2021 embedded VISION summit
VIRTUAL | MAY 25-28

Person Re-Identification and Tracking at the Edge: Challenges and Techniques

Morteza Biglari-Abhari
Department of Electrical, Computer and Software Engineering
University of Auckland, New Zealand

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

Outline

- Why Person Re-Identification and Tracking
- Key Challenges and Current Approaches
- Appearance Based One-shot / Unsupervised Re-Identification
- Spatio-Temporal Based Tracking
- Fused Appearance and Spatio-Temporal Approach
- Privacy Issues
- Summary and Conclusions

Why Person Re-Identification and Tracking

The aim is matching images of people as viewed through multiple cameras in different positions and locations and determine a unique identity.

Possible target applications:

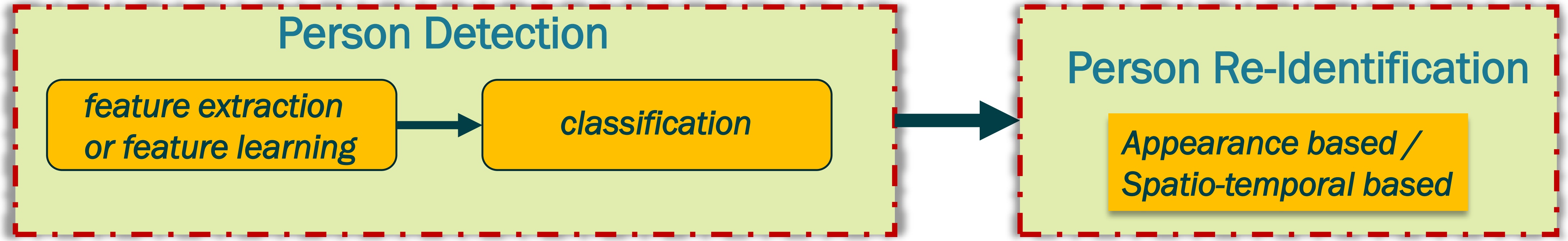
- Surveillance for Security and Public Safety
- Healthcare and Industrial Facilities
- Commercial Entities (such as supermarkets) to monitor customer behavior
- Intelligent Transportation System
- Smart Cities

Key Challenges

Challenges: variations in the appearance of a person (even in the same camera view)

(variations in pose, lighting, color, resolution, motion blur, obstacles, occlusions)

Current Approaches



- ***Person detection*** methods should be robust to detect people in different conditions.
- A ***person model*** needs to be robust against various conditions: varying lighting conditions, partially obscured views, different camera view angles

Person Detection Techniques

- Early person detection works relied on using ***blob detection*** [Krumm et al, 2000 – Everingham & Zisserman, 2006]
  - Low computational complexity but low accuracy
- ***Histogram of Oriented Gradients (HOG)*** and ***Support Vector Machine (SVM)*** algorithm [Krumm et al, 2000 – Dalal & Trigs, 2005]
- ***Deformable Parts Module (DPM)*** – uses HOG features but includes structural relationship between parts of the person [Cho et al, 2012 – Yan et al, 2014]
  - Calculating HOG features is very computationally expensive
  - Using background estimation can improve the accuracy
- ***Aggregate Channel Features (ACF)*** – improves detection speed through isolating features that have the largest contribution towards accurate person detection (focusing on gradient magnitude, HOG, and the LUV color channel). It may be slightly less accurate but faster than DPM. [Dollar et al, 2014 – De Smedt & Goedeme, 2015]

DPM versus ACF

An example of person parts being extracted using DPM

An example of ACF pedestrian detection [Benfold & Reid, 2011]

CNN-based Techniques for Person Detection

CNN-based Techniques:

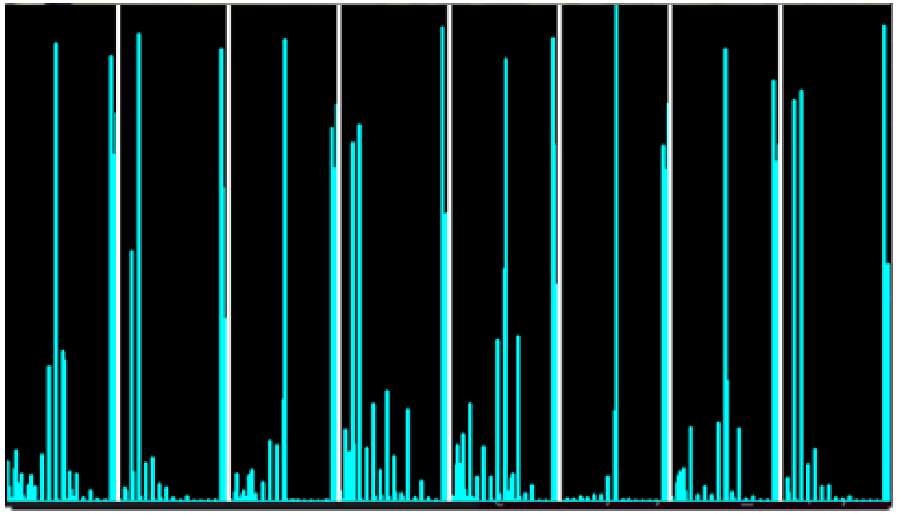
- ***R-CNN:*** regions of interest (ROI) are extracted that potentially have targets for further analysis and fed to CNN for feature extraction and classification. To improve the processing speed, Fast R-CNN and Faster R-CNN have been proposed. [Girshick, 2015 – Ren et al, 2017]
  - While this was faster than other CNN-based techniques, it could process 5 fps using a high-end GPU.
- ***Single Shot Detector (SSD):*** the image is only parsed once rather than processing multiple potentially overlapping windows. Achieving similar level of accuracy to Faster R-CNN but takes less processing time. YOLO is in this category.

- Pre-trained **ResNet-50** and **MobileNet-V2** have been used for person detection and re-identification (for specific datasets such as CUHK03 and DukeMTMC), but they are computationally very intensive for real-time detection on edge devices.

Feature Vectors (Person Detection)

The aim is to select features that allow for high inter-class variation (significantly different between multiple people) while maintaining low intra-class variation (similar for the same person).

- Features may include color (RGB, HSV, YCbCr), texture, and structure.
- Descriptors that include both color and texture perform better than either one alone. [Gou et al, 2017]

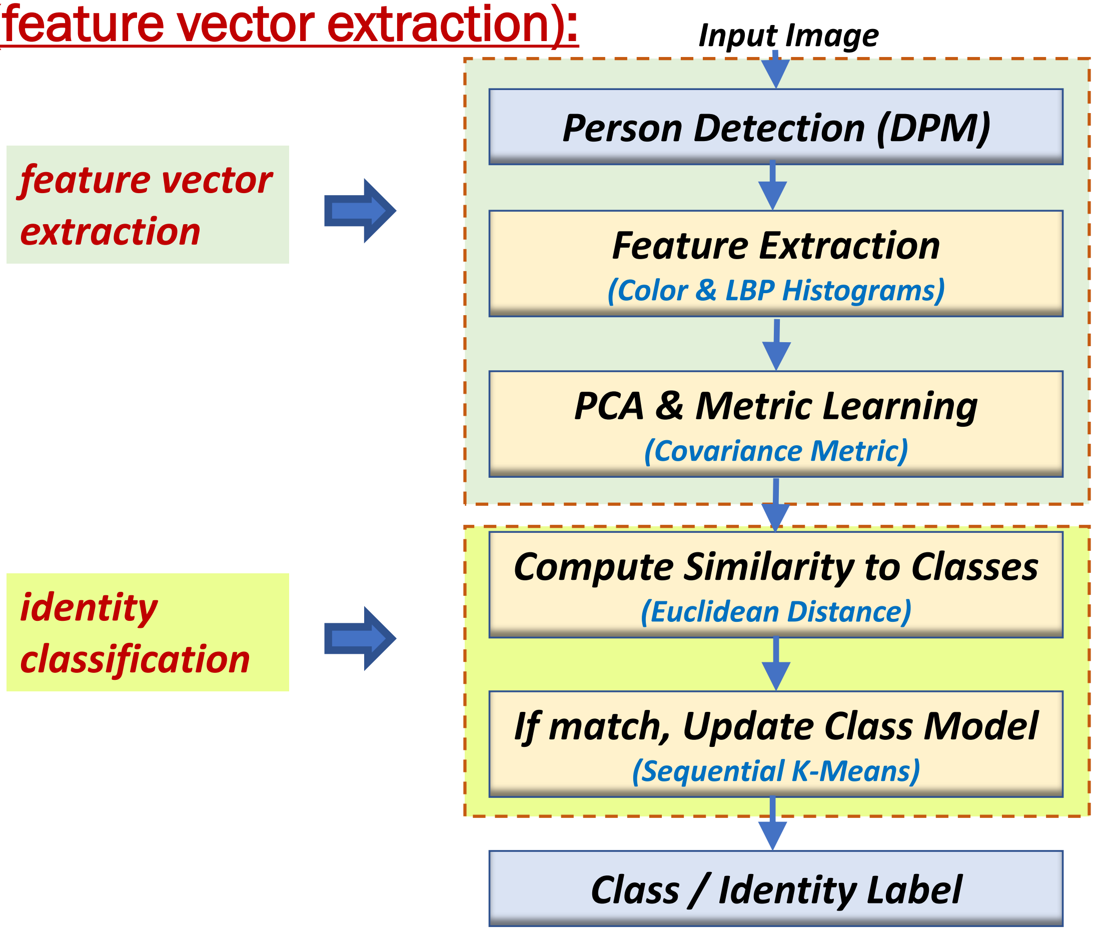


A visual representation of an example feature vector (made up of HSV color and LBP texture histograms) representing an entire person.

Appearance Based Re-Identification

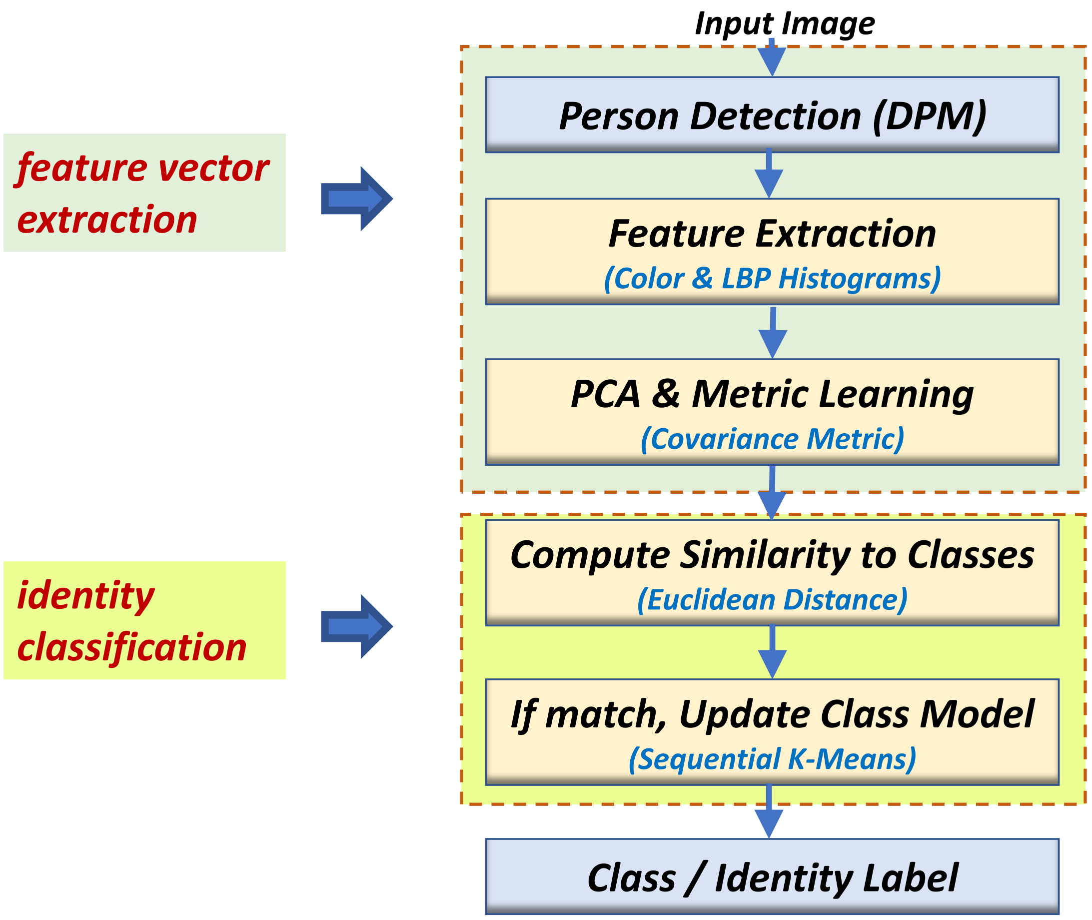
Fast one-shot/unsupervised re-identification (feature vector extraction):

- A combination of HSV for color and LBP (Local Binary Pattern) for texture are used to represent patches or parts of the detected people.
- Principal Components Analysis (PCA) can be used as an unsupervised method of determining the most important dimensions of the feature vectors in terms of variation
- Metric learning as a useful pre-processing step transforms the vectors so that they are more linearly separable into identity classes



Appearance Based Re-Identification (continued)

- Metric Learning reduces the computational complexity and improves the accuracy.
- Covariance Metric transformation used in this case study.



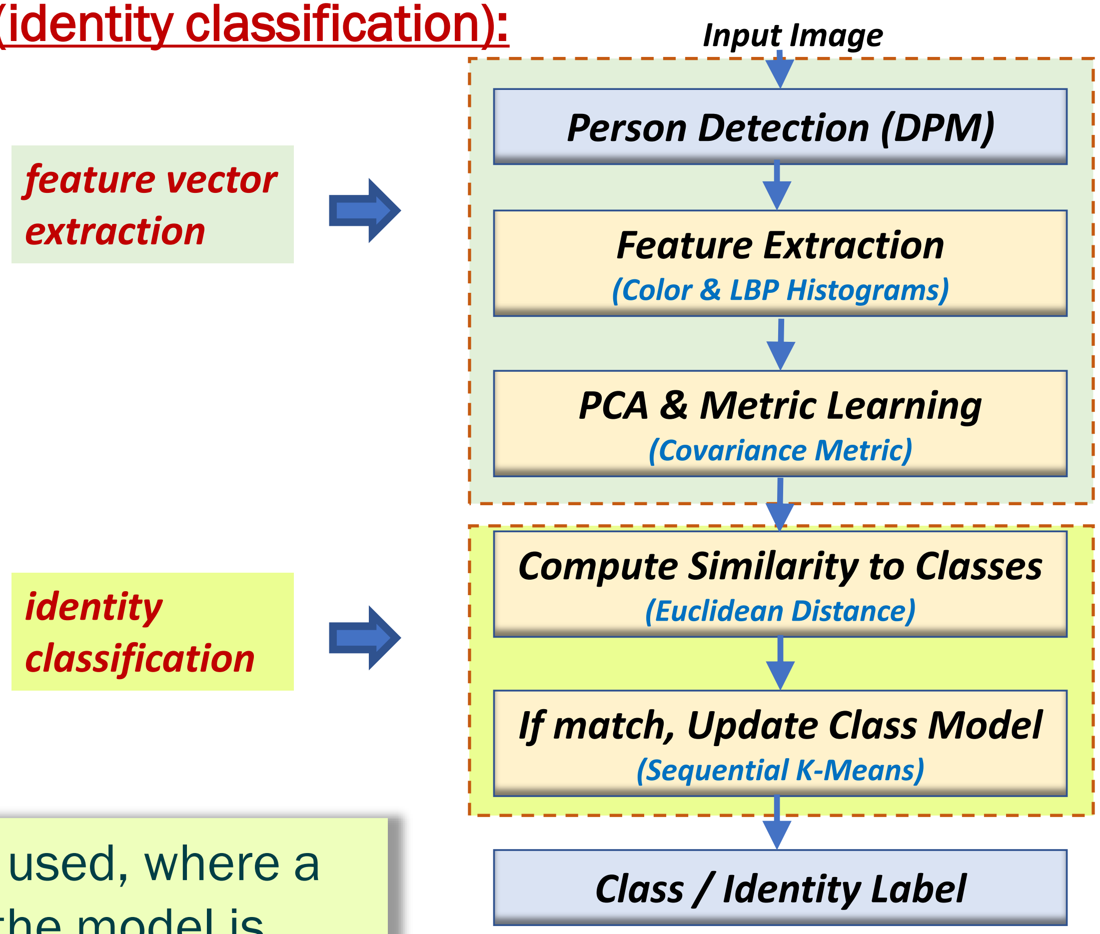
$$\Sigma_{ij} = \mu[X_i\, X_j] - \mu_i\, \mu_j$$

Σ is the output matrix, X is the input feature vector, μ is the mean, and i and j refer to the positions of elements within the vector/matrix.

Appearance Based Re-Identification (continued)

Fast one-shot/unsupervised re-identification (identity classification):

- Each class represents a single identity, and the aim is to classify the transformed feature vectors into classes
- Supervised learning requiring large training data may not be suitable in applications where a possible individual may enter an unconstrained camera view
- Unsupervised learning may not be suitable due to very poor accuracy



As a compromise, **one-shot learning methods** may be used, where a single sample (per class) is used during training and the model is updated at run-time.

Reducing the impact of misclassification in one-shot/unsupervised learning

Gallery Approach:

A *gallery* of N feature vectors is maintained for each identity class:
- Create a new class for a new person and use the extracted feature as an anchor
- Establish a gallery of *N* samples (initially all identical) for each class

Two main parts: Classification step and Model Update step

Classification

- Compare the new sample (probe) with each target sample in the gallery in each class (i.e. calculate the Euclidean distance)
- If the distance is below a specified threshold, then they match
- If the number of matches is more than a specified *numMin*, then the new sample is classified as part of that identity class

Model Update

- A random target sample in the class gallery is replaced with the classified probe sample.
- Constraint on *N* makes the model stable and *numMin* reduces the impact of mis-classification

Reducing the impact of misclassification in one-shot/unsupervised learning (continued)

Sequential k-Means Approach:

A modified from of ***k-Means clustering*** that supports online learning to classify feature vectors is used. Each class/cluster is a new person.

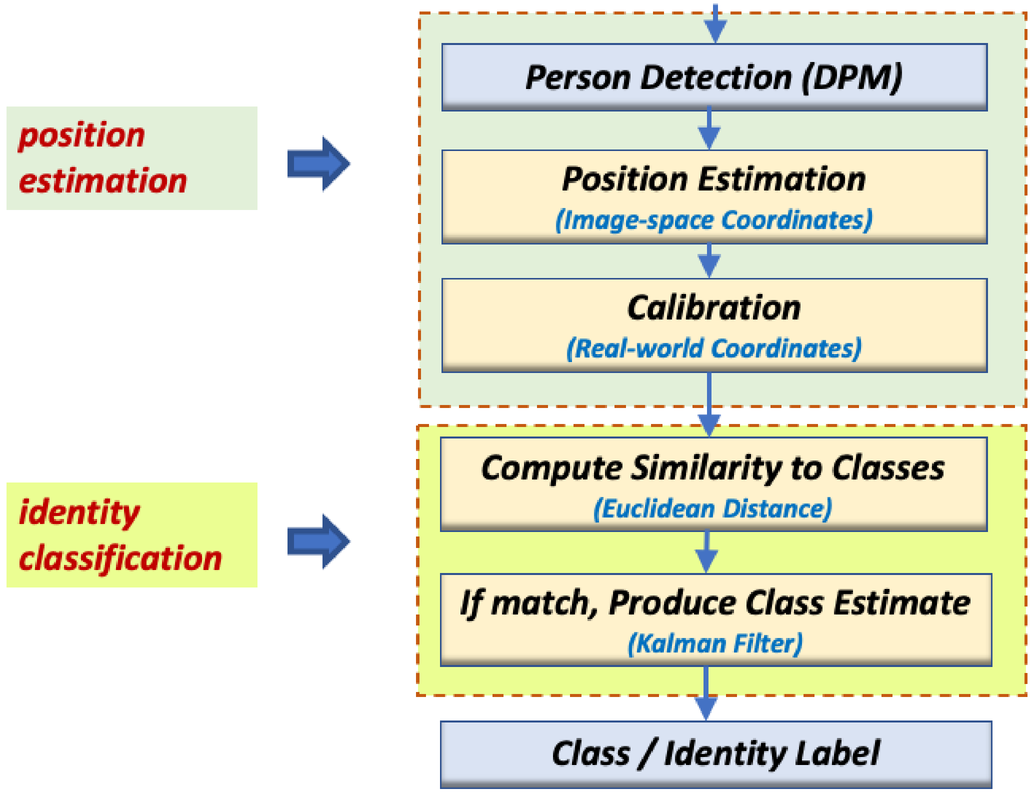
- Each class is represented only as a cluster mean (instead of retaining all the data points).

- Use the first sample's feature vector to initialize a new cluster center ($\mathbf{m}_c$ for class *c*)
- Compare a new probe feature vector X to the cluster mean $\mathbf{m}_c$ for each existing class
- The probe feature vector X is classified into class *c* with the lowest Euclidean distance $||\mathbf{m}_c - X||$
- Update the selected cluster mean using $\mathbf{m}_c = \beta.X + (1 - \beta)\,\mathbf{m}_c$

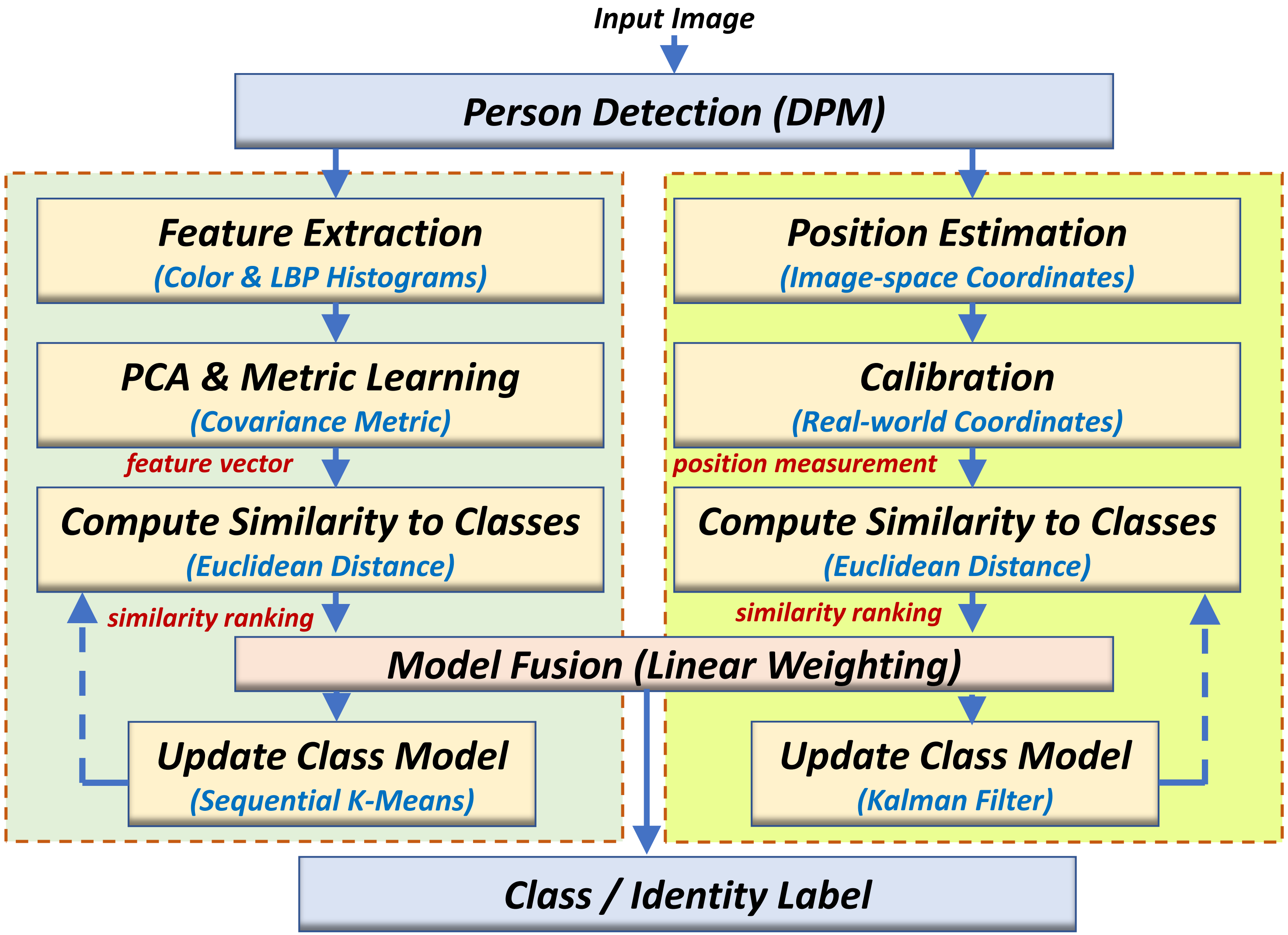
  Proper value for β can be determined through parameter sweeping for the specific data set.

Spatio-Temporal Based Tracking

- Camera calibration matrices are used to convert the image-space pixel coordinate for the person to a real-world coordinate on a map
- Position of each person detected in frame *N* (the current frame) is classified based on their proximity to each of the predictions in frame *N-1* (the previous frame).
- Kalman Filters are used to predict the next position of each track in a way that takes the kinematics of the person into account, with robustness against noise



Fused Appearance and Spatio-Temporal Approach



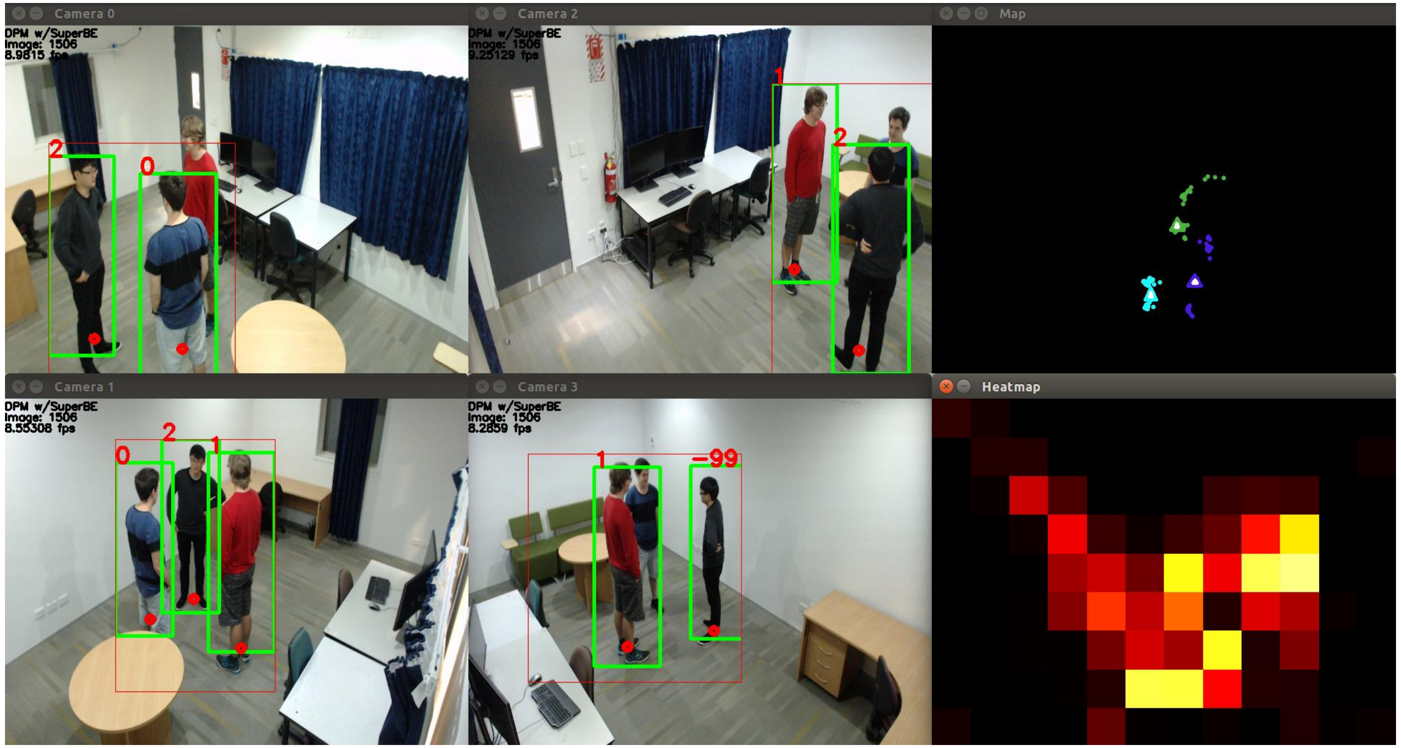
Initial Experimental Results - A Case Study

UoA-Indoor Dataset:

- Three hours of footage from four overlapping cameras (resolution 1920 x 1080 at 15 frames per second)
- 19 different identities annotated across 150000 frames

Experiments were conducted on two cases:

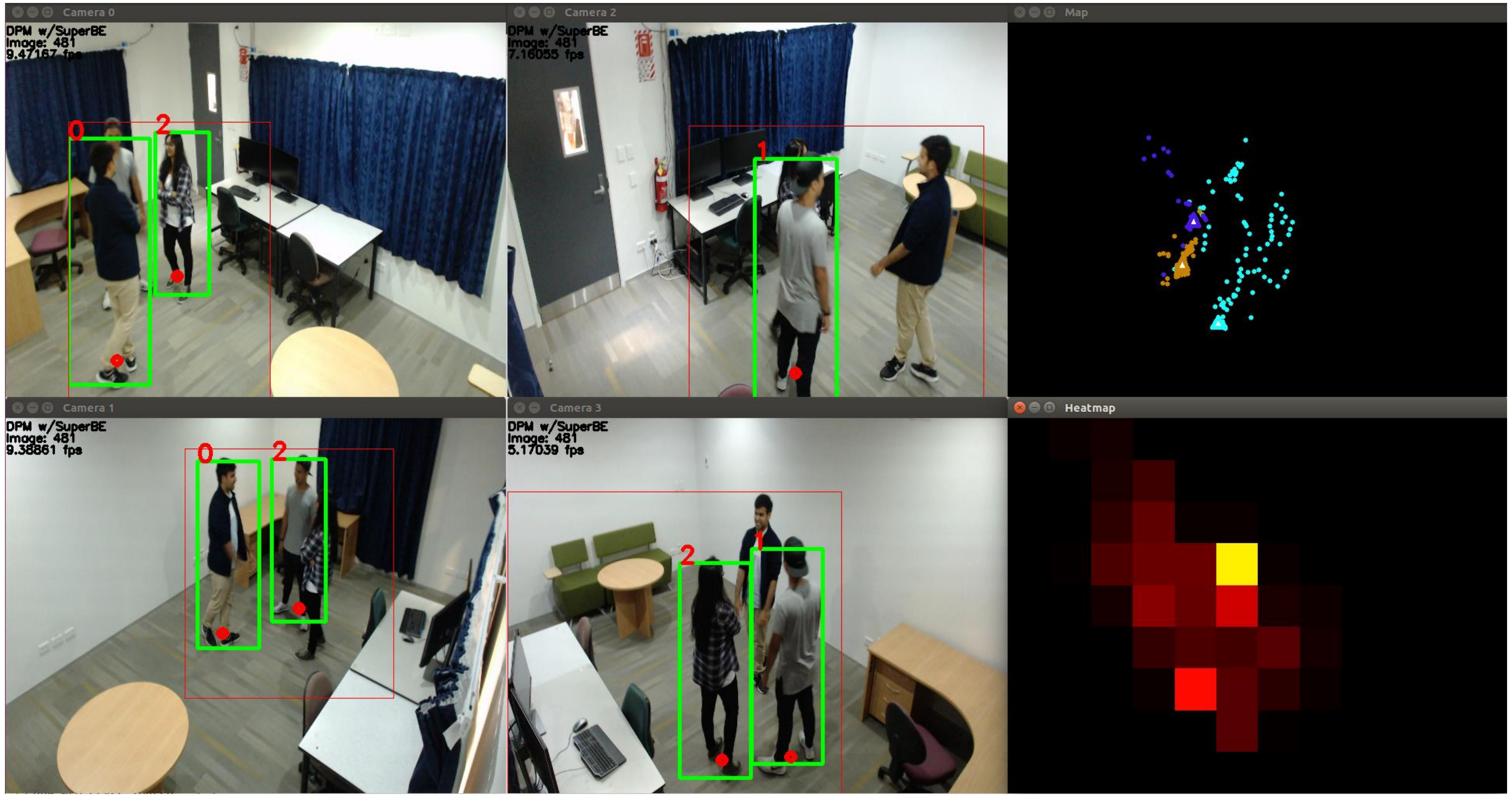
- ***Walk*** (where there is only one person in the room at a time)
- ***Group sequences*** (up to four people in the room at the same time, interacting with each other)



THE UNIVERSITY OF AUCKLAND NEW ZEALAND | ENGINEERING

Initial Experimental Results - A Case Study (Continued)

While people may not be in the room at the same time, the system remembers the identities of people it has seen before.



Initial Experimental Results - A Case Study (Continued)

Comparing DPM vs. ACF:

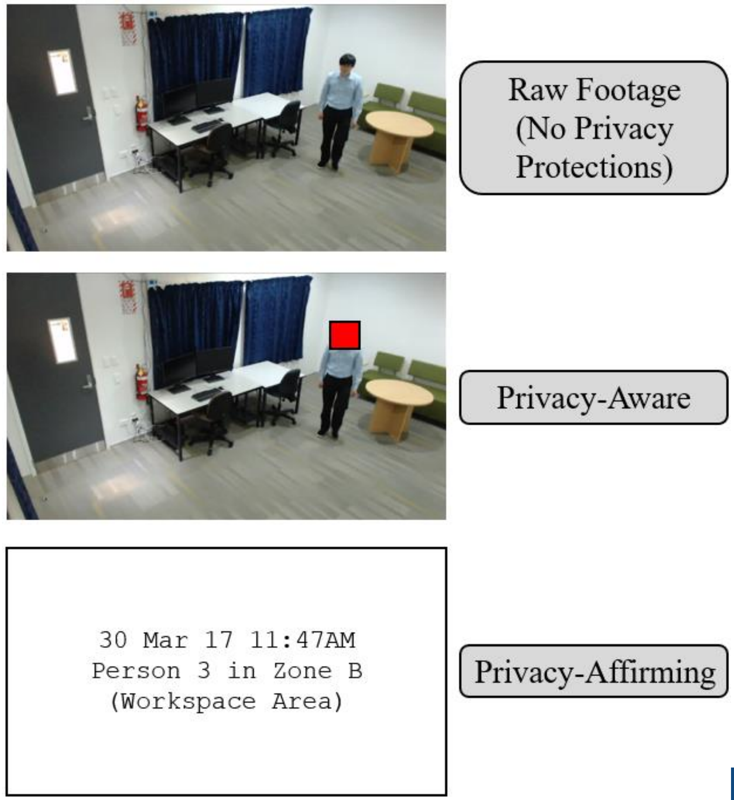
	Classification Model	One-shot learning accuracy %	Unsupervised learning accuracy %	Processing Speed (fps)
DPM	Appearance only	51.9	48.8	
	Spatio-temporal only	33.3	31.7	
	Fused	65.7	**61.6**	9.8
ACF	Appearance only	53.8	47.1	
	Spatio-temporal only	34.3	30.6	
	Fused	**69.4**	56.7	**22.3**

Privacy Issues

Privacy issues may be considered as ***protecting personal information*** and security vulnerabilities that may ***affect the sensitive information***.

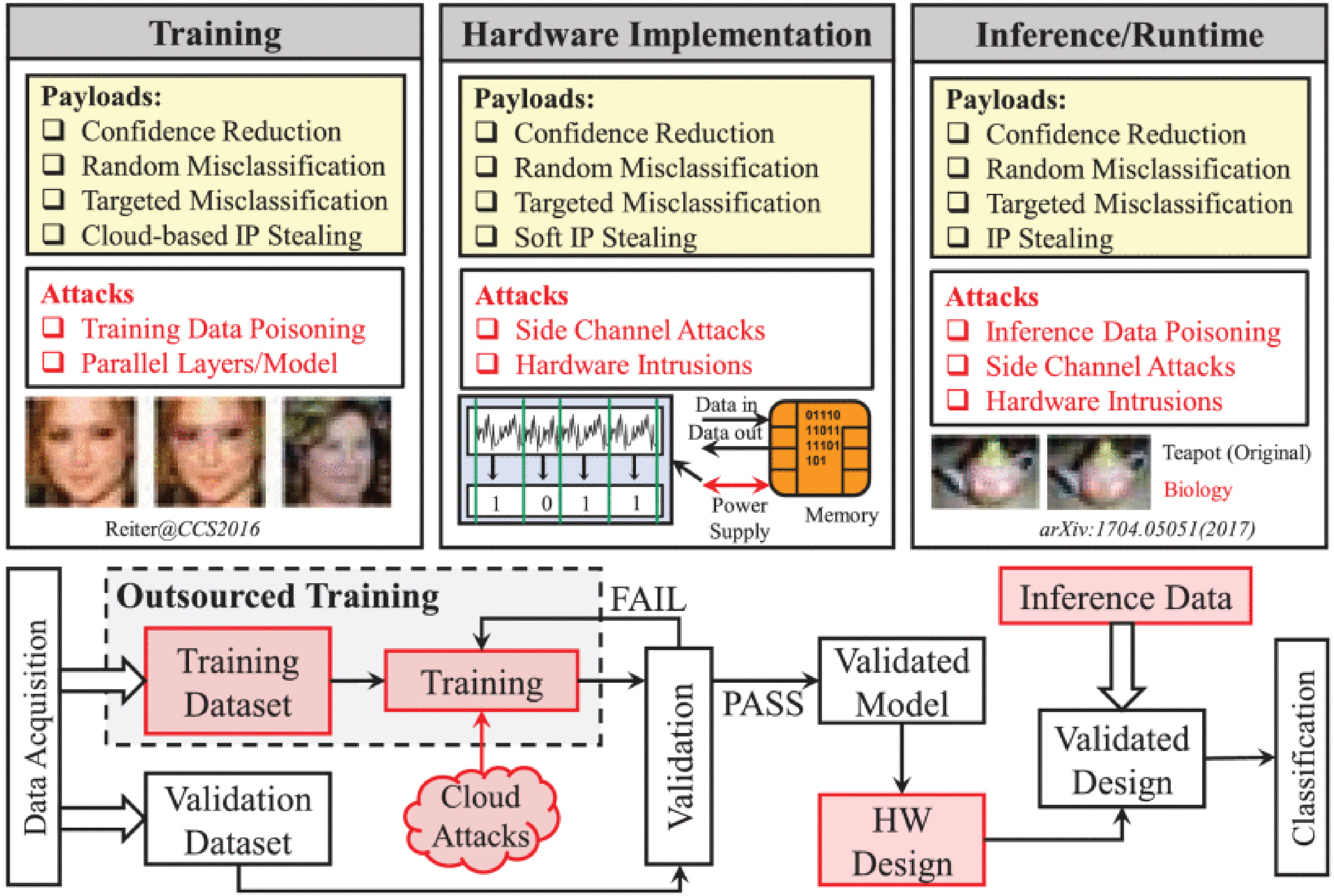
The first issue can be addressed through Privacy-by-Design.

- Privacy-Aware framework: Based on the target application requirements, parts of captured images may be censored to avoid individual identification (where not necessary).
- Privacy-Affirming framework: Only the necessary data is extracted from the input image though computer vision techniques.



Security Vulnerabilities (may affect sensitive information)

Security threats and attacks on machine learning based computer vision systems may significantly compromise the data integrity and robustness of object detection and tracking.



Source: [Hanif et al, 2018]

Summary and Conclusions

- An appearance-based person re-identification and tracking was presented considering some trade-offs between accuracy and computational complexities.
- A spatio-temporal model was discussed to further aid the classification of detected individuals into identity classes, using Kalman Filters to predict the future positions of people.
- A fused appearance-based and spatio-temporal approach was presented to improve the accuracy
- The effectiveness of existing approaches are application dependent.

- **Machine learning based and traditional image processing techniques can be employed for edge device implementations (depending on the required accuracy and application requirements).**
- **Some traditional image processing techniques may be more suitable for implementing at the edge devices (HOG based techniques are more energy efficient than CNN–based approaches).**
- **Privacy, security and data integrity are additional challenges for implementation at the edge.**

Acknowledgements

- Dr. Andrew Tzer-Yeu Chen
- Dr. Kevin I-Kai Wang

Resources: Related Publications

- Chen, A. T., **Biglari-Abhari, M.**, & Wang, K. I. K. **(2020)** *Fusing Appearance and Spatio-Temporal Models for Person Re-Identification and Tracking.* J. Imaging 2020, 6, 27. https://doi.org/10.3390/jimaging6050027
- Chen, A. T., **Biglari-Abhari, M.**, & Wang, K. I. K. **(2019)** *Investigating fast re-identification for multi-camera indoor person tracking*, Elsevier Journal of Computers & Electrical Engineering, Vol. 77, pp. 273 – 288, 2019. https://doi.org/10.1016/j.compeleceng.2019.06.009
- Chen, A. T-Y., **Biglari-Abhari, M.,** Wang, K. **(2018)** *SuperBE: Computationally-Light Background Estimation with Superpixels*, Journal of Real-time Image Processing, January 2018
- Chen, A. T., Gupta, R., Borzenko, A., Wang, K. I. K & **Biglari-Abhari, M. (2018).** *Accelerating SuperBE with Hardware/Software Co-Design*, in Journal of Imaging, 2018, 4(10), 122; doi: 10.3390/jimaging410012
- Chen, A. T., **Biglari-Abhari, M.**, & Wang, K. **(2018).** *Fast One-Shot Learning for Identity Classification in Person Re-identification and Tracking*, in Proceedings of the 15th IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV-2018), Singapore, 18-21 Nov. 2018
- Chen, A. T., **Biglari-Abhari, M.**, & Wang, K. **(2018).** *Context is King: Privacy Perceptions of Camera-based Surveillance*, in Proceedings of the 15th IEEE International Conference on Advanced Video and Signal-based Surveillance, Auckland - New Zealand, 27-30 November 2018
- Chen, A. T., **Biglari-Abhari, M.**, Wang, K. I. K., Bouzerdoum, A., & Tivive, F. H. -C. **(2018).** *Convolutional Neural Network Acceleration with Hardware/Software Co-Design*. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies. 48 (5), 1288-1301, doi:10.1007/s10489-017-1007-z
- Chen, A. T-Y., **Biglari-Abhari,** M., Wang, K. I-K., **(2017)** *Trusting the Computer in Computer Vision: A Privacy-Affirming Framework*, Proceedings of The First International Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (CV-COPS 2017), Honolulu, Hawaii — July 21, 2017
- Chen, A. T-Y., Fan, J., **Biglari-Abhari,** M., Wang, K. I-K., **(2017)** *A Computationally Efficient Pipeline for Camera-based Indoor Person Tracking*, Proceedings of Image and Vision Computing New Zealand (IVCNZ 2017), Christchurch, New Zealand — 4 – 6 Dec. 2017

Other Related Works (Our Research Team)

Other Related Embedded Computer Vision Systems publications:

- Hemmati, M., **Biglari-Abhari, M.**, & Niar, S. **(2019)** *Adaptive Vehicle Detection for Real-time Autonomous Driving System*, in Proceedings of the 2019 IEEE Conference on Design, Automation & Test in Europe (DATE), Florence, Italy, 25-28 March 2019, pp. 1034-1039, doi:10.23919/DATE.2019.8714818
- Porter, R., Morgan, S., **Biglari-Abhari, M. (2019)** *Extending a Soft-Core RISC-V Processor to Accelerate CNN Inference*, to appear in Proceedings of the Sixth Annual Conference on Computational Science & Computational Intelligence, Las Vegas, Nevada, 5-7 December 2019
- Hemmati, M., **Biglari-Abhari, M.,** Niar, S., Berber, S., **(2017)** *Real-Time Multi-Scale Pedestrian Detection for Driver Assistance Systems*, ACM/IEEE Proceedings of the 54th Design Automation Conference (DAC), Austin, TX, 18-22 June 2017
- Hemmati, M., **Biglari-Abhari, M.**, Berber, S., & Niar, S. **(2014)** *HOG Feature Extractor Hardware Accelerator for Real-time Pedestrian Detection*. Proceedings of 17th Euromicro Conference on Digital System Design (DSD), Verona, ITALY: 27 August - 29 August 2014. (543-550)

References

[Benfold & Reid, 2011] B. Benfold and I. Reid, “Stable multi-target tracking in real-time surveillance video,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3457–3464.

[Cho & Rybski, 2012] H. Cho, P. E. Rybski, A. Bar-Hillel, and W. Zhang, “Real-time pedestrian detection with deformable part models,” in Intelligent Vehicles Symposium (IVS), 2012, pp. 1035–1042.

[Dalal & Triggs, 2005] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 886–893.

[DeSmedt & Goedeme] F. DeSmedt and T. Goedemé,“Open framework for combined pedestrian detection,” in International Conference on Computer Vision Theory and Applications (VISIGRAPP), 2015, pp. 551–558.

[Dollar et al, 2014] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1532–1545, 2014.

[Everingham & Zisserman, 2006] M. Everingham and A. Zisserman, “Automated person identification in video,” in International Conference on Image and Video Retrieval (CIVR), 2006, pp. 289–298.

[Girshick, 2015] R. Girshick, “Fast R-CNN,” in International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.

[Gou et al, 2017] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke, “DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset,” in Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 10–19.

[Hanif et al, 2018] M. A. Hanif, F. Khalid, R. V. W. Putra, S. Rehman and M. Shafique, "Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks," 2018 IEEE 24th International Symposium on On-Line Testing And Robust System Design (IOLTS), pp. 257-260

[Krumm et al, 2000] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, “Multi-camera multi-person tracking for EasyLiving,” in International Workshop on Visual Surveillance, 2000, pp. 3–10.

[Ren et al, 2017] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137– 1149, 2017.

[Yan et al, 2014] J. Yan, Z. Lei, L. Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2497–2504.