A Computationally Efficient Pipeline for Camera-based Indoor Person Tracking

Andrew Tzer-Yeu Chen, Jerry Fan, Morteza Biglari-Abhari, Kevin I-Kai Wang
Embedded Systems Research Group, Dept of Electrical and Computer Engineering
The University of Auckland, New Zealand

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

IVCNZ December 2017

Person Tracking

Localising people over time

Targeting indoor market research applications

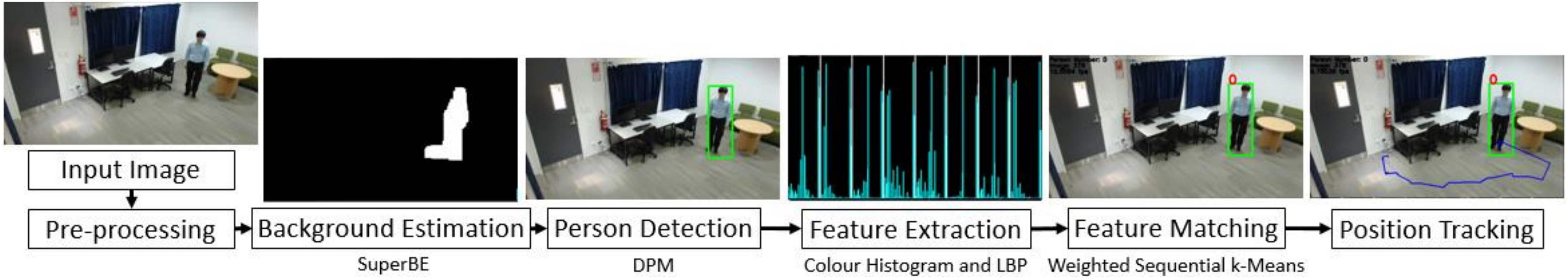
- How many customers are there?
- What parts of the shop are they in?
- Are they going past high value items?
- Do we have enough staff?

Primary Challenges

Computational Efficiency
- Most approaches too slow for real-time
- Balance accuracy vs speed trade-off
- Hard constraint in embedded systems

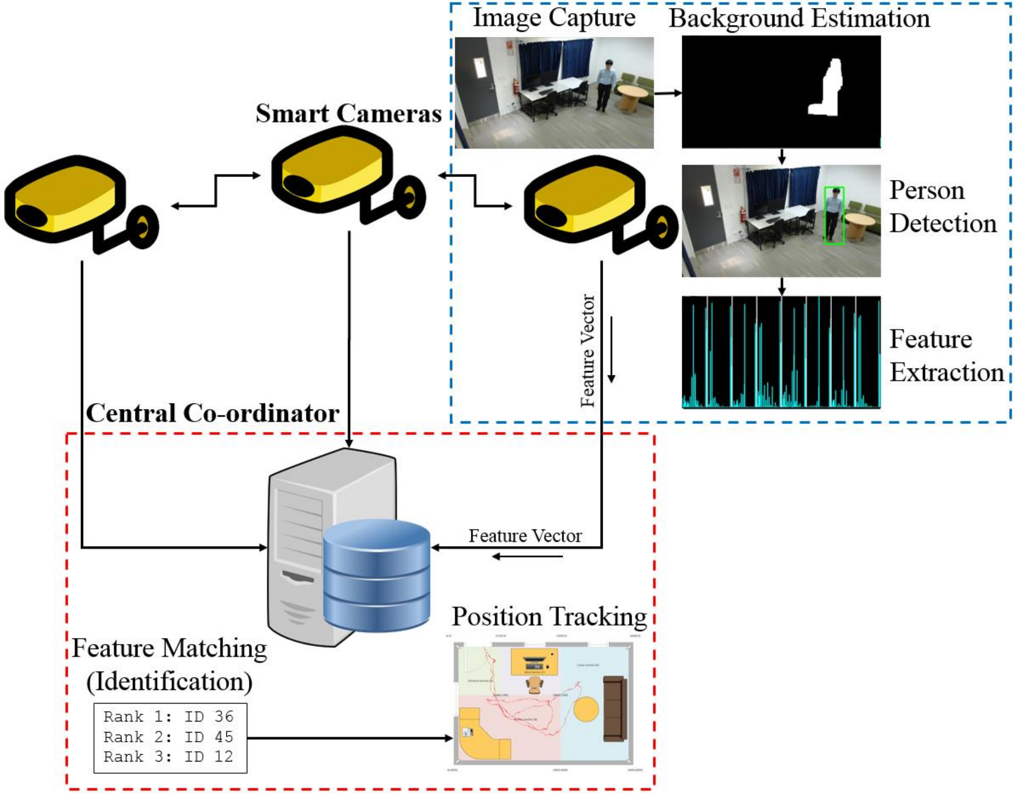
Unsupervised Learning
- Person re-identification across cameras
- Lack of training data at execution time
- Often low-quality cameras, low detail

Person Tracking Pipeline



- Plug-and-Play approach instead of end-to-end CNN
- Modular approach better for distributed processing
- Reduced memory and bandwidth requirements
- Supports privacy-affirming framework:
  potentially, no human ever sees the raw footage

A Distributed Privacy-Affirming Architecture



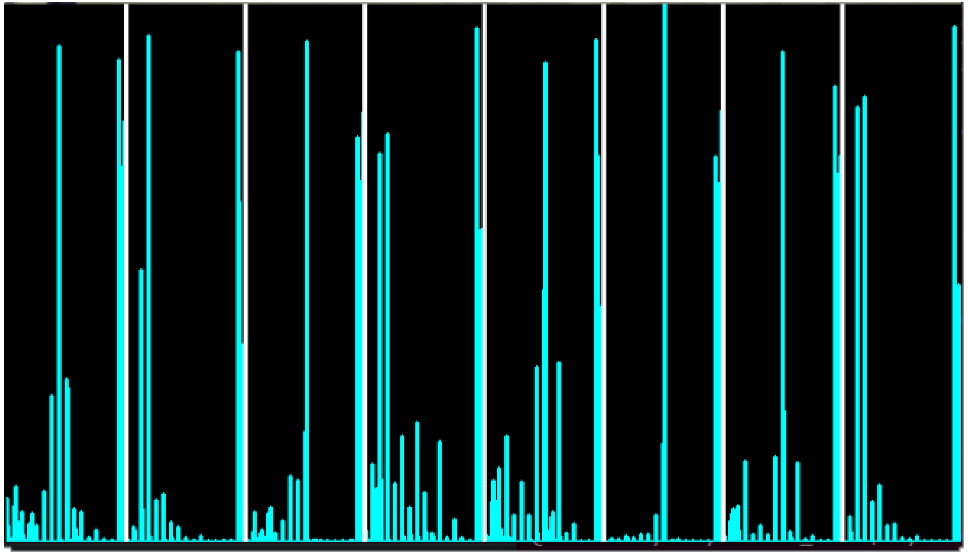
Background Estimation: SuperBE

- Superpixel-based Background Estimation
- Isolate foreground as region of interest
- Minimise unnecessary processing later in pipeline

Person Detection: DPM

- Deformable Parts Model
- Isolates out head, torso, arms, legs
- Helps us deal with obscured parts of the body

Feature Extraction



- Extract colour and texture features for each part
- Use HSL for a cylindrical colour space
- Use LBP for computationally efficient texture descriptor

Feature Matching

- Unsupervised sequential k-means model

1) Similarity and Classification

Use correlation distance to match histograms

Determine similarity of each part, ignoring obscured

Weight colour and texture features

$$S = \alpha_1 \frac{\sum_p d_{col_p} v_p}{\sum_p v_p} + \alpha_2 \frac{\sum_p d_{lbp_p} v_p}{\sum_p v_p}$$

Determine class with highest similarity

Feature Matching

- Unsupervised sequential k-means model

2) Model Update

We only maintain a class mean, not an entire cluster

Modify the class mean with the new sample

$$\mathbf{m}_c = \beta\mathbf{x} + (1-\beta)\mathbf{m}_c$$

Newer samples better indication of current person appearance than older samples

Need some robustness against noise/false positives

Feature Matching

- Unsupervised sequential k-means model

3) Feature Weighting

Try to minimise inter-class similarities
Try to maximise inter-class variation
Help achieve better discriminability

Suppress the common background
Exaggerate the different foreground
Weight the values in the feature vectors

Feature Matching

- Unsupervised sequential k-means model

3) Feature Weighting

$$d_{n_k} = |\mathbf{x}_k - \mathbf{m}_{n_k}|$$

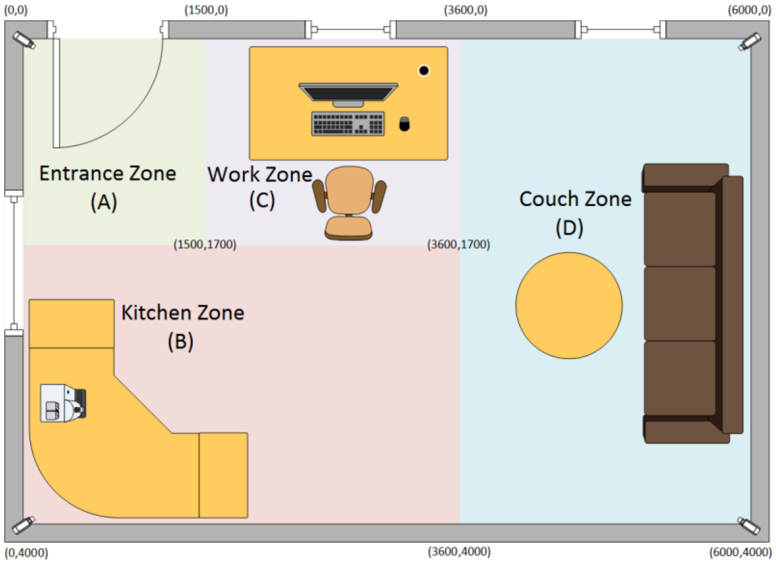
$$\mathbf{w}_k = sat(\mathbf{w}_k + \eta(d_{n_k} - T))$$

As each sample comes in, modify the weight vector
Apply weight vector during classification
Unsupervised learning, improves over time

Feature Matching and Position Tracking

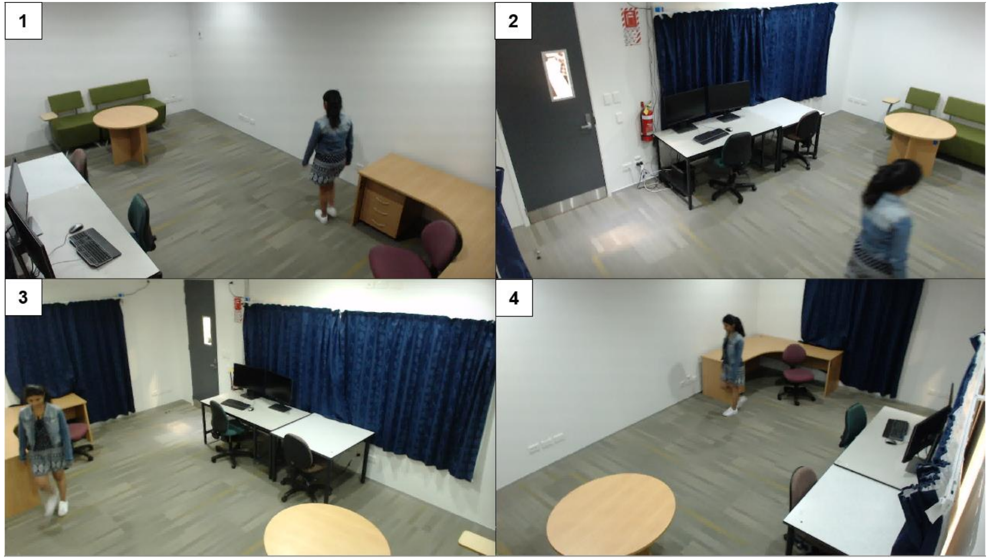
- Assign the class ID to the detected person
- Detect position by taking midpoint between feet parts
- Form a track over time by connecting position points

Dataset



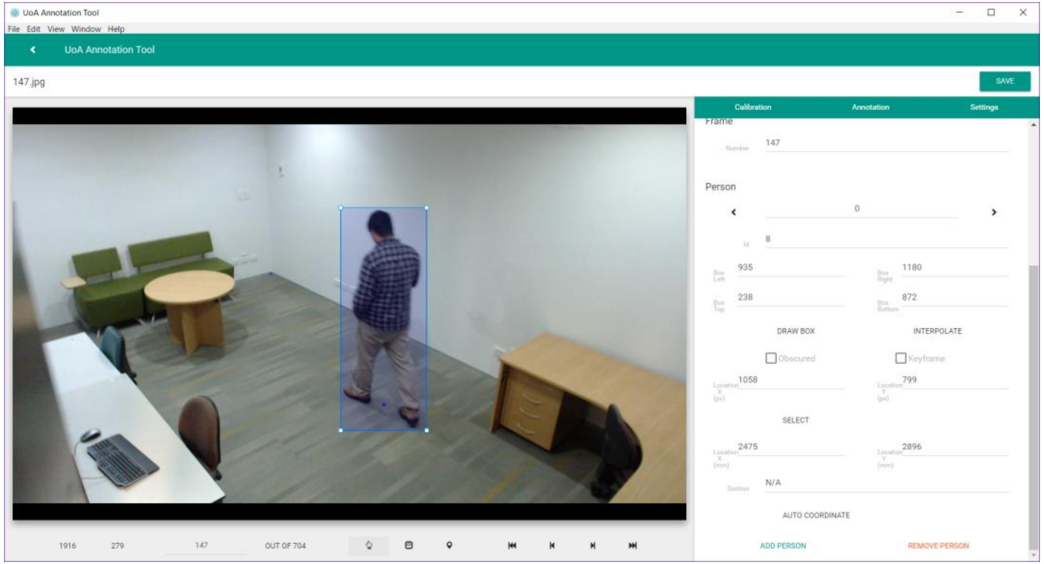
- Create a test environment, similar to an office space
- Humans approximate zones, not exact co-ordinates
- Loosens requirements on precision, better comparison

Dataset



- Four cameras, partly overlapping, different angles
- Cheap webcams to emulate real-world video capture
- Seven action categories: walking, sitting, groups, etc.

Annotation Tool



- Uses homographies for real-world co-ordinates and zones
- Use optical flow to predict boxes, reduce annotation time
- Annotation tool released as open-source on Github

Preliminary Results

TABLE I
SINGLE-CAMERA RESULTS USING IDENTITY-BASED MEASURES

Cam	Detections	IDP (Precision)	IDR (Recall)	IDF_1 (F-Score)
1	502	71.12	15.19	25.03
2	1277	93.50	49.24	64.51
3	1250	87.92	38.68	53.73
4	432	43.98	8.96	14.88
All	3461	82.25	35.84	49.14

- New identity-based metrics for person tracking
- Furniture blocking cameras significantly drops accuracy
- Comparable to state-of-the-art re-identification: 60-80%

Preliminary Results

TABLE II
AVERAGE COMPUTATION TIME (MS) BASED ON THE NUMBER OF PEOPLE DETECTED IN THE FRAME AND THOSE WITH FEATURES STORED

# of People Detected	# of People In Model	Background Estimation	Person Detection	Feature Extraction	Feature Matching
0	0	17.3	2.9	0.0	0.0
1	1	18.1	46.3	0.6	1.8
1	2	19.0	48.9	0.7	1.9
1	3	18.3	48.7	0.5	2.5
2	3	18.6	95.7	1.1	2.1
3	6	19.1	106.1	1.7	4.2
4	10	20.8	148.9	2.0	5.1

- Person detection is the bottleneck
- Each additional person requires approx. 50ms more
- Pipeline processes video with between 5-10fps

Summary

1. Computationally efficient person tracking
2. New online unsupervised learning approach to feature matching in real-time
3. Development of dataset and annotation tool
4. Development of pipelined system architecture to support future work

Future Work

1. Development of retail test scenario
2. Comparison with other re-identification and classification methods
3. Combine multiple camera views to localise position with high accuracy
4. Implementing distributed image processing architecture with smart cameras

Contact: andrew.chen@auckland.ac.nz