Processing Error Controller Input

Set of Potential Detections and Probabilities

Trained Deep NN Image Camera Sensor

Agent Kinematic Model

Camera/Agent Pose
Mobile Perception Results: Unreal Engine™
Mobile Perception Experiments
Ongoing and Future Work: Decentralized Perception, Identification, and Tracking

- Three cameras (2 fixed, 1 mobile) with different viewpoints, orientations, and scale
Future Work and Acknowledgements

**Future Work:**
- Deep CNN robustness guarantees
- Global and multi-level scene representation and reasoning
- Decentralized active camera control
- Performance analysis under variable network topologies

**Collaborators:**
- **Thomas A. Wettergren**, Ph.D.  
  Naval Undersea Warfare Center  
  Newport, RI
- **Jason Isaacs**, Ph.D.  
  Naval Surface Warfare Center  
  Panama City, FL

This research was funded by the ONR Decentralized Perception BRC Program, PMs: Behzad Kamgar-Parsi and Marc Steinberg  
Grant # N00014-17-1-2175
Cornell Project Team

* Sibley School of Mechanical and Aerospace Engineering and
# Department of Computer Science

Silvia Ferrari*
Professor

Mark Campbell*
Professor

Kilian Weinberger#
Associate Professor

Chang Liu*
Postdoctoral Associate

Matthew Davidow
Ph.D. student
Cornell Center for Applied Math

Jake Gemerek*
Ph.D. student

Brian Wang*
Ph.D. student

Lequn Wang#
Ph.D. student

Yan Wang#
Ph.D. student
Questions?

Thank you