CLASP Program Review

Institution:

- Boston University

Team Members:

- Advisors: Stan Sclaroff, Venkatesh Saligrama
- Post-doc: Hanxiao Wang

Contact:

- sclaroff@bu.edu, srv@bu.edu, hxw@bu.edu

This material is based upon work supported by the U.S. Department of Homeland security under Award Number 2013-ST-061-E001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

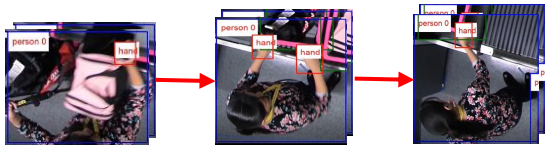
System Summary

Detection

- locate persons, bins, and hands

Tracking

- assign a single ID to bounding boxes of the same object



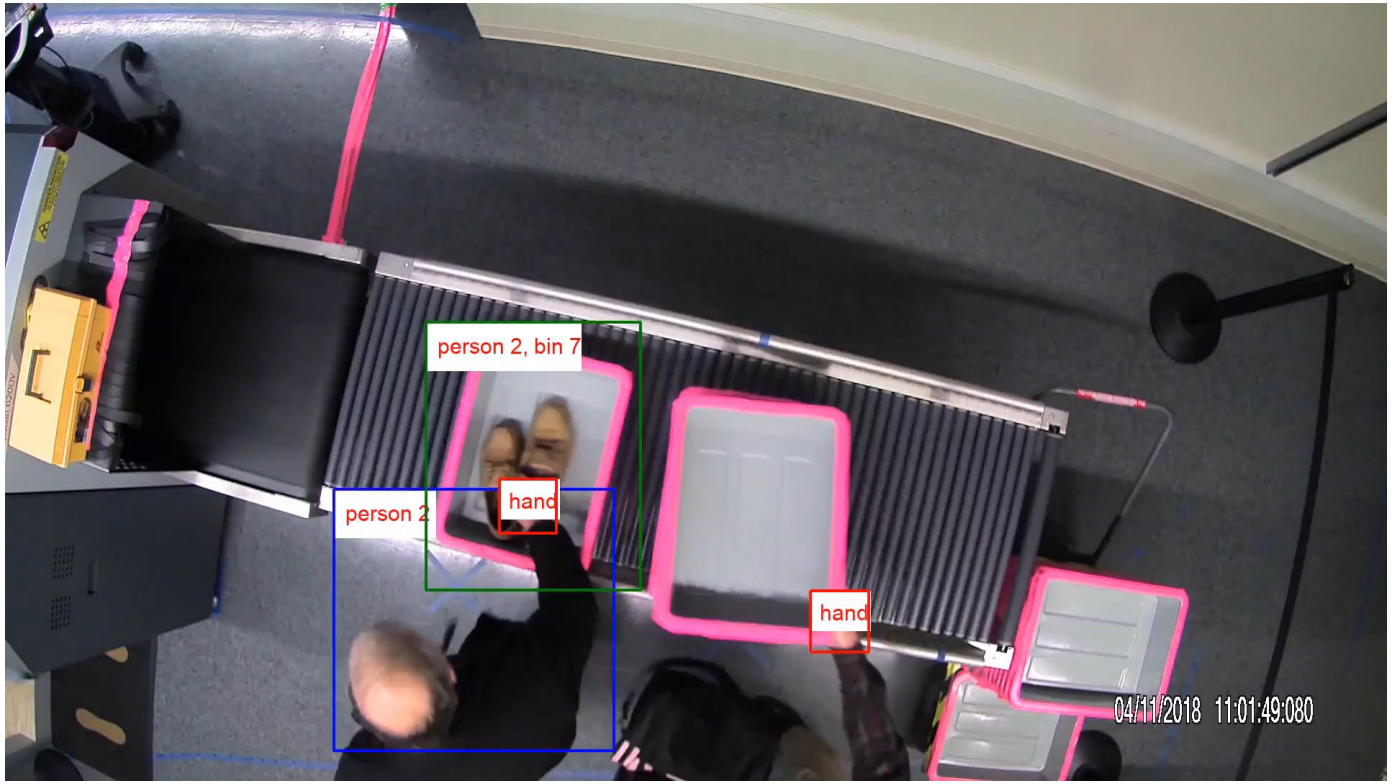
Association

- associate bins to corresponding persons

Event-Detection

- detect divesting/collection events
- verify ownership

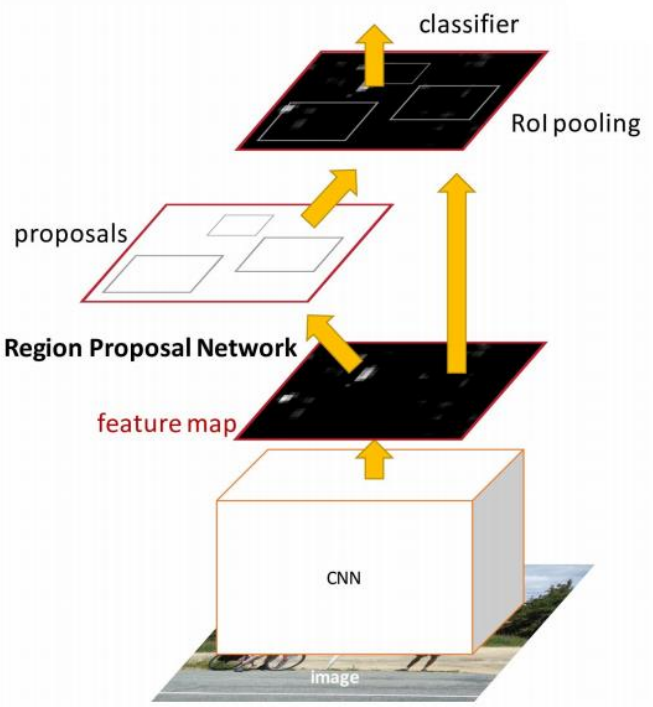
Demo Video



30 seconds video clip from exp5a, cam9

Detection

Faster-RCNN: state-of-art object detector by deep learning



- Feature network:
  - We chose ResNet-101
- Region proposal network:
  - We kept top 300 proposals per image
- Classifier:
  - Classify proposals into person/bin/hand
- Speed:
  - 106 ms per frame

Ren et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NIPS 2015

Detector Training

Detector:

- Faster-RCNN (state-of-art object detector by deep learning)

Pre-training:

- MS COCO (83K training images, 80 classes)

Training data for persons and bins:

- bin & person annotations from NEU, Purdue and Alert (1411 images, 2 classes)

Training data for hands:

- hand dataset released by Zisserman *et al* (13050 hand instances)

Zisserman *et al, "*Hand detection using multiple proposals", BMVC 2011

Tracking in a Single Camera

t=0 t=10 t=20

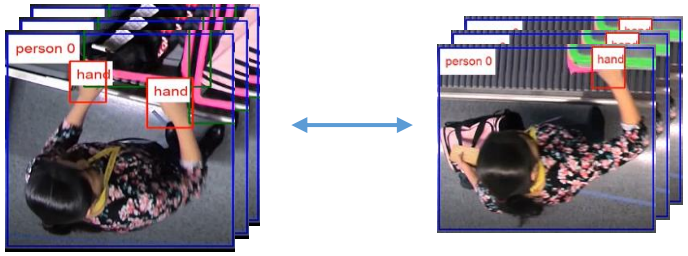
Three types of information used:

- Location (x, y, h, w)
- Speed (vx, vy)
- Visual Appearance (deep feature)

Method:

1. Speed and location model: Kalman Filter
2. Appearance model: Inception-V3 network trained on Market-1501 dataset (12936 images, 751 persons)
3. Linear assignment by Hungarian algorithm, with cost matrix given by above models

Tracking Across Multiple Cameras



Camera 9 Camera 11

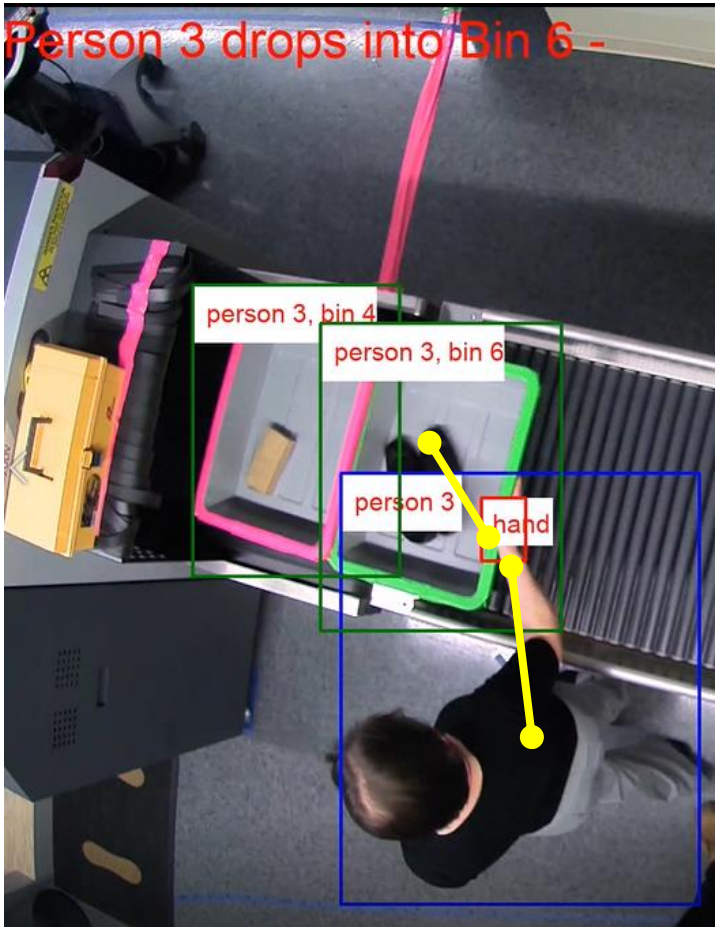
- Only appearance is used
- Time/Location information

  NOT considered

Method:

1. Appearance model: Inception-V3 network trained on Market-1501 dataset
2. Linear assignment by Hungarian algorithm, with cost matrix given by the above model

Association and Event-Detection



Distance-based association

- associate bins to nearest hands
- associate hands to nearest persons
- associate objects with people

An event is triggered when

- A hand box fully appears in a bin box

Results

↑ Higher is better

↓ Lower is better

	Exp 5a Results		Exp 5b Results	
	cam 9	all cams (2,5,9,11,13)	cam 9	all cams (2,5,9,11,13)
PD (PAX) ↑	96.3% (26/27)	95.0% (96/101)	100.0% (20/20)	90.8% (69/76)
PD (DVI) ↑	87.5 % (35/40)	90.7% (117/129)	77.1% (27/35)	78.1% (75/96)
PD (XFR) ↑	77.8% (14/18)	59.0% (23/39)	62.5% (10/16)	62.2% (23/37)
PFA (PAX) ↓	16.4% (6/41)	17.4% (39/207)	25.0% (10/41)	16.3% (33/204)
PFA (DVI) ↓	2.4% (1/41)	3.9% (8/207)	20.0% (8/41)	12.8% (26/204)
PFA (XFR) ↓	0.0% (0/4141)	0.0% (6/20705)	0.1% (5/4071)	0.1% (11/20355)
P (PAX switch) ↓	7.4% (2/27)	35.6% (36/101)	5.0% (1/20)	27.6% (21/76)
P (DVI switch) ↓	12.5% (5/40)	39.5% (51/129)	0.0% (0/35)	14.6% (14/96)
P (mismatch) ↓	0.0% (0/18)	23.1% (9/39)	0.0% (0/16)	24.3% (9/37)

Results

Higher is better

Lower is better

	Exp 5a Results		Exp 5b Results	
	cam 9	all cams (2,5,9,11,13)	cam 9	all cams (2,5,9,11,13)
PD (PAX) ↑	96.3% (26/27)	95.0% (96/101)	100.0% (20/20)	90.8% (69/76)
PD (DVI) ↑	87.5 % (35/40)	90.7% (117/129)	77.1% (27/35)	78.1% (75/96)
PD (XFR) ↑	77.8% (14/18)	59.0% (23/39)	62.5% (10/16)	62.2% (23/37)
PFA (PAX) ↓	16.4% (6/41)	17.4% (39/207)	25.0% (10/41)	16.3% (33/204)
PFA (DVI) ↓	2.4% (1/41)	3.9% (8/207)	20.0% (8/41)	12.8% (26/204)
PFA (XFR) ↓	0.0% (0/4141)	0.0% (6/20705)	0.1% (5/4071)	0.1% (11/20355)
P (PAX switch) ↓	7.4% (2/27)	35.6% (36/101)	5.0% (1/20)	27.6% (21/76)
P (DVI switch) ↓	12.5% (5/40)	39.5% (51/129)	0.0% (0/35)	14.6% (14/96)
P (mismatch) ↓	0.0% (0/18)	23.1% (9/39)	0.0% (0/16)	24.3% (9/37)

Results

Higher is better

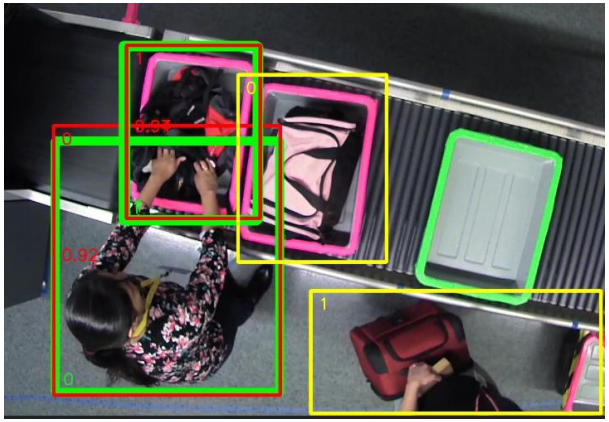
Lower is better

	Exp 5a Results		Exp 5b Results	
	cam 9	all cams (2,5,9,11,13)	cam 9	all cams (2,5,9,11,13)
PD (PAX) ↑	96.3% (26/27)	95.0% (96/101)	100.0% (20/20)	90.8% (69/76)
PD (DVI) ↑	87.5 % (35/40)	90.7% (117/129)	77.1% (27/35)	78.1% (75/96)
PD (XFR) ↑	77.8% (14/18)	59.0% (23/39)	62.5% (10/16)	62.2% (23/37)
PFA (PAX) ↓	16.4% (6/41)	17.4% (39/207)	25.0% (10/41)	16.3% (33/204)
PFA (DVI) ↓	2.4% (1/41)	3.9% (8/207)	20.0% (8/41)	12.8% (26/204)
PFA (XFR) ↓	0.0% (0/4141)	0.0% (6/20705)	0.1% (5/4071)	0.1% (11/20355)
P (PAX switch) ↓	7.4% (2/27)	35.6% (36/101)	5.0% (1/20)	27.6% (21/76)
P (DVI switch) ↓	12.5% (5/40)	39.5% (51/129)	0.0% (0/35)	14.6% (14/96)
P (mismatch) ↓	0.0% (0/18)	23.1% (9/39)	0.0% (0/16)	24.3% (9/37)

Improvements in future

In ground truth annotations:

- Label all visible bins and persons (to reduce PFA)
- Label empty bins



In algorithm:

- To reduce P (switch)
  - Include time information for cross camera tracking
  - Fine-tune appearance model
- To increase PD(XFR)
  - Include more information (e.g. motion, time, content check) other than hand for event detection