COLLEGE OF ENGINEERING
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Two Parts:
1. Video Analysis for Body Worn Cameras
2. Future of Datasets in Computer Vision

Jason J. Corso
Associate Professor
Electrical Engineering and Computer Science
University of Michigan

Joint work with various others, mentioned when appropriate.

http://web.eecs.umich.edu/~jjcorso jjcorso@eecs.umich.edu

September 29, 2015 @ NITRD Video and Image Analytics Working Group

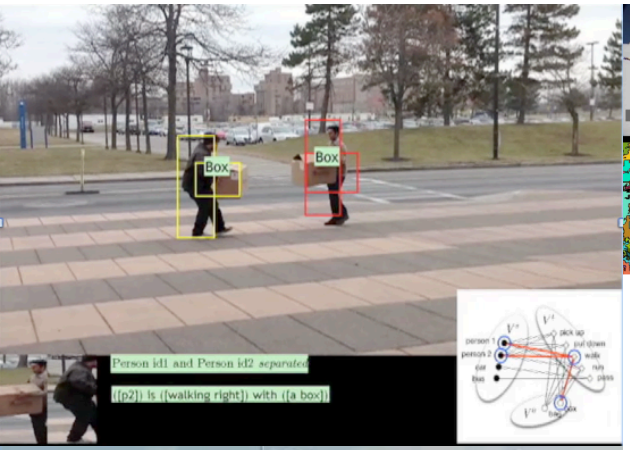
Jason Corso

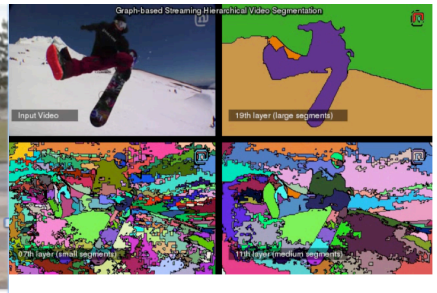
Associate Professor
Electrical Engineering and Computer Science
University of Michigan

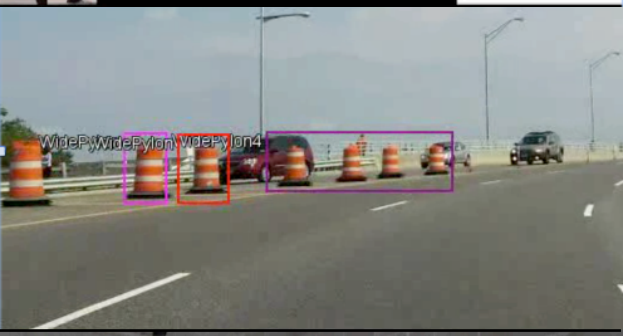
Primary Areas: **Computer Vision and Robotics**
Secondary Areas: Data Science and Machine Learning

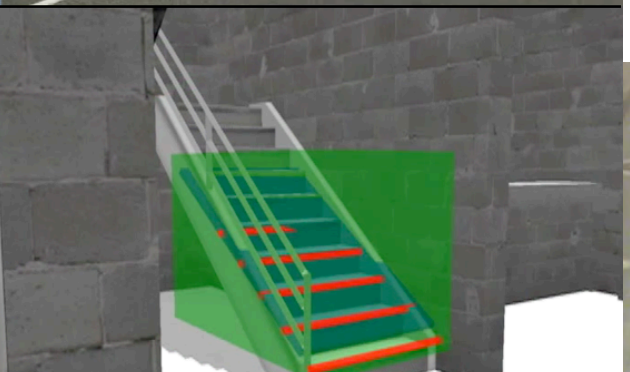
Biography and Distinctions

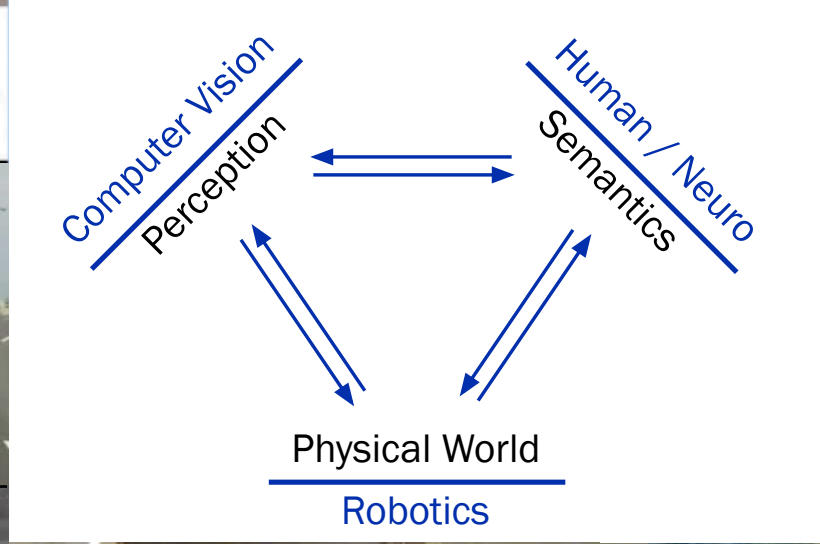
- 2006 Ph.D. Johns Hopkins (CS)
- 2008 NSF CAREER award
- 2009 DARPA CSSG award
- 2010 ARO Young Investigator award
- 2015 Google Faculty award
- Assc Editor for TPAMI and IJCV

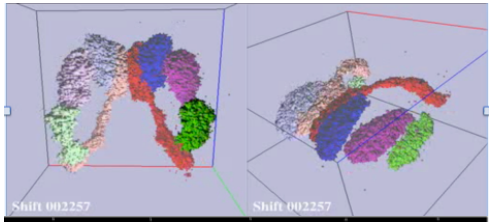


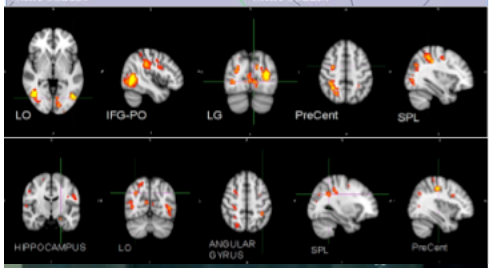


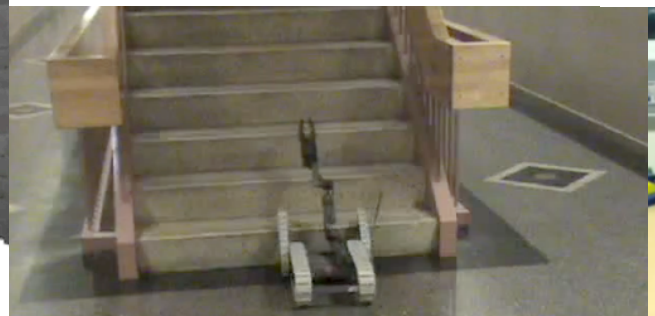












Jason Corso

Associate Professor
Electrical Engineering and Computer Science
University of Michigan

- Research Group (started in 2007)
  - Current: 9 PhD Students, 2 MSE Students, 2 Undergraduates, 1 Post-Doc, 1 Engineer
  - Past Peak: 2 Post-doctoral scholars and 1 visiting scholars, 12 PhD Students
- Computing Power
  - Dual Unix Servers (8-Core 3.4Ghz Xeon, 8GB memory)
  - 24+ Workstations (Varying power, max 24-core, 48GB memory)
  - 10+ Laptops (Varying power for on-road experimentation)
  - 25TB SAN
  - Integration with campus FLUX cluster
- Video Sensors
  - 4 VGA (640x480x3 30Hz Bayer), 1 VGA with programmable zoom
  - 2 SOC Videre Stereo cameras
  - 1 XGA with active lighting
  - 6 Asus Xtions, 3 Microsoft Kinects, 2 laser scanners
  - A bushel of web-cameras
- Robots
  - 6 Turtlebots from ClearPath Robotics.
  - 1 Aldebaran NAOs Humanoid Biped.
  - 1 Kuka YouBot (mobile manipulators).
  - 2 ClearPath Husky.
  - 3 Home-Built Grasping Arms.

Funded Projects As Faculty

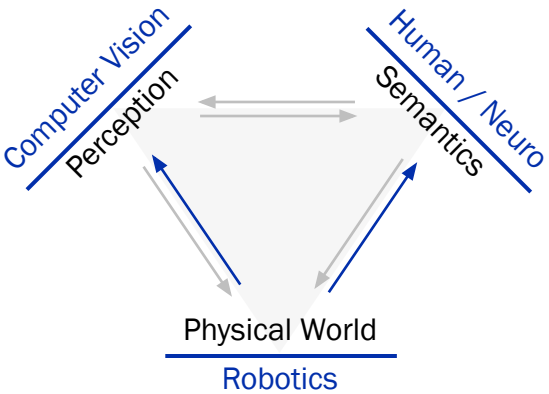
DARPA MINDSEYE (PI)	$2.3M
IARPA ALADDIN (Sub-PI)	$1.0M
DARPA CSSG (PI)	$900K
NSF CRI x2 (PI)	$740K
NSF NRI (PI)	$650K
DARPA STTR (Sub-PI)	$602K
NSF CAREER (PI)	$540K
CIA/IC PF (PI)	$357K
ARO Core (PI)	$290K
FHWA EAR (Sub-PI)	$260K
ARO DURIP (PI)	$250K
NAVY POSTGRAD (PI)	$190K
NIH NCR (Sub-PI)	$164K
ARO YIP (PI)	$150K
HP IRP (Co-I)	$130K
Google (PI)	$65K
ARL TARDEC (PI)	$25K
	$8.42M

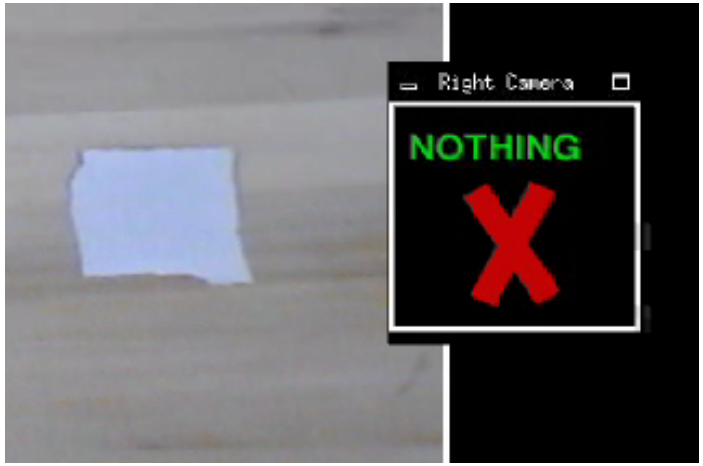
UNIVERSITY OF MICHIGAN

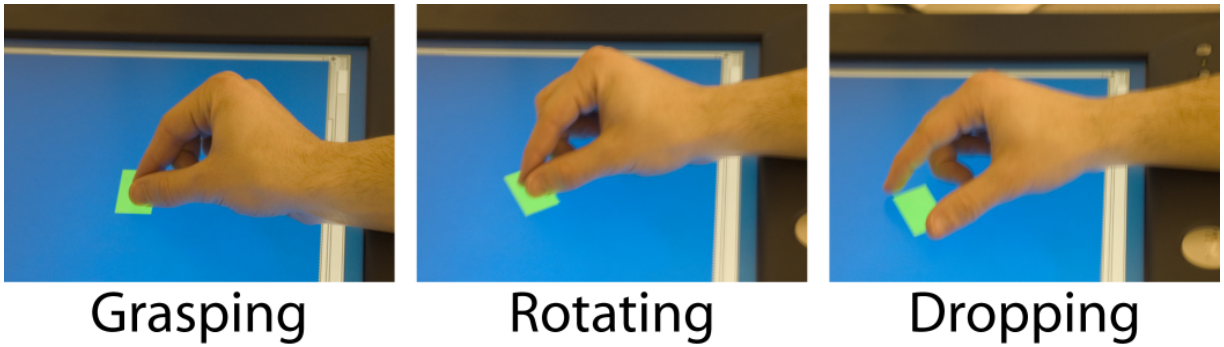
Jason Corso

Associate Professor
Electrical Engineering and Computer Science
University of Michigan

Vision-Based Human-Computer Interaction In Shared Perceptual-Physical Workspaces



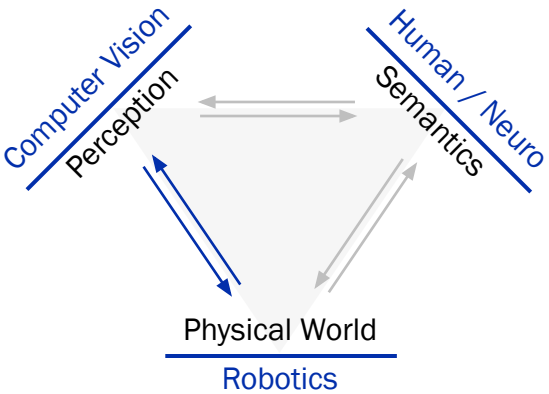




Generated a Language of Interaction.

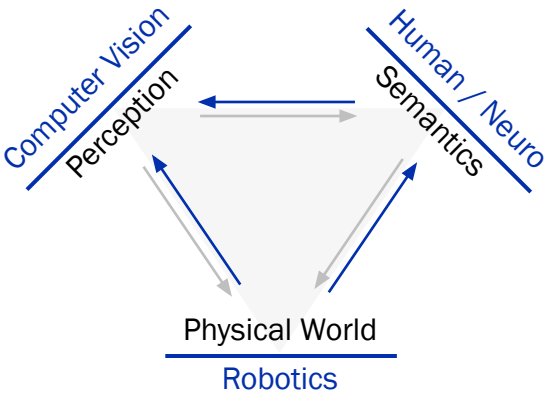
[Corso et al., CVPR 2003, VR 2004]

Physically-Grounded Robot Perception Enabling New Robot Behaviors



[Delmerico and Corso, IROS 2012, ICRA 2013]

Language-Grounded Computer Vision



[Das, …, Corso, WSDM 2013, CVPR 2013]

Video Analysis for Body Worn Cameras

Discussion of a whitepaper that grew out of a CCC-sponsored working group at CVPR 2015.

With: Alex Alahi, Kristen Grauman, Greg Hager, Louis-Philippe Morency, Harpreet Sawhney, and Yaser Sheikh

Body-Worn Cameras: The Potential

- Transparency
  - Increase public trust and confidence in the police.
- Protection
  - Protect officers from false allegations.
  - Positively influence behavior of officer and those being recorded.
- Investigative
  - Cameras supplement officers' recall and document events.
- Training
  - Recorded real-life situations will aid in educating both green and experience officers.

Body-Worn Cameras: Technology Drivers

- **Redaction**
  - Cited as one of the most urgent needs for police departments that are adopting bodycams.
  - Traditional redaction, such as blurring faces, only a first step.
  - Subtler information, such as a logo, a tattoo, furniture, may also need to be redacted, depending on context.
- **Freedom of Information Act servicing**
  - Redaction aside, FOIA requests can include various open-ended queries such as time of day, number of officers present, etc.
  - Current video indexing tools do not meet the semantic richness such FOIA queries require.

- **Forensic search and triage**
  - Abilities to index, search and triage large repositories of body camera video footage will be a critical forensic capability.
  - Different levels of specificity.
  - Incorporate video, audio and multimodal aspects to search.
  - Geospatial and temporal localization.
- **Training systems**
  - Curation of videos suitable for use in training scenarios.
- **Early warning systems**
  - Monitor officer behavior to detect early warning signs, such as premature use of force.
  - Currently, the officer largely self-reports this information.

Body-Worn Cameras: Technology Enablers

- **Computer vision recognition community has made huge strides in recent years.**
  - Discriminable tasks like sporting events and face detection.
  - Works less well in open-ended tasks, such as FOIA servicing, when various criteria are beyond system capabilities.
  - Work in egocentric vision, although normally slower paced everyday-type activities.
  - Summarizing long, first-person videos into a shorter video.
  - Customized compression schemes for audio.
- **Challenges**
  - Bodycam video from law enforcement will be more challenging: video will be shaky, fast motion, occlusions.
  - Hard to evaluate summaries of long-videos.
  - Limited work in fusing audio and video signals for bodycams.
  - Practical challenges of storage, battery, etc.

Body-Worn Cameras: Cross-Cutting Challenges

- Closed and proprietary versus open and standardized
  - Current bodycam acquisition and storage is closed and based on proprietary platforms.
  - Access is controlled through proprietary interfaces defined by vendors based on their needs and goals.
  - Means it is difficult to pull and share data from these systems.
  - **Need to cultivate an open ecosystem of development around these platforms.**
- Bootstrap development with curated video
  - Data drives development in experimental research communities.
  - Sources of such curated bodycam video are not currently known.
  - Multimodal labeling correlating audio and video is critical to success.

Body-Worn Cameras: Timelines

- 2-Year Timeline
  - Refinement of existing technologies to the specific task.
  - Privacy filters in video and audio.
  - Simple summarization and indexing.
- 5-Year Timeline
  - Detection of certain classes of entities fully automatically with high fidelity. No human verification needed.
  - Modeling and queries on complex events involving many parts.
  - Interconnecting front-end officer and back-end HQ.
- 10-Year Timeline
  - Real-time redaction and indexing.
  - Full situation awareness.
  - Large scale indexing, combining visual elements and language elements for the query.

Body-Worn Cameras: Policy Recommendations

- **Usage Protocol**
  - Recommended best practices developed and distributed.
  - When to turn on devices, how much history to buffer, narration guidelines, debriefing guidelines, and a clear explanation of how the data will be used.
- **Public Education**
  - Recommend developing a plan for educating the public and journalists on how conclusions can be drawn from the data.
  - E.g., it is **never** possible to guarantee the camera viewpoint is that of the officer or the officer's attention.
    - There is hence no good reason to limit the capture viewpoint.

Body-Worn Cameras: Technology Recommendations

- **Multimodal Sensing**
  - Recommend providing minimal sensor guidelines for audio, visual and metadata sensors.
  - Recommend the following technologies
    - Stereo pair of wide field-of-view, high-resolution cameras
    - Microphones with sufficient dynamic range for human speech
    - Inertial measurement unit
    - GPS
    - Timestamp all sensors to GPS-locked clock.
- **Media Central**
  - Recommend developing a central medial facility for use by police departments across the country for storage and analysis of the data.
  - Secure, with state of the art cloud tools for indexing and searching the video.
  - Attention to data quality, e.g., compression, is critical.

Body-Worn Cameras: Technology Recommendations

- **Indexing**
  - Data should be indexed against the state of the art visual and audio indexing technologies.
  - Original data should be stored indefinitely for later re-indexing and analysis when better technologies become available.
- **Open Standards**
  - Open, community-driven standards for data representation in video and audio are paramount to establishing an industry around improving the use of body-worn camera video.
  - Minimal technical barriers to export, access and exchange data.

Body-Worn Cameras: Research Recommendations

- **Standards, Datasets and Benchmarks**
  - Open standards and established datasets and benchmarks move toward a high-level of consistency to ensure good data is available across jurisdictions.
  - Cultivate an ecosystem of innovation around bodycams.
  - Dataset Desiderata
    - Thorough annotation
    - Size: covers many scenarios
    - Similar characteristics to end-game scenarios.
- **Research Funding**
  - Community policing initiative provides financial support for the acquisition of body-worn cameras and storage.
  - It does not account for the very many unsolved questions we have discussed, nor does it account for the expected high cost of new personnel to manage and make use of the data.
  - Hence, both basic and applied research funding are needed.

Body-Worn Cameras: Research Recommendations

- **Technology Transition**
  - Bringing research prototypes to a level of readiness for field study requires further investment in time and money.
  - Open standards would reduce such transition burden.
- **Continued Involvement Among Video Processing Research Community**
  - A mechanism for establishing involvement among research community is important.
    - Workshops, working groups and committees are possibilities.
    - Standards committee is important.
  - Research community not well-suited for the transition-efforts.
  - However, long-term 10-20 year vision requires more basic research funding along these lines.
  - Potential for establishing a Center of Excellence in Video Analysis and Analytics for Law Enforcement.

Future of Datasets in Computer Vision

Discussion of an NSF-funded CRI Seedling and the recent workshop held at CVPR 2015.

With: Kate Saenko

Datasets drive progress

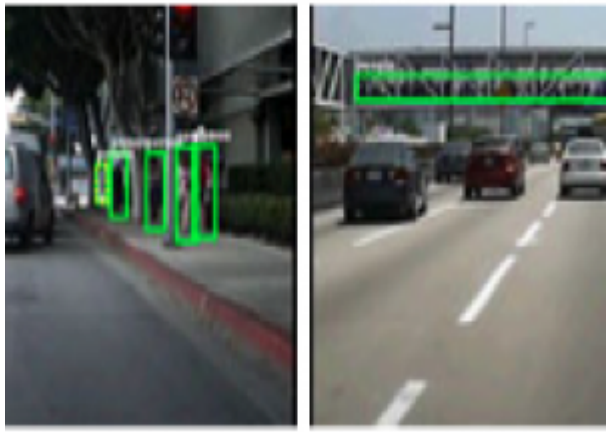
- Many recent advances in computer vision have been driven by **labeled data.**

- PASCAL VOC, Caltech 101/256, ImageNet, LabelMe, SUN, TrecVID-MED, HMDB51, NYU RGB-D, MS COCO, just to name a few...

Dataset Life Cycle

1. Researcher decided to tackle new problem
2. A dataset is born
3. Publish dataset with 20% accuracy results
4. A race ensues…
5. Teams achieve 90%+ accuracy
6. Death/rebirth

YouCook



Caltech Pedestrian Detection

CompCars: 163 car makes with 1,716 car models

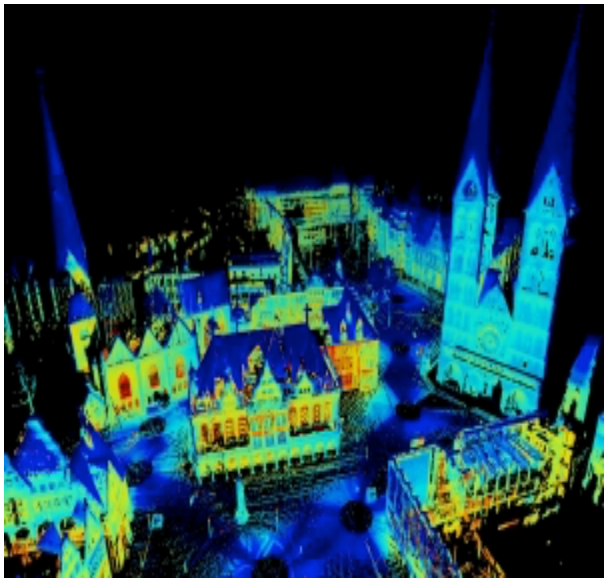
MTFL: Multi-Task Facial Landmark: 12,995 face images with landmarks

DogCentric Activity Dataset: first-person videos from a camera mounted on a dog.

UCF50: Action Recognition in Realistic Videos

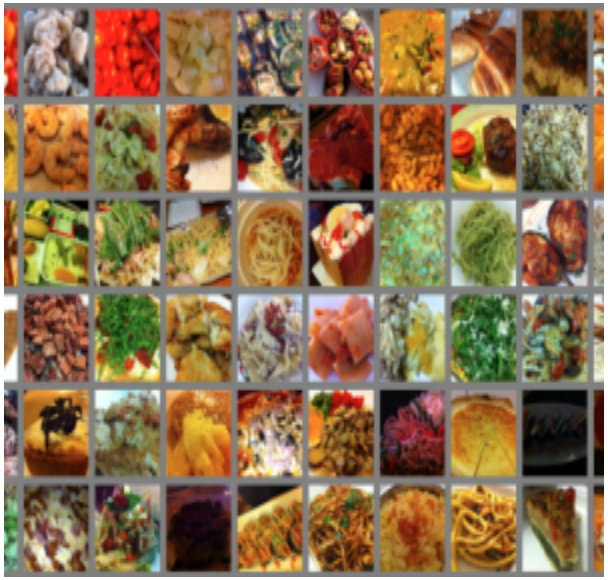
[Facial Expressions in the Wild (SFEW / AFEW)]

[3DPES - PEople Surveillance Dataset]



[Robotic 3D Scan Repository]

[FlickrLogos-32]



[UNICT-FD889]: 889 distinct plates of food.

[PASCAL-Context Dataset]: augments PASCAL VOC10 dataset with 400+ additional categories.

CV Datasets on the web

COMPUTER VISION
O . N . L . I . N . E
YOUR GATEWAY TO THE COMPUTER VISION WORLD

Yet Another Computer Vision Index To Datasets (YACVID)

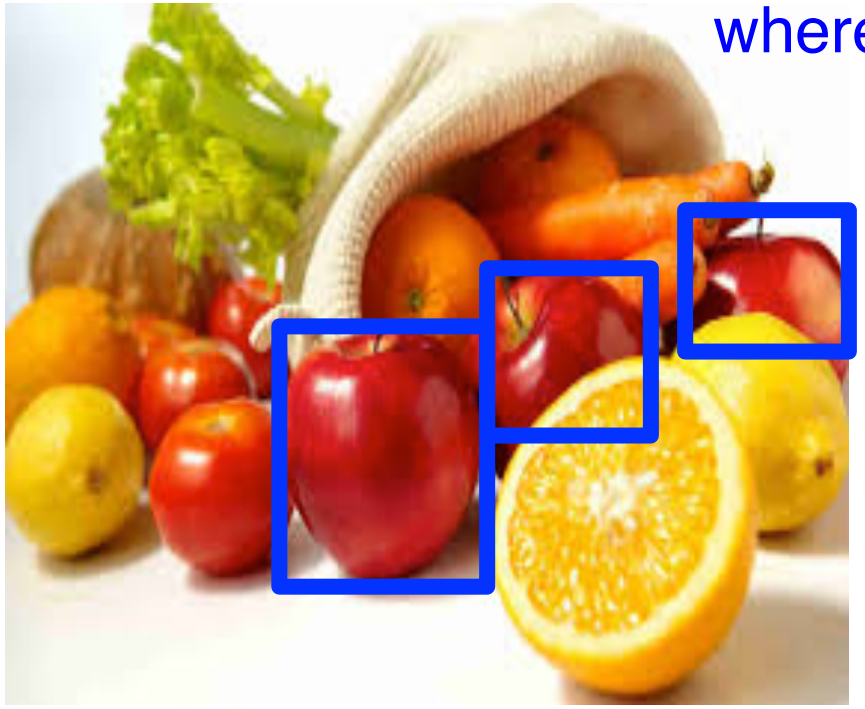
What are the problems with datasets?

Too small

To continue progress in new tasks we must have much more data.

Most address one aspect

Particular but arbitrary view of the broader image/video understanding problem.



where is the apple?

Disconnected

- No central database with common format.
- Only good for isolated tasks.
- No mechanism for mapping across problems and across datasets to understand and measure progress in broad visual understanding.

Biased

- selection bias
- capture bias
- semantic bias

Federated Data Set Infrastructure for Recognition Problems in Computer Vision

Original CRI-New NSF Proposal

T. Berg (UNC), J. Corso (SUNY Buffalo), T. Darrell (UCB), A. Efros (UCB), J. Hockenmaier (UIUC), F.-F. Li (Stanford), J. Malik (UCB), K. Saenko (UMass Lowell), and A. Torralba (MIT)

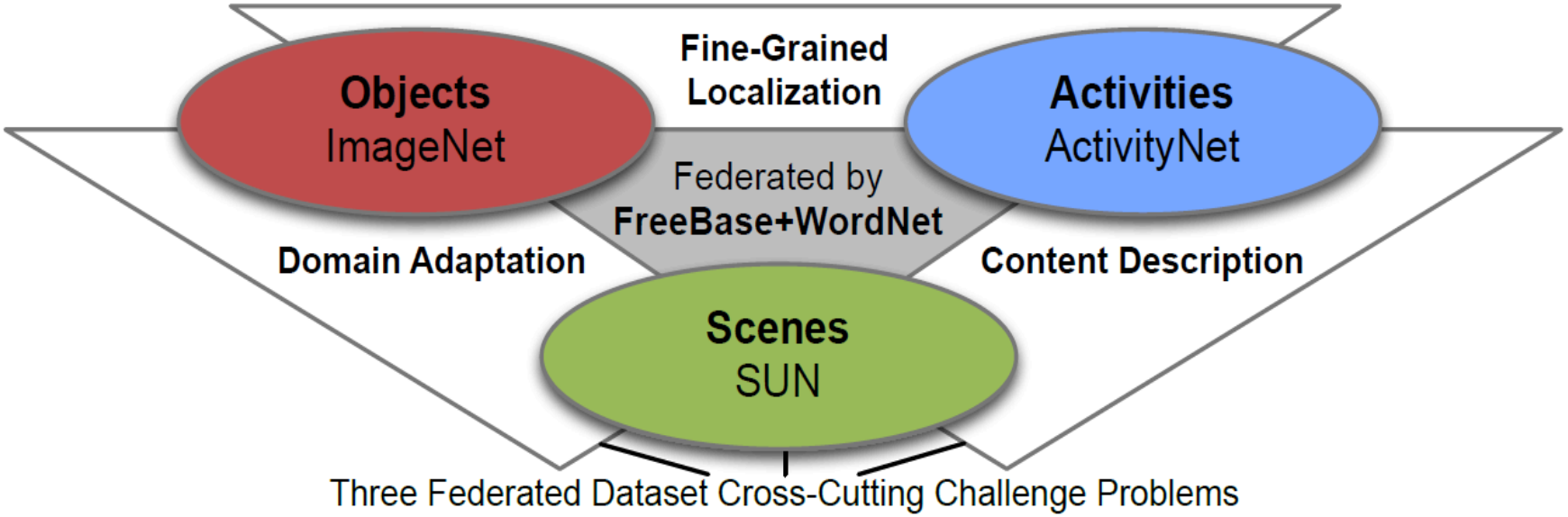
Objectives

- Federate datasets in a **single infrastructure**.
- Map entities to a **common semantic namespace** to allow meaningful translation and cross-pollination.
- For and by the **community**.

Proposed infrastructure

- **software APIs** for curating and accessing vision data
- **database** maps entities within each dataset to a common semantic space (WordNet)
- **crowd-sourcing APIs** to gather annotations from coarse level labels to fine-grained annotations

Three core recognition problems



Outcome of NSF proposal

- NSF gave funding to organize a workshop to solicit community feedback
  - how would community use this infrastructure?
  - what research would it enable?
  - what is the response to a prototype?

- We held that workshop at CVPR 2015
  - Invited Speakers: JC Niebles, M Shah, S Pradhan, L Zitnick.
  - 20 posters, from 26 institutions, 10 countries and 84 individuals.
  - And, we received valuable feedback!

    Through a distributed questionnaire at the workshop.



What is the Future?

UNIVERSITY OF MICHIGAN

COVE: Computer Vision Exchange
of Data, Annotations and Tools

- Workshop Guidance
  - Overwhelming recommendation to focus on a shared and community-driven infrastructure for data storage, annotation representation, and tools to manipulate these.
- Plans and Community Involvement
  - Now working with the Computer Vision Foundation to have them host COVE on cv-foundation.org
  - Prototype implementation of the first version of COVE
    - Dataset browser and annotation translator
    - Query for datasets/annotations by constraints
    - Plan to ingest as many datasets as possible.