Person Detection from a Top-View Perspective

Mameli Marco
Paolanti Marina
Pazzaglia Giulia
Frontoni Emanuele

Baldascino Giovanni - 1097405
Squarcella Loisi - 1096539

UNIVERSITÀ POLITECNICA DELLE MARCHE

INTRODUCTION

Person Detection is performed from a Top-View Perspective by using most recent object detection frameworks.

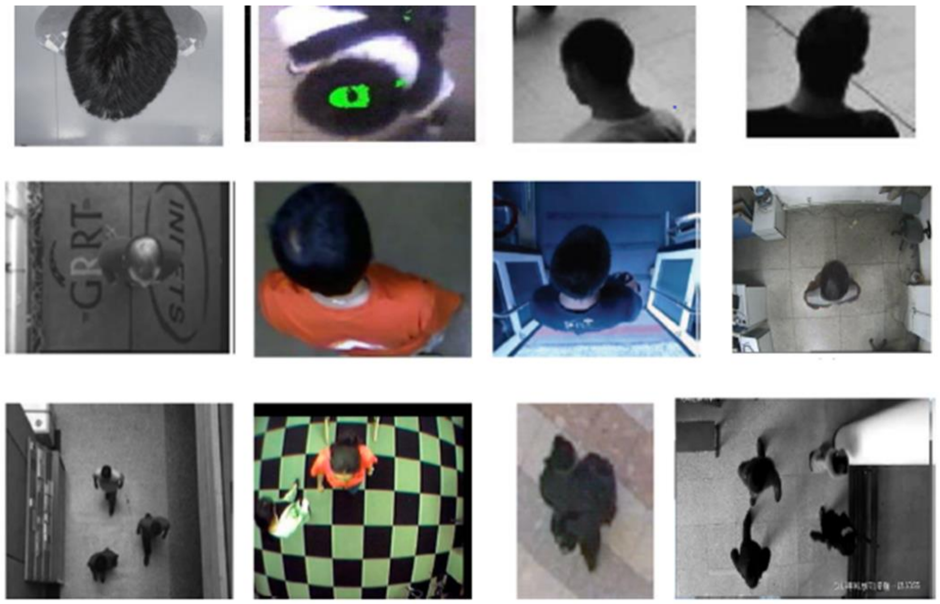
The aim is to build an efficient model to detect the people in the scene.

This framework is evaluated on a real-world scenario which is a retail environment.

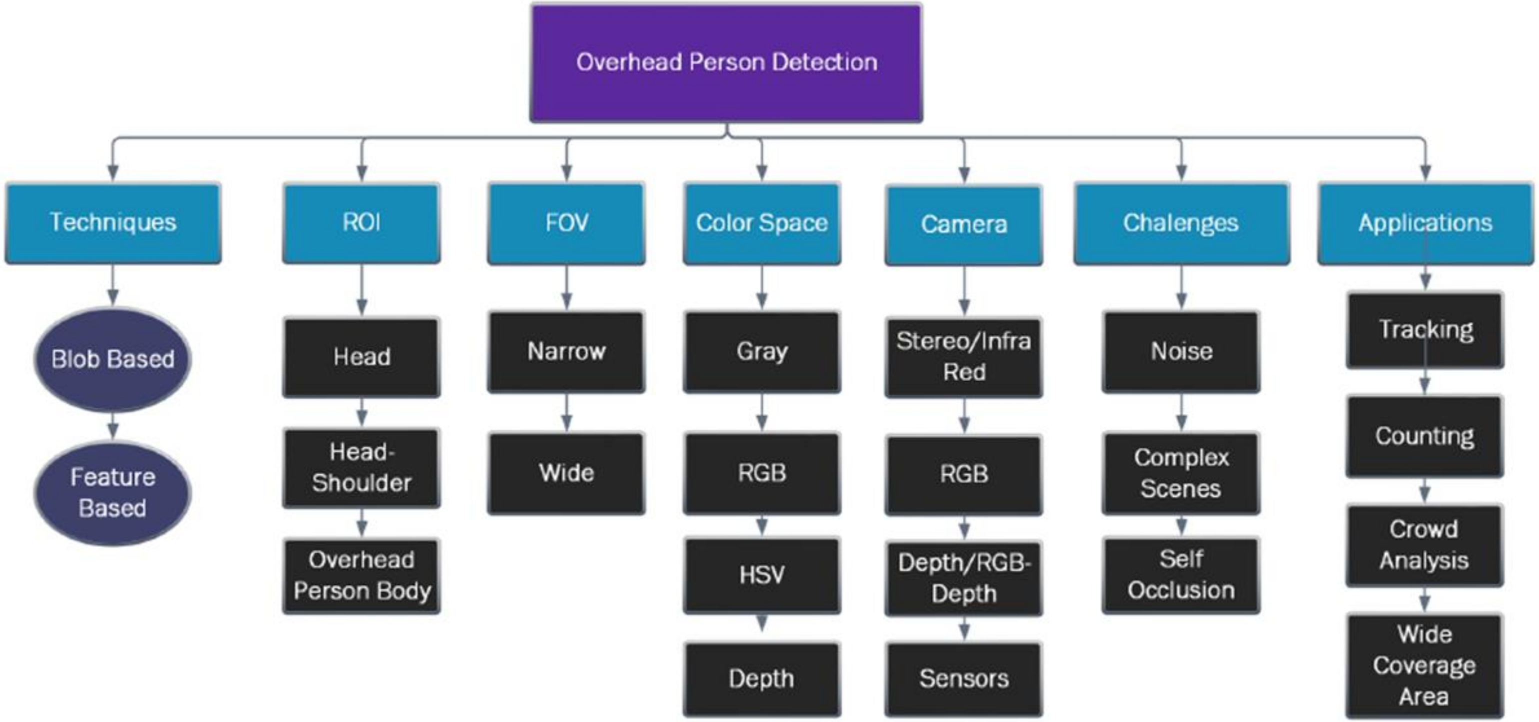
INTRODUCTION

MAIN STEPS:

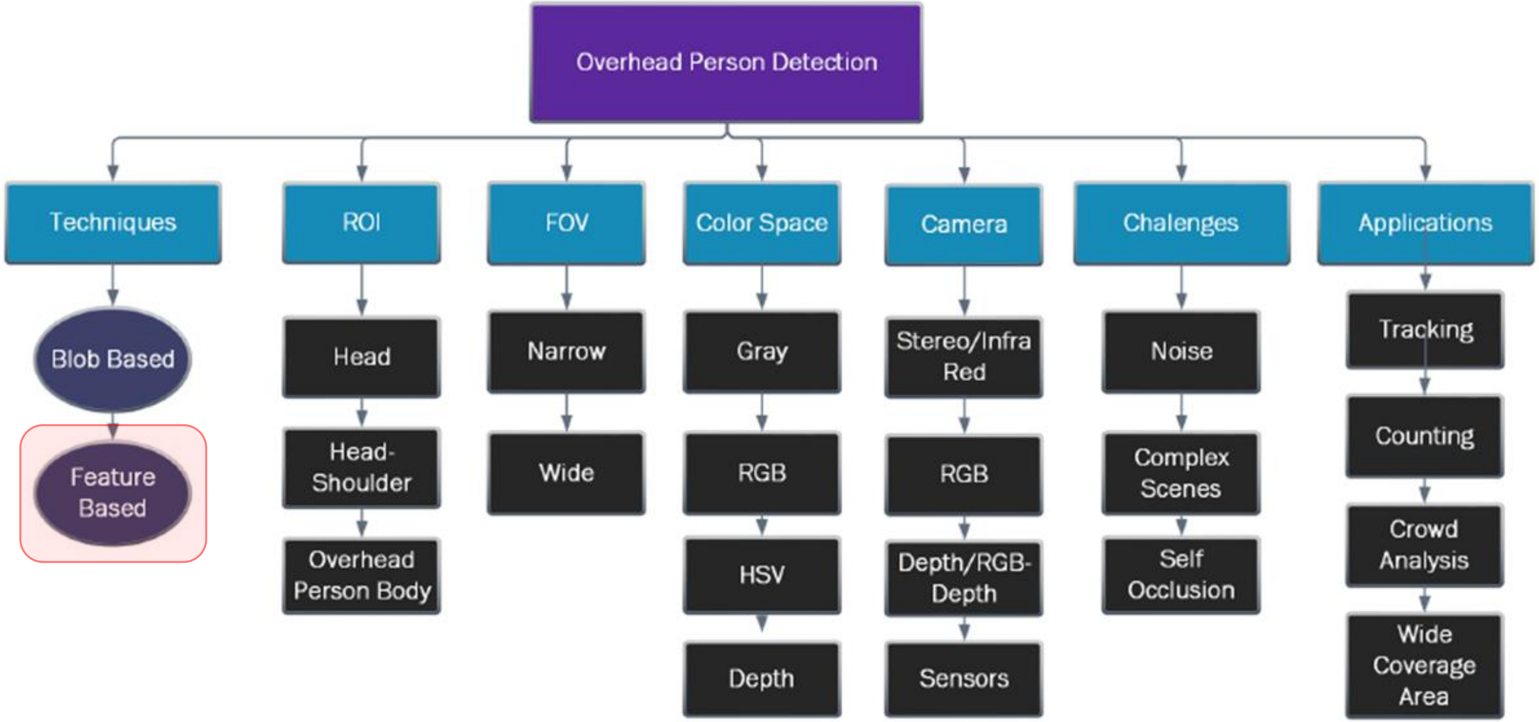
- Definition of the **Region of Interest** (ROI);
- People localization.



STATE OF ART – Person Detection



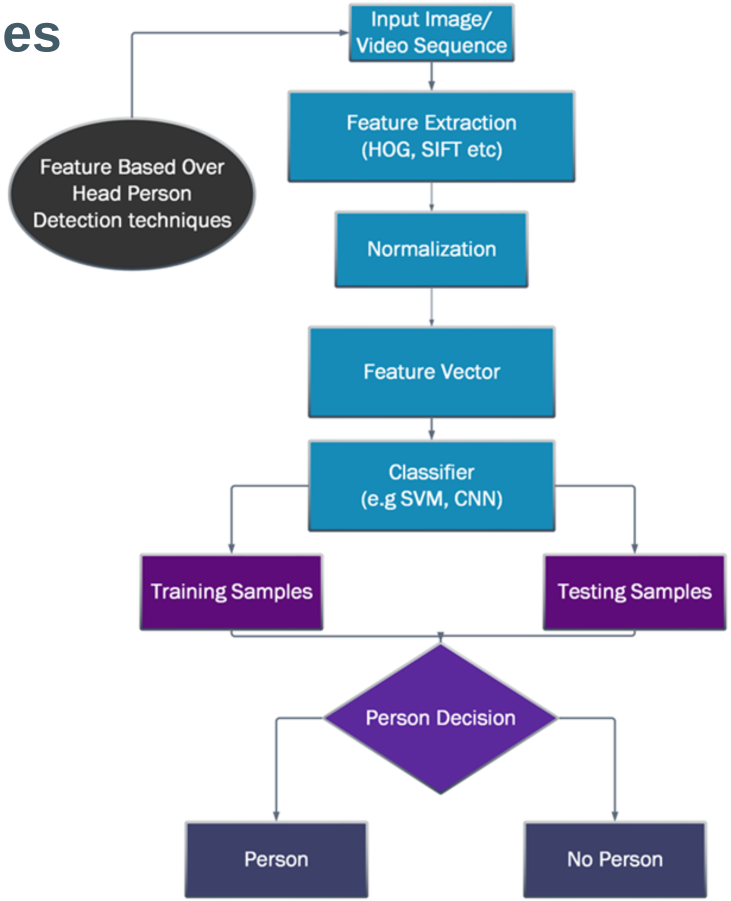
STATE OF ART – Person Detection



STATE OF ART – Feature based Techniques

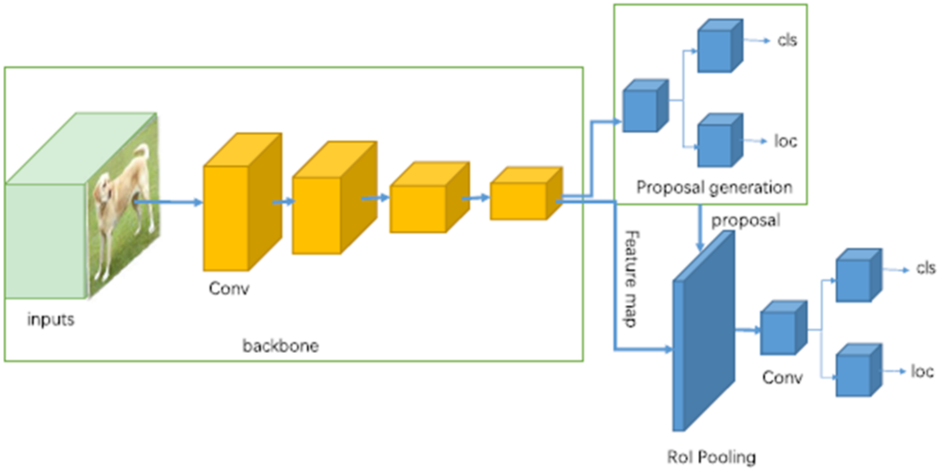
These **Feature based Techniques** operate on *features extracted* from overhead view videos and images.

The extracted features contain shape, color, texture, etc.… The images are often divided into samples for training and testing.



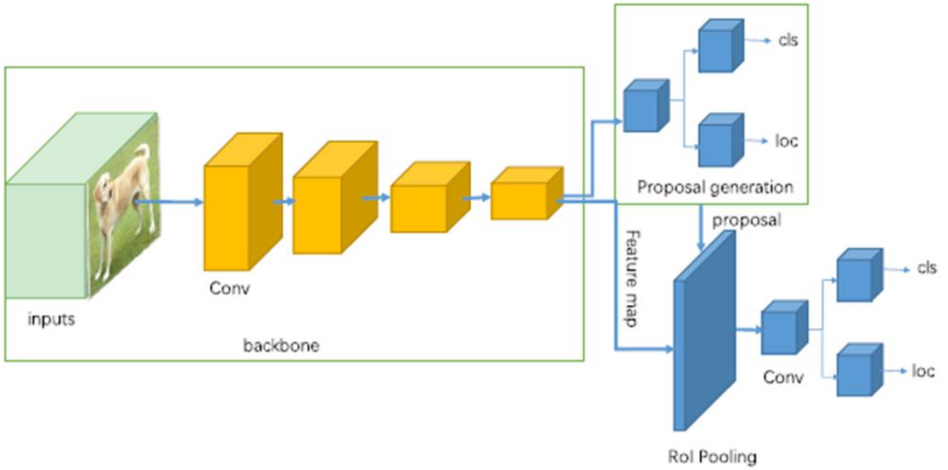
STATE OF ART – Feature based Techniques

Two-stage Detectors (R-CNN, Faster R-CNN, etc..) use a *Region Proposal Network* to generate regions of interests.

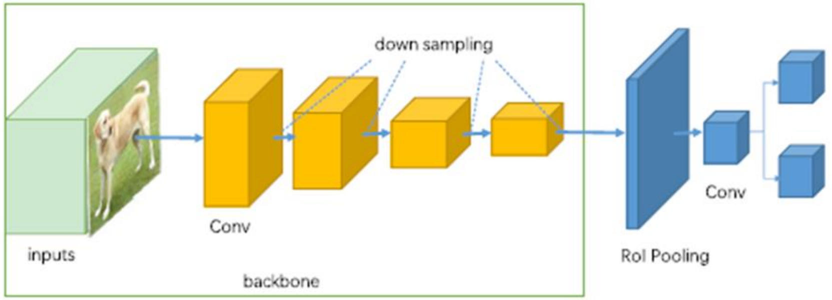


STATE OF ART – Feature based Techniques

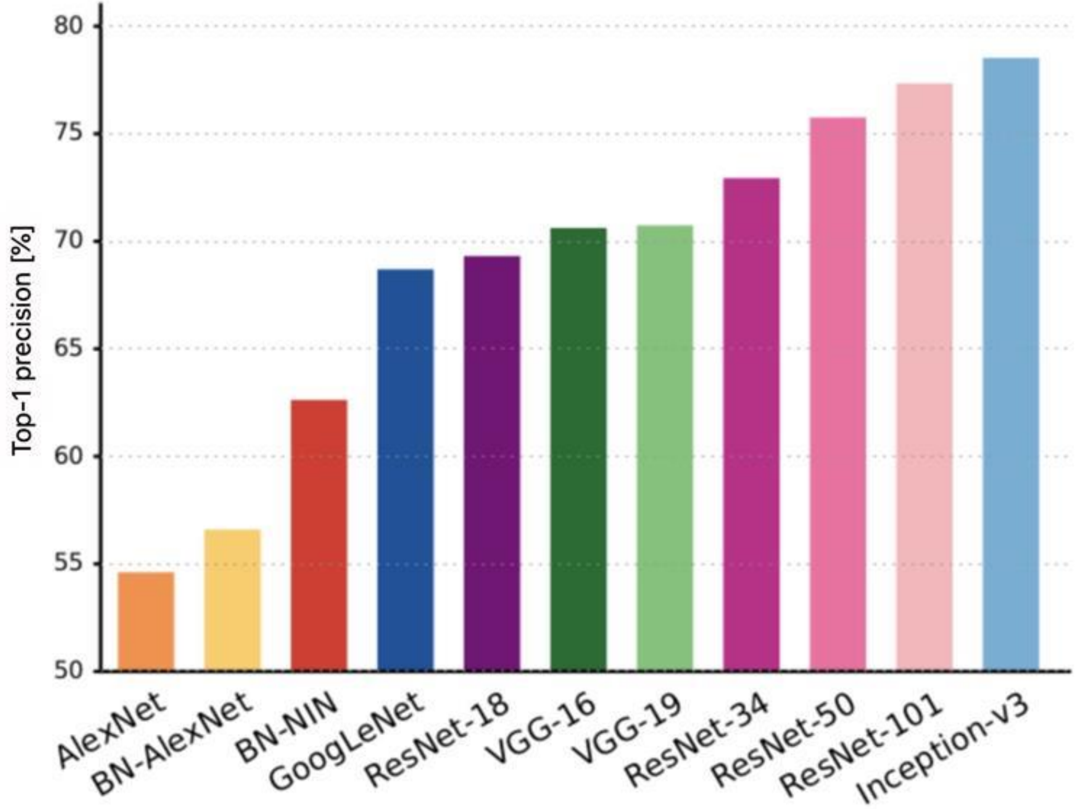
Two-stage Detectors (R-CNN, Faster R-CNN, etc..) use a *Region Proposal Network* to generate regions of interests.



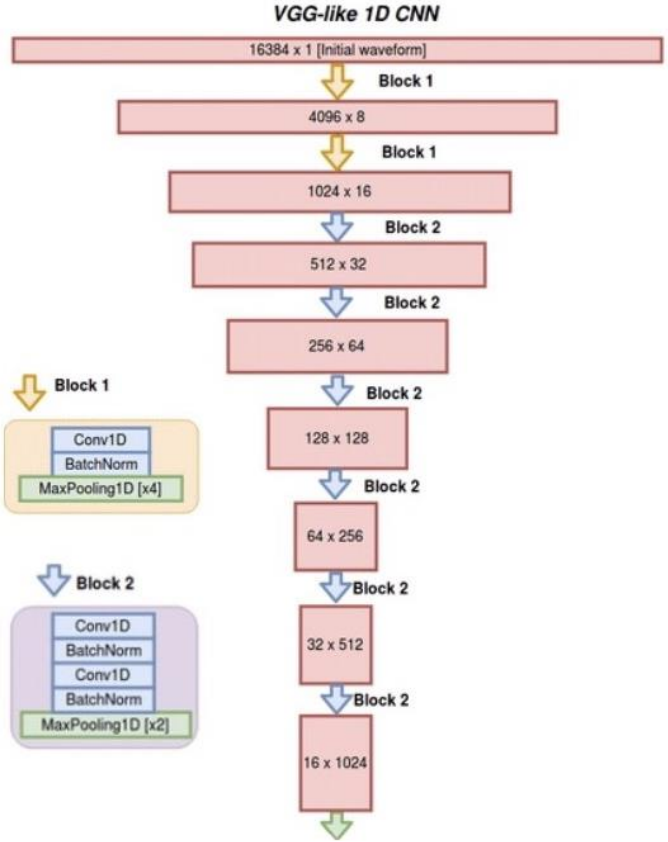
One-stage Detectors (YOLO, SSD, etc..) treat object detection as a *simple regression problem*.

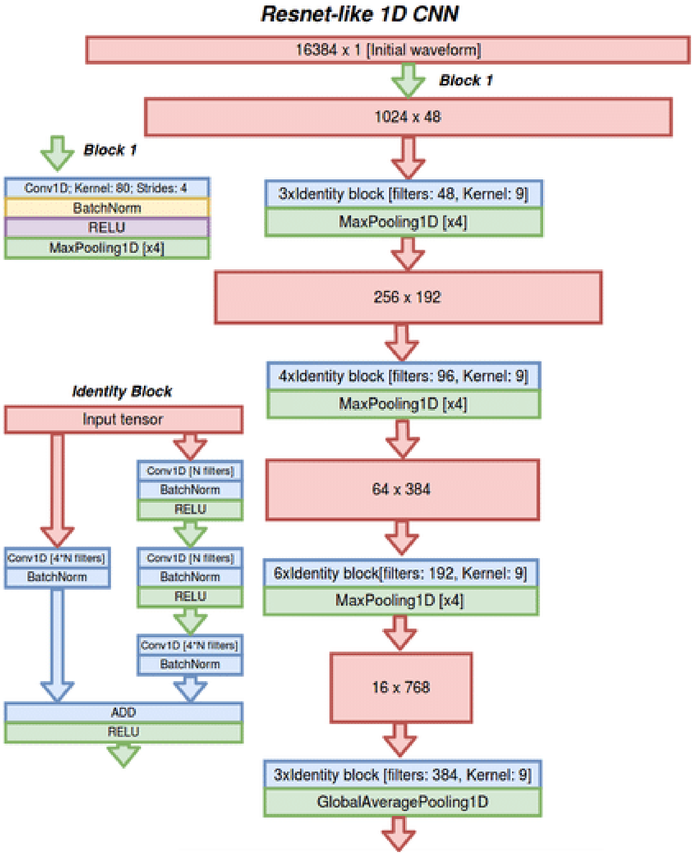


STATE OF ART – Feature extraction architectures



STATE OF ART – Feature extraction architectures





MATERIALS AND METHODS – Data Collection

The provided *Dataset* comes from video frames recorded by different cameras from a top-view, within a real environment: a supermarket.

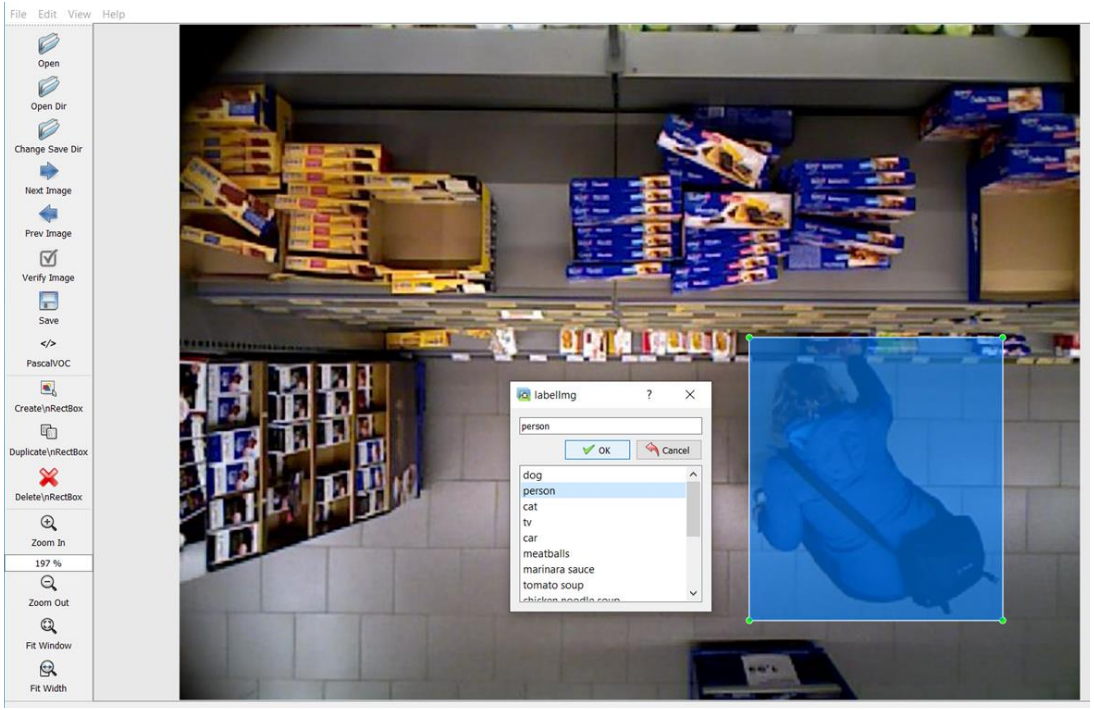
These 30.000 images needed a skim because some did not contain people.

MATERIALS AND METHODS – Labelling

Dataset annotation using **LabelImg**. LabelImg is a *graphical image annotation tool*.

About 17.000 images divided into:

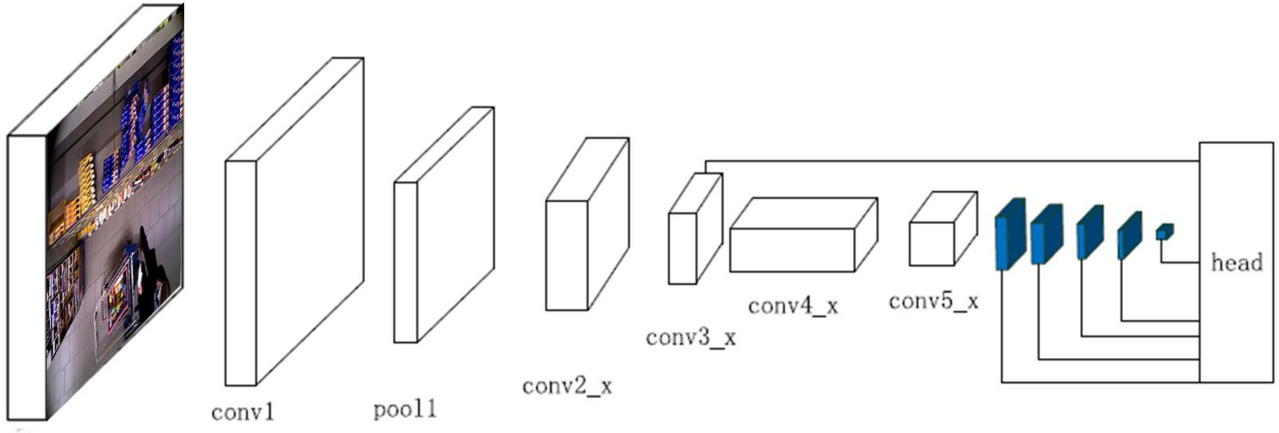
- **Testing set** → 30%
- **Training set** → 70%
- **Validation set** → 30% of training set



MATERIALS AND METHODS – Single-shot Detector

SSD (Single-shot Detector) discretizes the output space of *bounding boxes* into a set of default boxes over different aspect ratios and scales per feature map location.

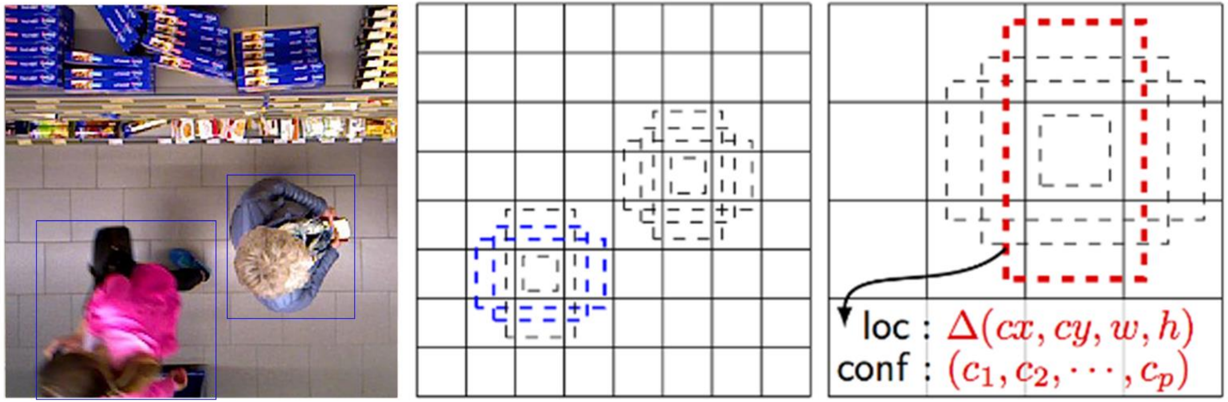
In SSD the *prediction layer* is acting on fused features of different levels. Head module consists of a series of *convolutional layers* followed by several classification layers and localization layers.



[Wei Liu et al, 2016] Present a method for detecting objects in images using a single deep neural network. The approach, named SSD, discretizes the output space of bounding boxes into a set of default boxes. over different aspect ratios and scales feature map location.

UNIVERSITÀ POLITECNICA DELLE MARCHE

MATERIALS AND METHODS – Single-shot Detector



Each prediction is composed of:

- Bounding box with shape offset (Δcx, Δcy, w and h);
- *Confidences* for all object categories or all the classes.

MATERIALS AND METHODS – Settings

Three types of Single-shot Detectors were used: **SSD300**, **SSD512** and **SSD7**. The architecture of the first two is almost the same (9 and 10 layers respectively) while the SSD7 provides a simplified approach (7 layers).

Method	mAP	FPS	batch size	Input resolution
SSD300	74.3	46	1	300 × 300
SSD512	76.8	19	1	512 × 512
SSD300	74.3	59	8	300 × 300
SSD512	76.8	22	8	512 × 512

The default *backbone* was based on **VGG16**, then replaced with **ResNet50**.

Furthermore, a subsequent approach was fine tuning starting from a pre-trained model.

MATERIALS AND METHODS – Settings

Three types of Single-shot Detectors were used: **SSD300**, **SSD512** and **SSD7**. The architecture of the first two is almost the same (9 and 10 layers respectively) while the SSD7 provides a simplified approach (7 layers).

		Batch size	Steps per epoch
VGG16	**SSD300**	32	260
	SSD512	16	520
ResNet50	**SSD300**	32	260
	SSD512	12	690

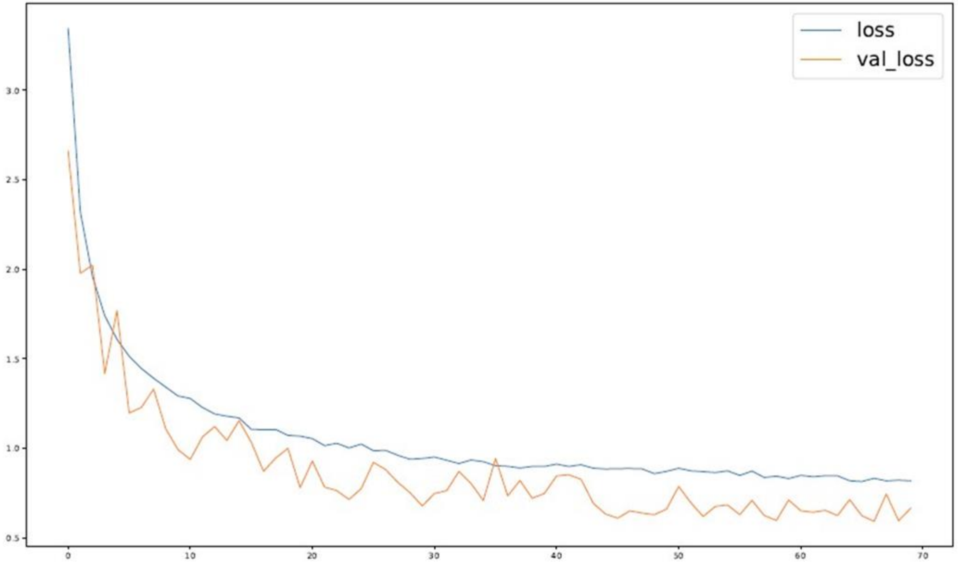
$$\text{Steps} = \frac{\text{Trainig set}}{\text{Batch size}}$$

Epochs = 40

Learning rate = 0,001

MATERIALS AND METHODS – Settings

Three types of Single-shot Detectors were used: **SSD300**, **SSD512** and **SSD7**. The architecture of the first two is almost the same (9 and 10 layers respectively) while the SSD7 provides a simplified approach (7 layers).



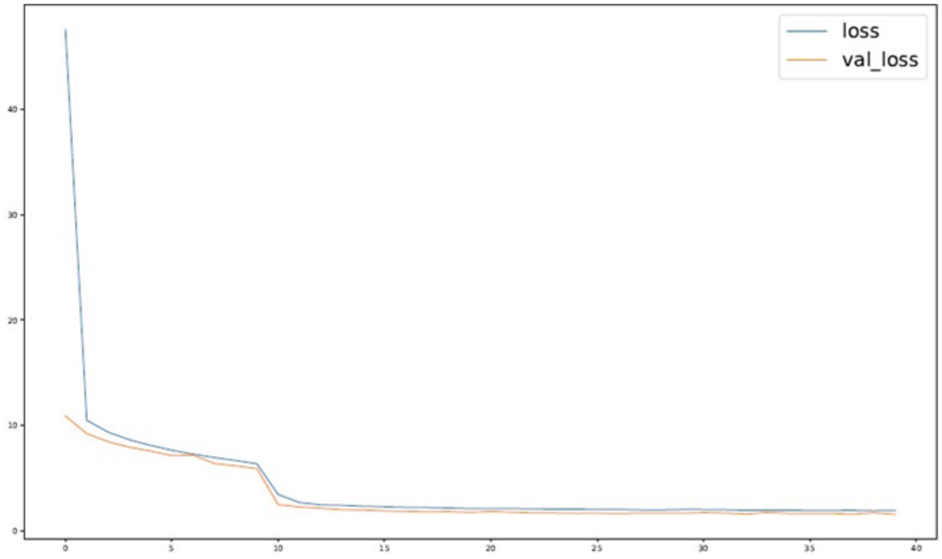
SSD7

Epochs = 70

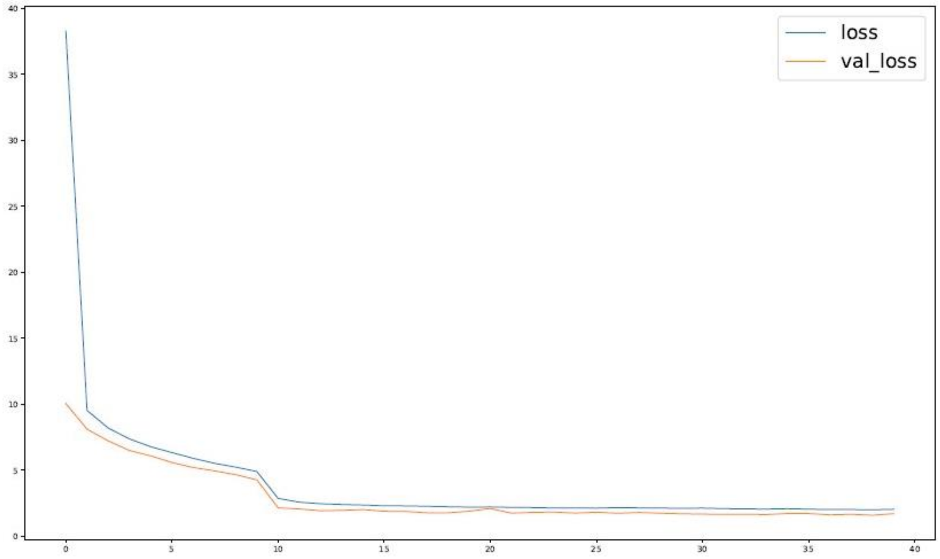
Learning rate = 0,001

MATERIALS AND METHODS – Settings

SSD300



SSD512



VGG16

UNIVERSITÀ POLITECNICA DELLE MARCHE

VRAI

MATERIALS AND METHODS – Loss Function

The **Loss function** consists of two terms: L_{conf} and L_{loc} where N is the matched default boxes.

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g))$$

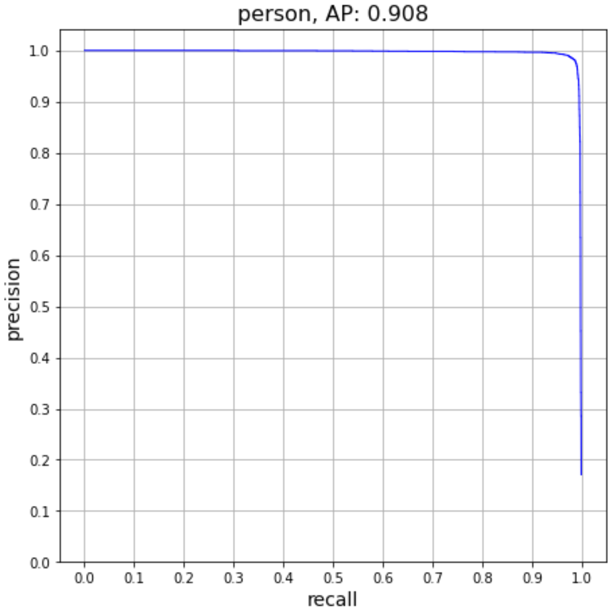
L_{loc} is the **Localization Loss** which is the smooth L1 loss between the predicted box (l) and the ground-truth box (g) parameters. L_{conf} is the **Confidence Loss** which is the softmax loss over confidences (c).

RESULTS AND DISCUSSIONS

The metrics used for the evaluation are **AP** (*Average Precision*), **Recall**, **F1 Score** and **IoU** *(Intersection over Union).*

SSD300	AP	Recall	F1 Score	IoU
VGG16	0,908	0,818	0,861	0,879
ResNet50	<u>0,909</u>	0,618	0,736	0,842

SSD512	AP	Recall	F1 Score	IoU
VGG16	0,908	<u>0,912</u>	0,910	0,801
ResNet50	<u>0,909</u>	0,872	0,890	0,846



SSD512 (VGG16)

UNIVERSITÀ POLITECNICA DELLE MARCHE

RESULTS AND DISCUSSIONS

The metrics used for the evaluation are **AP** (*Average Precision*), **Recall**, **F1 Score** and **IoU** *(Intersection over Union).*

SSD300	AP	Recall	F1 Score	IoU
VGG16	0,908	0,818	0,861	0,879
ResNet50	0,909	0,618	0,736	0,842

SSD512	AP	Recall	F1 Score	IoU
VGG16	0,908	0,912	0,910	0,801
ResNet50	0,909	0,872	0,890	0,846

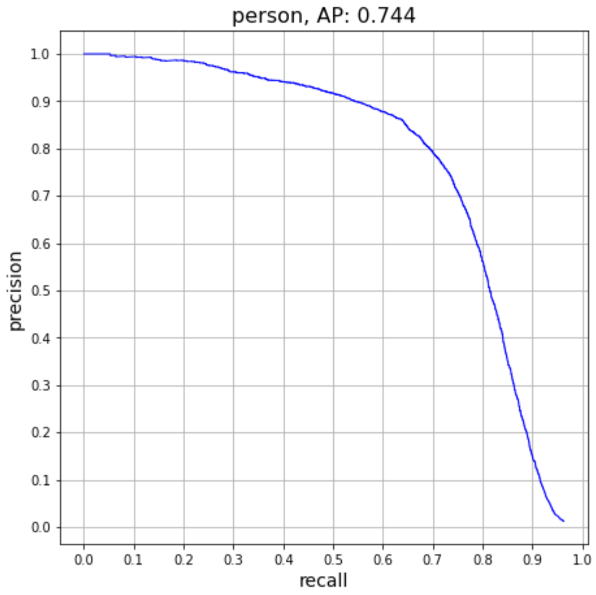
SSD512 (VGG16)

UNIVERSITÀ POLITECNICA DELLE MARCHE

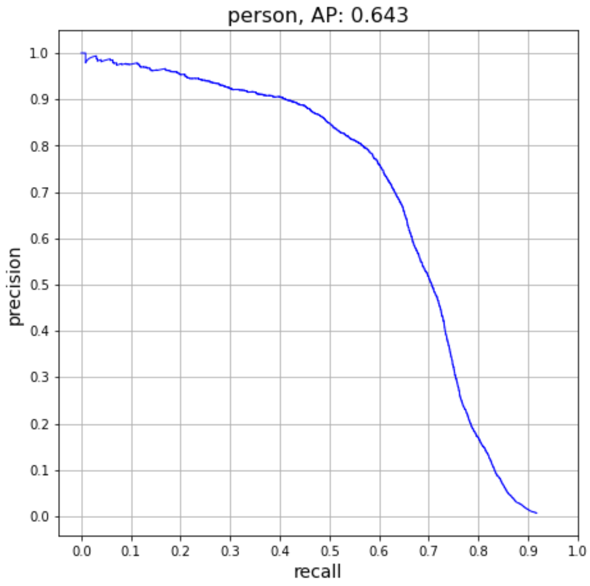
VRAI

RESULTS AND DISCUSSIONS

The SSD300 and SSD512 presented *weights* related to models already trained on the Pascal VOC 07 + 12 dataset; a **fine tuning** of the weights was made to go from 20 classes to the only <**person**> class.



SSD512

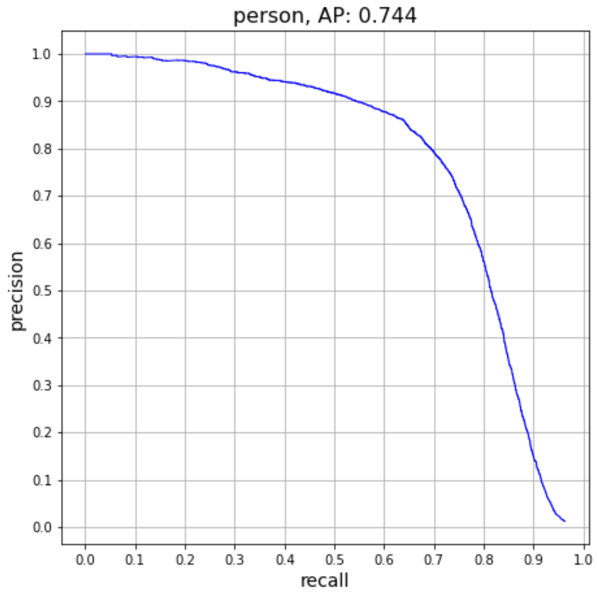


SSD300

UNIVERSITÀ POLITECNICA DELLE MARCHE
VRAI

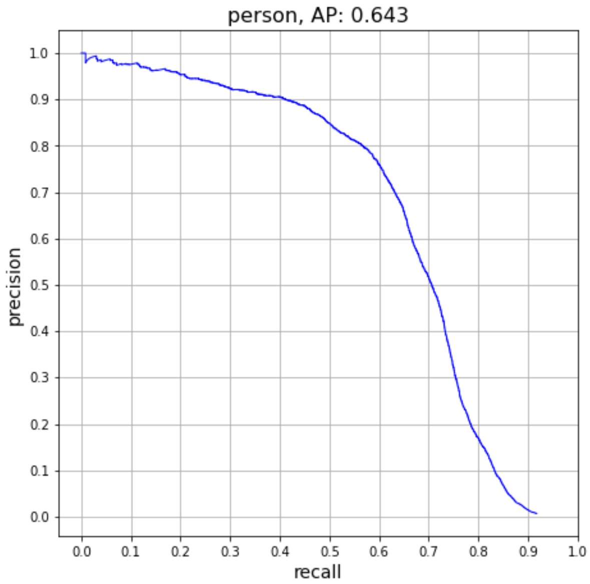
RESULTS AND DISCUSSIONS

The SSD300 and SSD512 presented *weights* related to models already trained on the Pascal VOC 07 + 12 dataset; a **fine tuning** of the weights was made to go from 20 classes to the only <**person**> class.



SSD512

FINE-TUNE	SSD300	SSD512
AP	0,643	0,744
Recall	0,887	0,934
F1 Score	0,746	0,828
IoU	0,712	0,885



SSD300

CONCLUSION AND FUTURE WORKS

- ✓ Labelling the Dataset (~17.000);
- ✓ Replacement of VGG16 with ResNet50;
- ✓ Training from scratch and fine-tuning;
- ✓ Average Precision not less than 90%.

The proposed approach, i.e. replacing the feature extractor, proved to be excellent to solve the detection problem. The best architecture remains the SSD512 with the VGG16.

CONCLUSION AND FUTURE WORKS

- Changing the feature extractor with other architectures ResNet-like, for example DenseNet;
- Changing the type of feature extractor, i.e. Feature Pyramid Networks (FPN)

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

REFERENCES

[1] Misbah Ahmad, Imran Ahmed, Kaleem Ullah, Iqbal khan, Ayesha Khattak, Awais Adnan, "Person Detection from Overhead View: A Survey".

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single Shot MultiBox Detector".

[3] Zhengxia Zou, Zhenwei Shi,Member, IEEE, Yuhong Guo, and Jieping Ye,Senior Member, "Object Detection in 20 Years: A Survey," IEEE.

[4] L. Jiao et al., "A Survey of Deep Learning-Based Object Detection," in IEEE Access, 2019, vol. 7, pp. 128837-128868.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp.580-587.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region based convolution with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

“Some people call it Artificial Intelligence but the reality is, this technology will enhance us. So, instead of Artificial Intelligence I think we will Augment our Intelligence”

Ginni Rometty, President & CEO of IBM