UNIVERSITÀ
POLITECNICA
DELLE MARCHE

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

VRAI

People counting on low cost embedded hardware during the SARS-CoV-2 pandemic

Giulia Pazzaglia, Marco Mameli, Luca Rossi, Marina Paolanti,
Adriano Mancini, Primo Zingaretti, Emanuele Frontoni

Dipartimento di Ingegneria dell'Informazione (DII)
Università Politecnica delle Marche, Ancona, Italy
https://vrai.dii.univpm.it/

Introduction

- Detecting and tracking people is a challenging task in a persistent crowded environment as retail, airport or station, for human behavior analysis of security purposes.

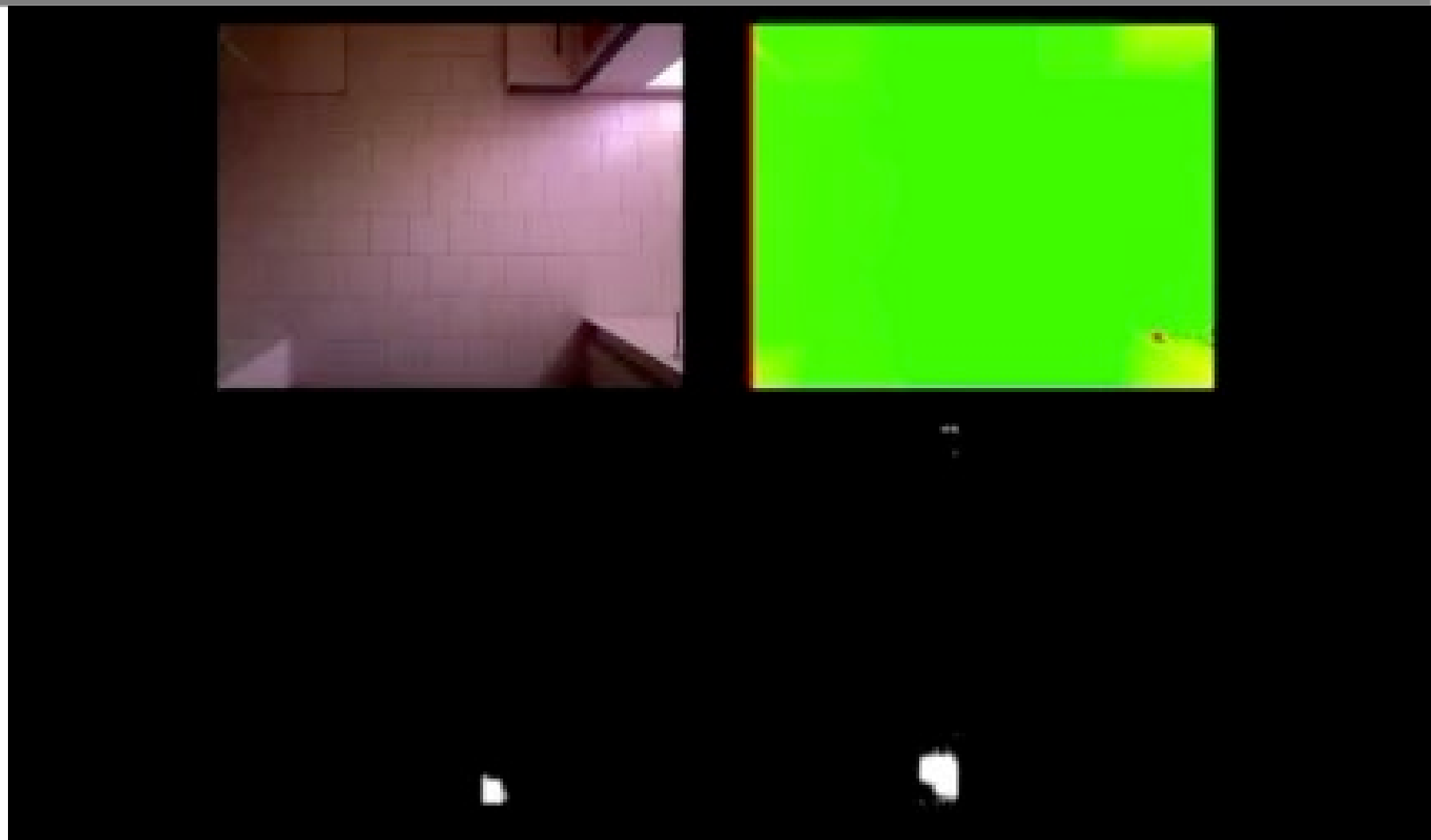
- During the global spread of SARS-CoV-2 virus that has become part of everyday life in every country people counting is a mandatory access to regulate the access to buildings / shops

- In this context it is therefore useful to create systems for the control of gates and queues, which help to automatically manage the density of people within an environment, by acting on traffic lights, automatic ticket machines, turnstiles and automatic doors to control the flow and prevent overcrowding.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

Introduction – Previous Works



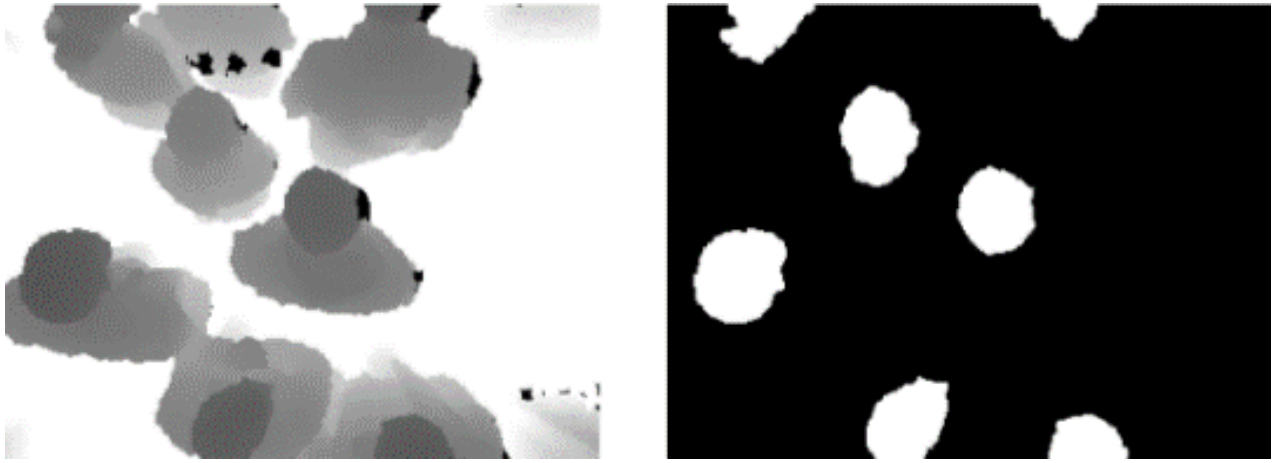
Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., & Zingaretti, P. (2018, August). Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment. In 2018 24th international conference on pattern recognition (ICPR) (pp. 1384-1389). IEEE.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

TVHEADS - TOP-VIEW HEADS DATASET

- 1815 depth images (16 bit) with dimensions of 320 × 240 pixels captured from an RGB-D camera in a top-view configuration.
- The images collected in this dataset represented a crowded retail environment with at least three people per square metre and physical contact between them.

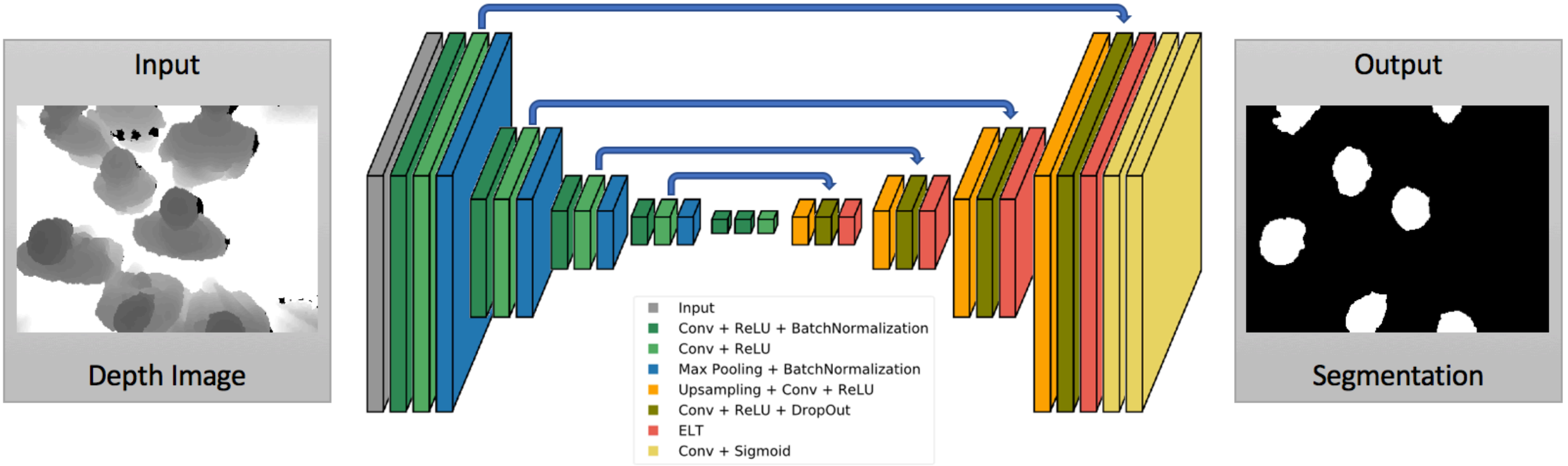


Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., & Zingaretti, P. (2018, August). Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment. In 2018 24th international conference on pattern recognition (ICPR) (pp. 1384-1389). IEEE.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

VRAI-NET 1 FOR PEOPLE COUNTING IN CROWDED ENVIRONMENT



Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., & Zingaretti, P. (2020). Deep understanding of shopper behaviours and interactions using RGB-D vision. Machine Vision and Applications, 31(7), 1-21.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

VRAI-NET 1 FOR PEOPLE COUNTING IN CROWDED ENVIRONMENT: SETTINGS

- **Optimizer:** Adam (lr =0.01)
- **Loss:** Dice Loss
- **Regularisation:** Dropout e Batch normalization
- **Metric for weights saving:** Dice Loss

$$\text{Dice Loss} = 2|X \cap Y|/(|X| + |Y|)$$

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., & Zingaretti, P. (2020). Deep understanding of shopper behaviours and interactions using RGB-D vision. Machine Vision and Applications, 31(7), 1-21.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

VRAI-NET1 FOR PEOPLE COUNTING IN CROWDED ENVIRONMENT: RESULTS

Approach	Jaccard	Dice	Accuracy	Precision	Recall	F1-Score
Fractal	0.9477	0.9732	0.9444	0.9927	0.9933	0.9930
SegNet	0.8277	0.9058	0.9927	0.9462	0.9533	0.9497
ResNet	0.8482	0.9179	0.9938	0.9688	0.9693	0.9690
U-Net	0.8695	0.9302	0.9926	0.9450	0.9490	0.9469
U-Net 2	0.9382	0.9681	0.9932	0.9679	0.9706	0.9691
U-Net 3	0.9299	0.9637	0.9946	0.9894	0.9894	0.9894
VRAI-Net 1	**0.9290**	**0.9642**	**0.9946**	**0.9893**	**0.9895**	**0.9894**

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., & Zingaretti, P. (2020). Deep understanding of shopper behaviours and interactions using RGB-D vision. Machine Vision and Applications, 31(7), 1-21.

Challenge

- Main challenges
  - This work focuses on the field of People Counting and in particular on the Line of Interest (LOI) approach as the RGB video streams are captured from a camera with top-view perspective
  - The developed solution is designed for real retail environments with great variability of acquired data derived from a large experience minizing as much as possible the overall cost;
  - People counting has to work with contemporary shoppers;
  - Dataset of shoppers acquired during the pandemic situation in Marche Region, in the center of Italy.

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

Materials and Methods

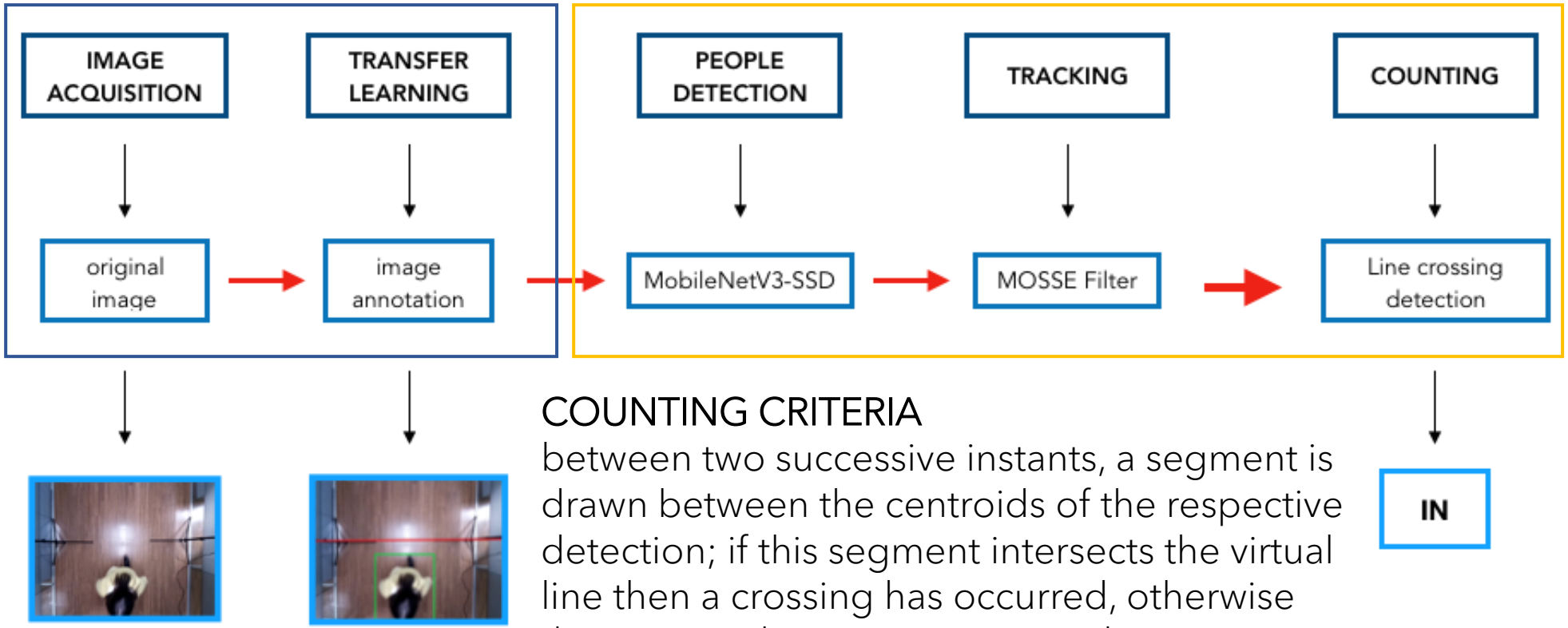
UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

Framework

The proposed framework is organized into five main stages:

- Image Acquisition
- Transfer Learning
- People Detection
- Tracking
- Counting



COUNTING CRITERIA

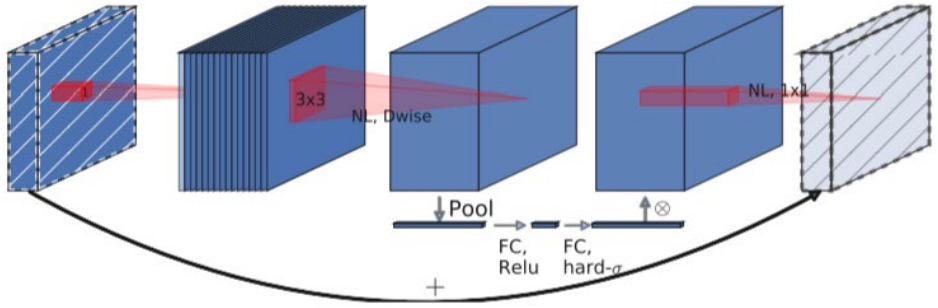
between two successive instants, a segment is drawn between the centroids of the respective detection; if this segment intersects the virtual line then a crossing has occurred, otherwise the crossing has not yet occurred.

UNIVERSITÀ POLITECNICA DELLE MARCHE

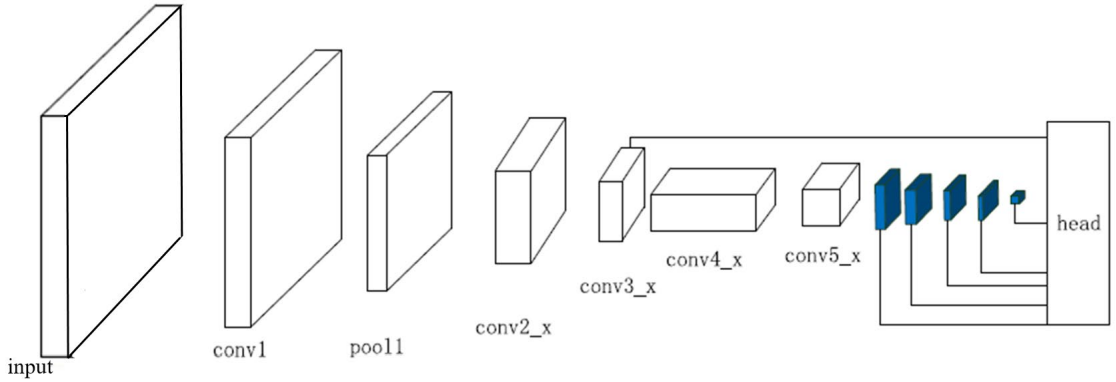
VRAI

People Detection

For the People Detection task, we used MobileNetV3[1] as base architecture for SSD[2] Network.



MobileNetV3 Network



SSD Network

[1] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu,Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1314–1324 (2019)
[2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision.pp. 21–37. Springer (2016)

Tracking

For the Tracking phase, **Minimum Output Sum of Squared Error** (MOSSE)[3] filter was adopted, that is a stable correlation filter which can be initialized on a single frame of a video. It is described by the following formula:

$$\min_{H^*} \sum_i |F_i \odot H^* - G_i|^2$$

- $F_i \odot H^*$ is the filtered training input
- G_i is the training output

[3] Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 2544–2550. IEEE (2010)

Performance Evaluation Metrics

Taken a range from 1 to n in the ranking:

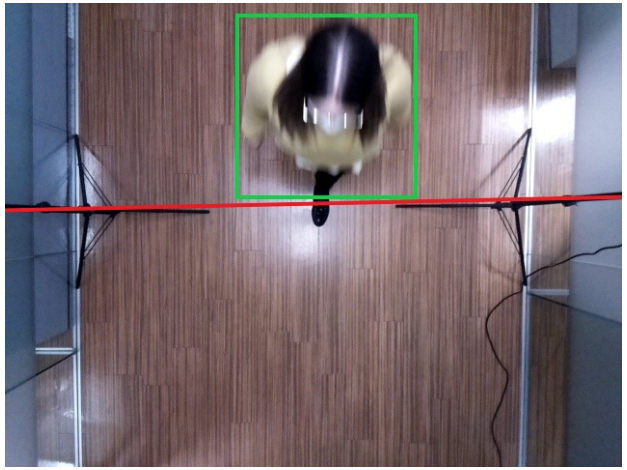
- Average Precision :$AP = \sum_{i=1}^{n} \frac{Precision_i}{n}$
- Average Recall : $AR = \sum_{i=1}^{n} \frac{Recall_i}{n}$
- Mean Average Precision : $mAP = \frac{1}{N} \sum_{i=1}^{n} AP_i$

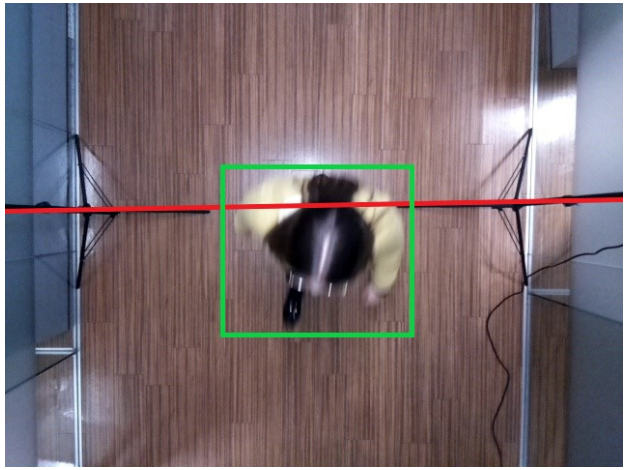
UNIVERSITÀ POLITECNICA DELLE MARCHE

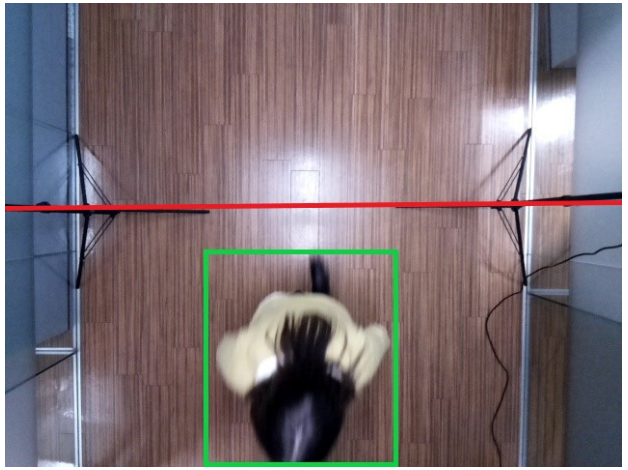
VRAI

Dataset

- The dataset specifically collected for this work, was acquired in real retail installation for 1 class “top-view person”.
- Using an RGB camera in top-view perspective, about 8000 image acquisitions were performed. The images have been subsequently manually annotated.
- The annotation of the images was carried out in three different stage:
  - before the virtual red line;
  - below the virtual red line;
  - after the red line.





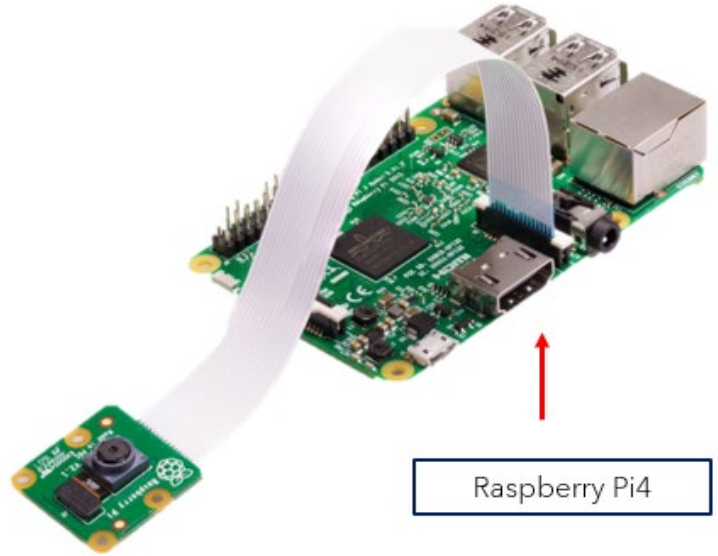


UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

Hardware Setup

Hardware setup



The image processing takes place directly on the Raspberry with **inference times** equal to **0.11** seconds for Raspberry, using the **MOSSE Tracking**, and **taking** a **frame every 3** in order to **process 20 fps**.

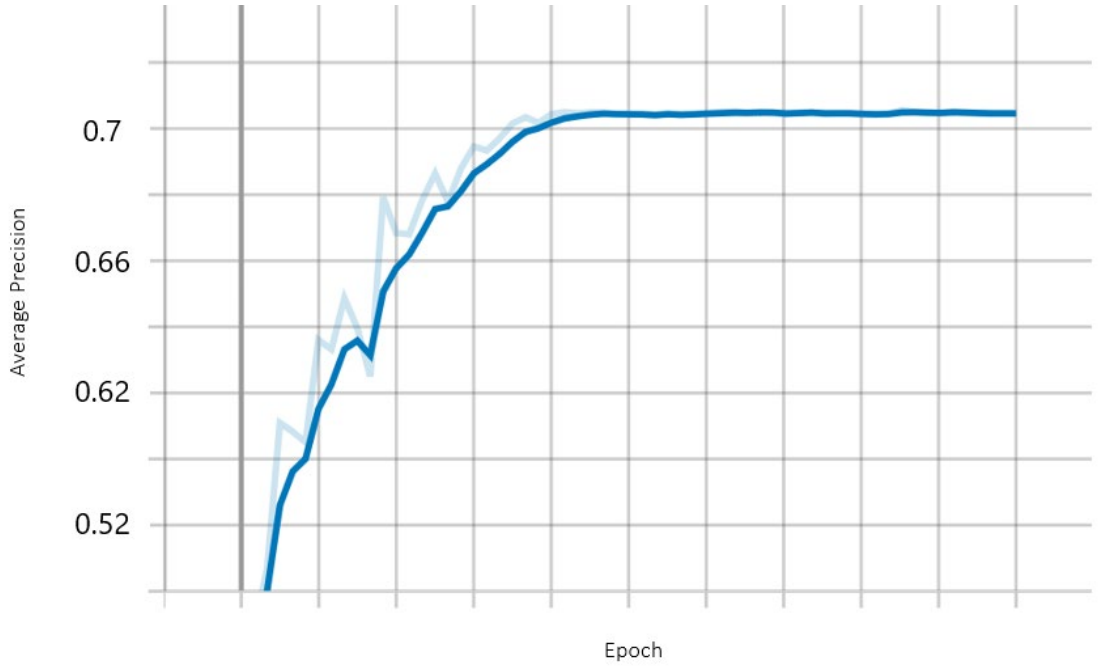
Results and Discussion

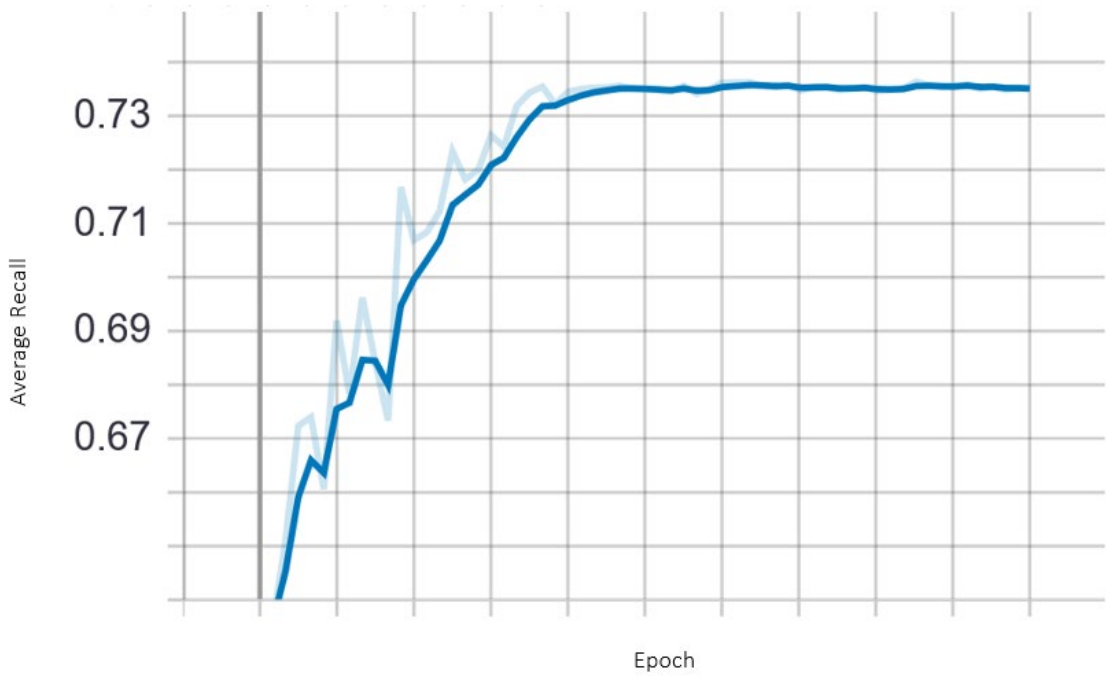
Results

For the training of MobileNetV3 we used:

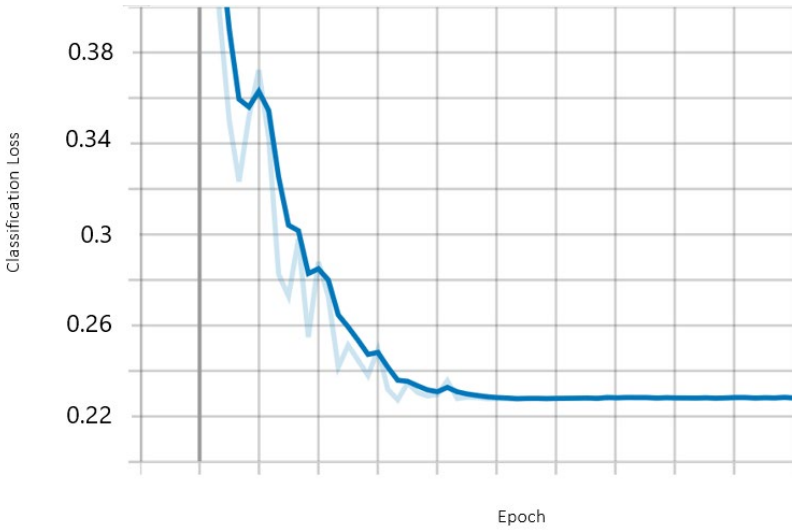
- **ReLU6** as an activation function;
- **l2-regularized** as a regularizer;
- **weighted sigmoid focal** as classification loss with parameters $\alpha = 0.75$ and $\gamma = 0.2$;
- **weighted smooth l1** as localization loss with $\delta = 1.0$;
- **sigmoid function** as final loss of the model;
- batch size = 128
- **momentum** with parameter $\mu = 0.9$ and adaptive learning rate was used as an optimizer starting from a base of 0.05 and applying the cosine decay learning rate.

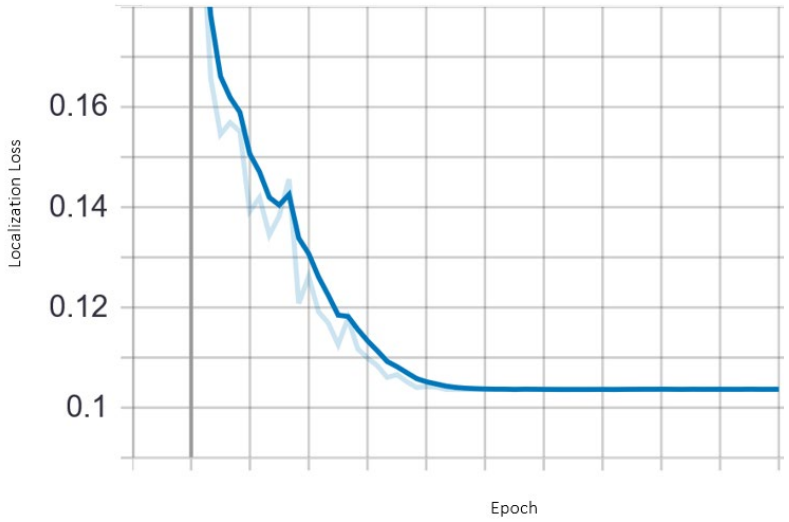
Results – AP & AR

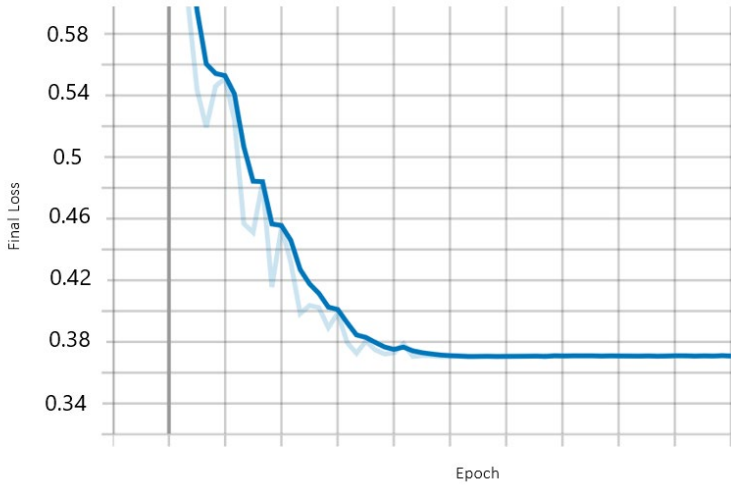




Results – Loss







Conclusions

- Deep Learning approaches, and in particular CNNs, have been used to overcome the problem of counting people that cross a virtual line and the problem of counting people who are within a certain area of interest.
- MobileNetV3 has been chosen for the task of People Detection and a MOSSE filter has been adopted for the tracking phase.
- The results of MobileNetV3 for the detection are promising; the AP was increased from 67% to 70%.
- Top-view configuration preserves the privacy of users.
- The tool developed for the acquisition of RGB images and for people counting is inexpensive, as it is formed by Raspberry Pi4 and a Picamera module v2.
- These devices are already installed in real environments, such as some museums and supermarkets in the Marche region (Italy).

Future Works

- Improve the performance of the network, by increasing the dataset in terms of locations, environmental conditions (e.g. light)
- Testing other networks for the detection task;
- Evaluate the system in other real environments.

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Thanks for your attention !!!

https://vrai.dii.univpm.it/

Contact: a.mancini@univpm.it

UNIVERSITÀ
POLITECNICA
DELLE MARCHE