

Project Acronym: STAR
Grant Agreement number: 956573 (H2020-ICT-2020-1 – Research and Innovation Action)
Project Full Title: Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines
Project Coordinator: Netcompany-Intrasoft



Funded by the Horizon 2020
Framework Programme of the
European Union

DELIVERABLE

D5.6 – Visual Scene Analysis for Safety Zones Detection-Final version

Dissemination level	PU -Public
Type of Document	Demonstrator
Contractual date of delivery	31/03/2023
Deliverable Leader	THALES SIX GTS GRANCE
Status - version, date	Final – v1.0,
WP / Task responsible	WP5/T5.3
Keywords:	Visual Scene Analysis, Safety zones, localisation

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 956573. It is the property of the STAR consortium and shall not be distributed or reproduced without the formal approval of the STAR Management Committee. The content of this report reflects only the authors' view. The European Commission is not responsible for any use that may be made of the information it contains.

Executive Summary

The aim of this deliverable is to describe the final version of our demonstrator to detect moving elements in the project testbed at DFKI smart factory.

For this purpose, this deliverable illustrates some video analytics approaches based on object detection and classification. The deliverable also provides justifications to support the approach followed for the demonstrator development. The aim of the demonstrator described in this deliverable is to improve safety in the context of factories in which automatic or semi-automatic robots work together with humans.

The component presented here will extend the STAR system in order to analyse the scene and monitor the robots deployed in the next generation work floor, using a video analysis module detecting empty areas for secure robot displacements.

This system should be able to detect the obstacles to avoid collision, feeding another module presented in D5.7 that will dynamically provide a robot path.

The elements of the scene to detect are: moving items, static object/obstacle on the navigation path and human occupying the robot's neighbourhood.

Deliverable Leader:	THALES
Contributors:	Andreina Chietera, Jean-Emmanuel Haugeard, Solène Blasco Lopez
Reviewers:	Mihail Fontul (IBER) Spyros Theodoropoulos (UPRC)
Approved by:	Charalampos Ipektsidis, John Soldatos (INTRA)

Document History			
Version	Date	Contributor(s)	Description
V0.1	05/06/2023	Thales	TOC
V0.2	16/06/2023	Thales	First Draft of the deliverable available
V0.3	26/06/2023	Thales	Second Draft of the deliverable available
V0.4	05/07/2023	UPRC	First review
V0.5	10/07/2023	IBER	Second review
V0.6	13/07/2023	Thales	Integration of comments
V1.0	17/07/2023	INTRA	QA and creation of the final submitted version

Table of Contents

EXECUTIVE SUMMARY	2
TABLE OF CONTENTS.....	4
TABLE OF FIGURES.....	5
DEFINITIONS, ACRONYMS AND ABBREVIATIONS	6
1 INTRODUCTION.....	7
2 MODERN COMPUTER VISION APPROACHES FOR SITUATION AWARENESS	9
2.1 DETECTION OF OBJECTS OF INTEREST IN VIDEO STREAMS	9
2.1.1 <i>Moving Object Detection Using Background Modelling.....</i>	<i>9</i>
2.1.2 <i>Objects detection and classification Using CNNs</i>	<i>12</i>
2.1.3 <i>Object Identification Based on Few-Shot and One-Class Approaches.....</i>	<i>14</i>
2.1.4 <i>Human Detection Based on Deep Learning</i>	<i>15</i>
2.2 OBJECT GEOLOCATION USING 3D CALIBRATION	17
2.3 OBJECT TRACKING USING 3D GEOLOCALISATION	19
3 IMPLEMENTATION OF THE STAR DEMONSTRATOR	21
3.1 INPUT DFKI PLATFORM:	21
3.2 OUTPUT	24
3.2.1 <i>Heatmaps for the path planning implementation</i>	<i>24</i>
4 CONCLUSION.....	26
5 BIBLIOGRAPHY	27

Table of Figures

FIGURE 1: AN OVERVIEW OF THE INTEGRATION FROM THE POINT OF VIEW OF THE ARCHITECTURE. 7

FIGURE 2: VISUAL SCENE ANALYSIS COMPONENTS 8

FIGURE 3: BASIC STEPS FOR BACKGROUND SUBTRACTION ALGORITHMS – THALES LABORATORY EXAMPLE10

FIGURE 4: RESULTS OF DIFFERENT BACKGROUND SUBTRACTION ON CDNET DATASET11

FIGURE 5: SUBTRACTION EVALUATION.....12

FIGURE 6: OBJECT CLASSIFICATION USING YOLO ALGORITHM IN THE CONTEXT OF ROBOT-HUMAN COHABITATION. IN THIS RESULT, THE FINAL DETECTION IS ROBOT, HUMAN AND STOOL.13

FIGURE 7: RESULT OF ROBOTINO IDENTIFICATION (GREEN) USING RESNET FEATURES AND ONE-CLASS SVM APPLIED TO BACKGROUND SUBTRACTION’S DETECTED MOVING OBJECTS (BOUNDING BOXES)15

FIGURE 8: EXAMPLE OF HOG FEATURE ON THE STAR PROJECT IMAGE.....16

FIGURE 9: DEFORMABLE PART MODEL: MODEL FOR THE PERSON CATEGORY.16

FIGURE 10: THE FIGURE SHOWS AN EXAMPLE OF THE RESULTS OBTAINED USING DIFFERENT DETECTORS: 1) MOVING OBJECT DETECTION, (GMM SUBTRACTION) 2) OBJECT DETECTOR TO IDENTIFY THE STOOL AND THE ROBOT (YOLO), 3) HUMAN DETECTOR (OPENPOSE)17

FIGURE 11: SKELETON DETECTION AND 3D LOCATION WITH KEY POINTS (FEET, HIP, SHOULDER) PROJECTION.....18

FIGURE 12: SKELETON DETECTION AND 3D POSITIONS ALONG THE WHITE LINES GROUNDTRUTH19

FIGURE 13: COMPARISONS BETWEEN 3D POSITIONS ESTIMATED BY FEET PROJECTION, HIP PROJECTION, SHOULDER PROJECTION AND THE GROUNDTRUTH19

FIGURE 14: HUMAN TRACKING OVER TIME FROM 3D LOCALISATIONS OF ONE CAMERA’S DETECTIONS IN OUR THALES LABORATORY (LEFT: VIDEO FRAMES AT DIFFERENT TIME WITH ID DISPLAY – RIGHT: DETECTIONS OVER TIME (POINTS) WITH ESTIMATED MOTION (ARROWS)).....20

FIGURE 15: DFKI PLATFORM – 3 WORKSTATIONS, ROBOTS AND 2 CAMERAS21

FIGURE 16: THALES CALIBRATION TOOLS BASED ON VANISHING POINTS.....21

FIGURE 17: HUMAN 3D POSITION BASED ON POSE ESTIMATION MODEL AND CALIBRATION22

FIGURE 18: BACKGROUND SUBTRACTION - MOVING OBJECT DETECTION BASED ON SUBSENSE22

FIGURE 19: BACKGROUND SUBTRACTION - MOVING OBJECT DETECTION BASED ON GAUSSIAN MODEL. SOME FALSE ALARMS - SHINY REFLECTIONS IN LEFT WORKSTATION23

FIGURE 20: EXAMPLE OF HUMAN DETECTION, 3D GEOLOCALISATION AND TRACKING WITH MERGE FROM 2 CAMERAS. 24

FIGURE 21: EXAMPLE OF HUMAN AND OBJECT DETECTION, 3D GEOLOCALISATION, TRACKING (DONE SEPARATELY FOR OBJECTS AND HUMANS) AND ROBOT’S IDENTIFICATION WITH MERGE FROM 2 CAMERAS IN WHICH THE ROBOT IS MOVING.24

FIGURE 22: HEATMAP EXAMPLE25

FIGURE 23: SAFETY ZONES DETECTION DETECTS AND LOCATES HUMAN AND OBJECT AND PUBLISHES THE POSITION ON MQTT BROKER PRESENT IN THE HDT SYSTEM25

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AGMM	Adaptive Gaussian Mixture Model
AGV	Autonomous Ground Vehicles
CIFAR-FS	CIFAR100 few-shots (CIFAR: Canadian Institute For Advanced Research)
CNN	Convolutional Neural Network
DPM	Deformable Part Model
GMM	Gaussian Mixture Model
HDT	Human Digital Twin
HOG	Histogram of Oriented Gradients
R-CNN	Region Based Convolutional Neural Networks
SSD	Single Shot Detector
STAR	Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines
SVDD	Support Vector Data Description
VGG	Visual Geometry Group
YOLO	You Only Look Once

1 Introduction

Nowadays with the Fourth Industrial Revolution (or Industry 4.0), the automation of traditional manufacturing and industrial practices require the deployment of mobile robots that are involved to accomplish several tasks to assist workers in a modular production line. The robots are equipped with several embedded sensors (radar, camera) to analyse the nearby environment, in order to move safely and avoid obstacles. Unfortunately, this technology does not provide a dynamic global view of the work floor. Thus, the cohabitation between humans and robots remains unoptimised and can lead to a partial exploitation of the production line or worst to dangerous situations. The software we will describe in the following sections has the aim to detect dynamically security or empty zones throughout the infrastructure using a global situation assessment. For that, we will implement AI based algorithms to analyse the scene using the global point of view of the camera network already deployed in the factory.

Video analytics allows to automatically exploit the video streams in real time to detect anomalies and immediately raise an alarm. To this end, the algorithms detect and track elements of interest (such as people, robots, and new objects occupying the scene) over time, and alert the robots of the presence of any obstacles in the surrounding area. Whenever a human is detected close to the robot, his movements will be monitored. Based on a human behaviour analysis, the system will decide whether a new robot's path should be calculated to reach the docking station or to stop completely to avoid any collision.

To improve the understanding of the global picture of the STAR project aim, the Safety Zones Detection System, presented in this deliverable, will complete the global awareness of the factory proposed in the WP5 of the project. Indeed, it will be integrated into the HDT system described in D5.1 and presented in the figure below (Figure 1). To support the Privacy by Design principle, the Safety Zones Detection System will publish on the HDT system only computation results as "spatial heatmaps" containing information on objects, Robotino and workers' positions in the work floor on the IoT Middleware to update the HDT picture.

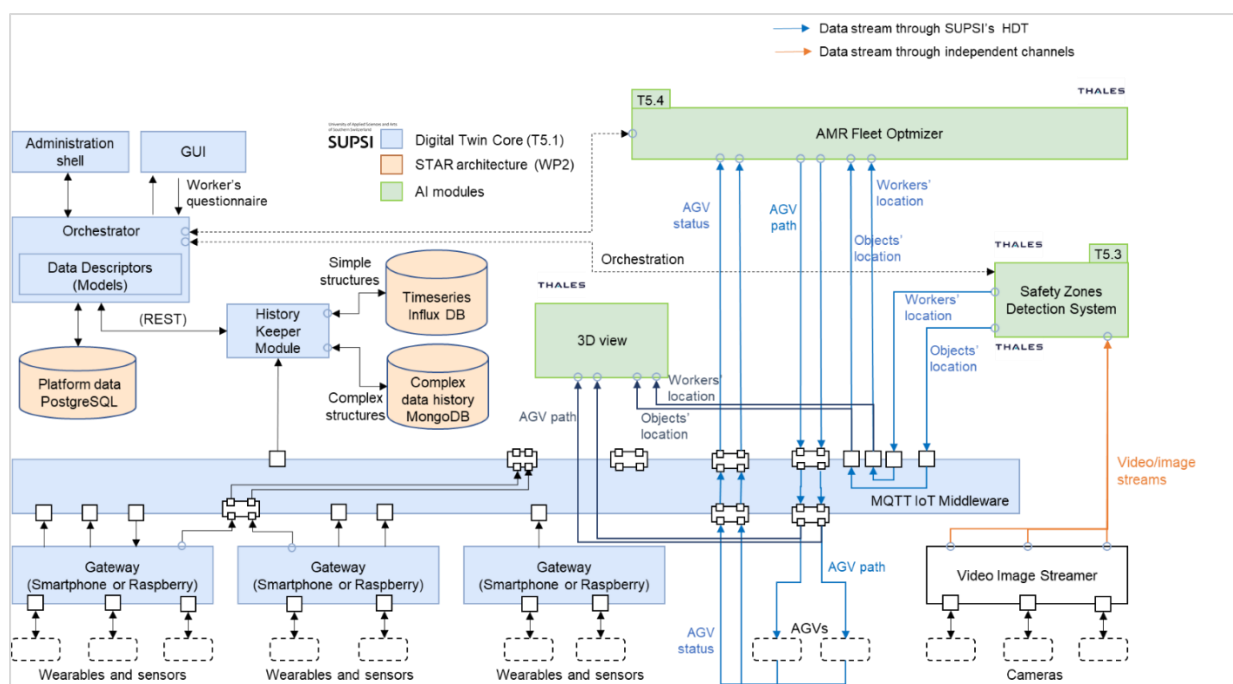


Figure 1: an overview of the integration from the point of view of the architecture.

Besides this integration with the HDT system to complete the global picture of the factory, the results of this video awareness system will feed directly to the AGV (Autonomous Ground Vehicles) Fleet Optimizer that will be presented in D5.7. The latter module will access the heatmaps via the HDT as shown in the figure above. This data will allow the AGV module to dynamically define and update the best paths for the AGVs, avoiding obstacles and introducing factually safety when humans and robots share the same spaces. From the HDT Middleware, the AGVs will have access to the results of this module and consequently, they will adapt their behaviour.

Specifically, the Safety Zones Detection System exploits video footprints as input and will deliver the spatial heatmaps as results of the analytics. This process combines 2 main components as presented in Figure 2:

- The elements extraction module;
- The 3D object localisation.

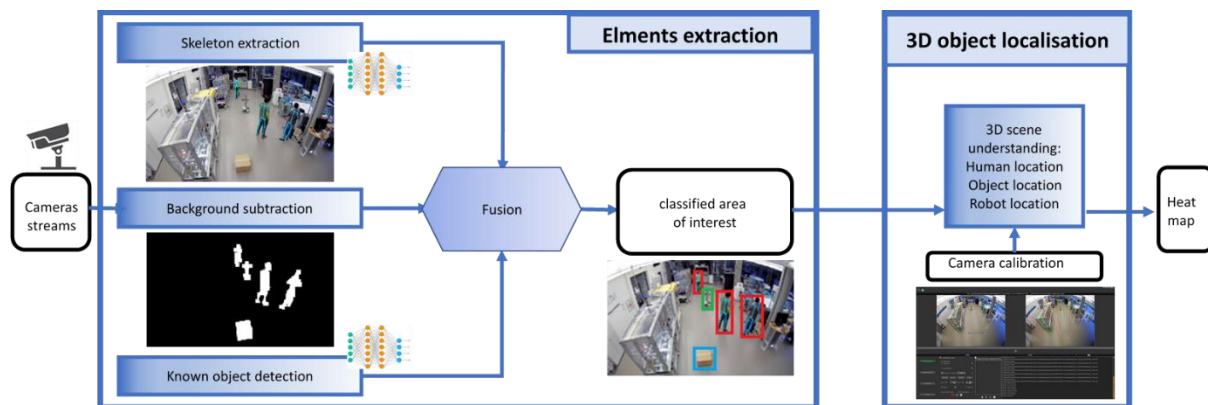


Figure 2: Visual Scene Analysis Components

Particularly, the elements extractor engine merges two deep learning algorithms, one for the skeleton reconstruction to follow the human gesture and pose and the other for the detection and classification of the non-static object in the scene, with a background subtraction module.

This latter method allows to assess the difference between the background model and the current image to infer moving elements in the scene under observation.

The 3D object localisation takes as input the results of the elements extractor in order to localise them using the Euclidian reference system. To achieve this task a calibration software is developed to make a correspondence between the camera pixels and the physical world.

The following sections, after an introduction of the most common computer vision approaches, will describe the component and the methodologies implemented to realise the final prototype of the Safety Zones Detection System.

2 Modern Computer Vision approaches for situation awareness

One of the main goals of STAR is to ensure the optimisation of a production line to increase the efficiency of the manufacturing process. It is considered that efficiency and safety go hand in hand in a complex environment such as the production lines, in which operators, robots and automatic systems share dynamically the same physical workspace.

The aim of this module is to take advantage of modern computer vision approaches in order to recognise the postures and motion of workers and locate them as well as the items positioned in the environment. The main output will be an “average spatial heatmap” representing a probabilistic occupancy of the production lines based on fixed RGB cameras deployed in the factory. The purpose of this module is to feed a “planner” indicating dynamically which areas should be avoided by the robots’ fleet operating in the production lines.

The solution we imagine is conceived by merging the following technologies:

- Detection of objects of interest in video streams:
 - Moving object detection using background modelling
 - Dynamic object detection via a convolutional neural network (CNN)
 - Skeleton extraction by human pose detection CNN
- 3D Object geo-localisation and motion in the infrastructure and estimation of human-robot distances using the geometric calibration of fixed RGB cameras.

2.1 Detection of Objects of Interest in Video Streams

2.1.1 Moving Object Detection Using Background Modelling

The image segmentation into background regions and moving objects is a crucial stage in these video applications. The segmentation result is often used as an input for object detection/classification. Background subtraction methods are based on the premise that the difference between the background model and the current image is due to the presence of moving objects in the scene under observation.

The proposed approaches are based on background modelling of the observed scene (“background”) as a first step, then on the analysis of the differences between each image and the estimated background (Figure 3).

The foreground segmentation is possible under certain conditions:

- The camera is static (properties do not change);
- The background is statically visible most of the time;
- The background is quasi-stable and can be modelled statistically over time;
- Objects of interest are different (colour/texture) from the background model in order to detect the difference between the current image and the background model.

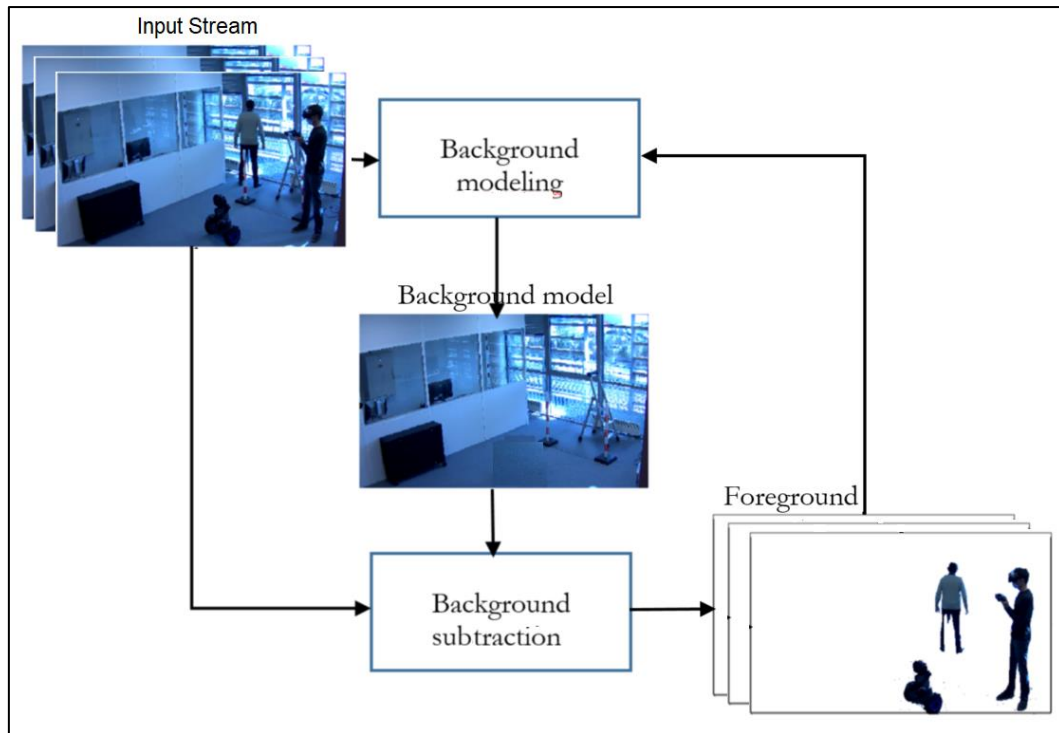


Figure 3: Basic Steps for Background Subtraction Algorithms – Thales laboratory example

A detailed survey of various background modelling methods in video analysis applications can be found in [1]. The background subtraction approaches can be divided in 4 categories:

- Basic methods: define the background as the mean or median of the observed values;
- Filtering methods (Wiener filter [2], Kalman filter [3], ...): design dynamic backgrounds by adapting the model using a filter;
- Clustering methods (K-Means [4], Codebooks [5], ...): compare the current pixel and the different clusters at every point in the image;
- Stochastic methods (Gaussian model [6], Gaussian mixture model – GMM [7], Kernel density estimation – KDE [8], ...): use probabilistic modelling of the background.

The principle of background subtraction methods is to discriminate the pixels of moving objects (foreground) from those of the static scene (background), by detecting pixels which are significantly deviating from the model of the frame sequence. Stochastic methods (GMM approaches) are more commonly used in the video applications. For example, Adaptive Gaussian Mixture Model (AGMM) uses Gaussian distribution models for background and foreground pixels and update the models continuously.

The background subtraction allows to extract the “foreground” of the scene, namely the silhouettes or contour of new or moving objects (people, vehicles, objects newly occupying the camera point of view) in the scene, but also extracts areas in which lighting changes appeared due to the variations of the lighting conditions during the day. Moreover, if the objects are close to each other and/or they hide each other, their silhouettes are merged together as a single element and the resulting foreground is difficult to analyse by its shape. This type of approach, therefore, makes it possible to detect all the changes in the scene,

which may correspond to the presence of a new element (objects or people), but also to the presence of moving people/robot.

2.1.1.1 The STAR approach and development Status

In the STAR project, we have implemented and evaluated several state-of-the-art solutions on the ChangeDetection dataset (CDNet dataset - Figure 4).

To evaluate the results (Figure 5), we use commonly used metrics and compare each of the predicted masks with the ground truth for a given dataset. We select 3 metrics:

- **Precision** describes the purity of our positive detections relative to the ground truth;
- **Recall** describes the completeness of our positive predictions relative to the ground truth;
- **F1-score** is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{Recall} = \frac{TP}{TP+FN} \quad \text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$



Figure 4: Results of different background subtraction on CDNet dataset

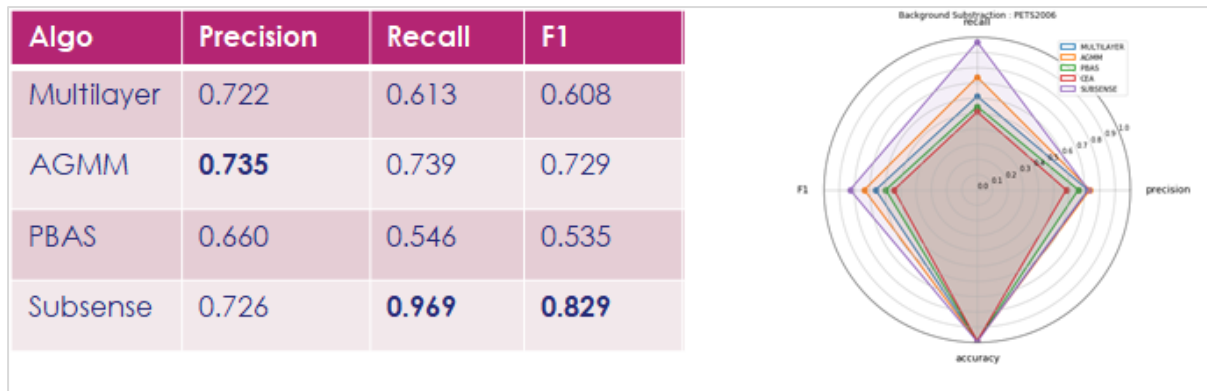


Figure 5: Subtraction Evaluation

The best results (Figure 5) are obtained by approaches based Gaussian model (GMM) and the SuBSENSE approach. The SuBSENSE [9] method combines colour and local binary similarity pattern (LBSP) features to improve the spatial awareness. In the rest of the project, we have chosen a method based on GMM and SuBSENSE methods.

2.1.2 Objects detection and classification Using CNNs

Today, as in the field of image classification, object detection approaches are all based on Convolutional Neural Network architecture (CNN). These solutions based on CNN architecture consist of two parts: a "feature extractor" called backbone and a "feature classifier". In the field of object detection based on deep learning [10], the architectures usually can be divided into two categories: two-stage and single-stage approaches.

- Two-stage detector

Two-stage networks use the "Region Proposal Network" algorithm as a first step to quickly select the best candidate windows. These windows (from a few hundred to a few thousand) are then processed by a classification model (the second step) to decide whether or not they contain an object from the list considered. The most cited examples are the R-CNN model (Regions with CNN features [11]) and its derivatives: Fast R-CNN [12]), Faster R-CNN [13] and Mask R -CNN [14].

- One-stage detector

The one-stage detectors propose predicted boxes from input images directly without the region proposal step, thus they are time efficient and can be used for real-time applications. The one-stage detectors apply the classification directly to dense window grids ("anchors") of different sizes (cf. Figure 6). The two main representatives of this family are the YOLO model (You Only Look Once [15]) and its derivatives: Yolov2, Yolov3 ([16]), Yolo9000, and the SSD model (Single Shot Detector [17]).

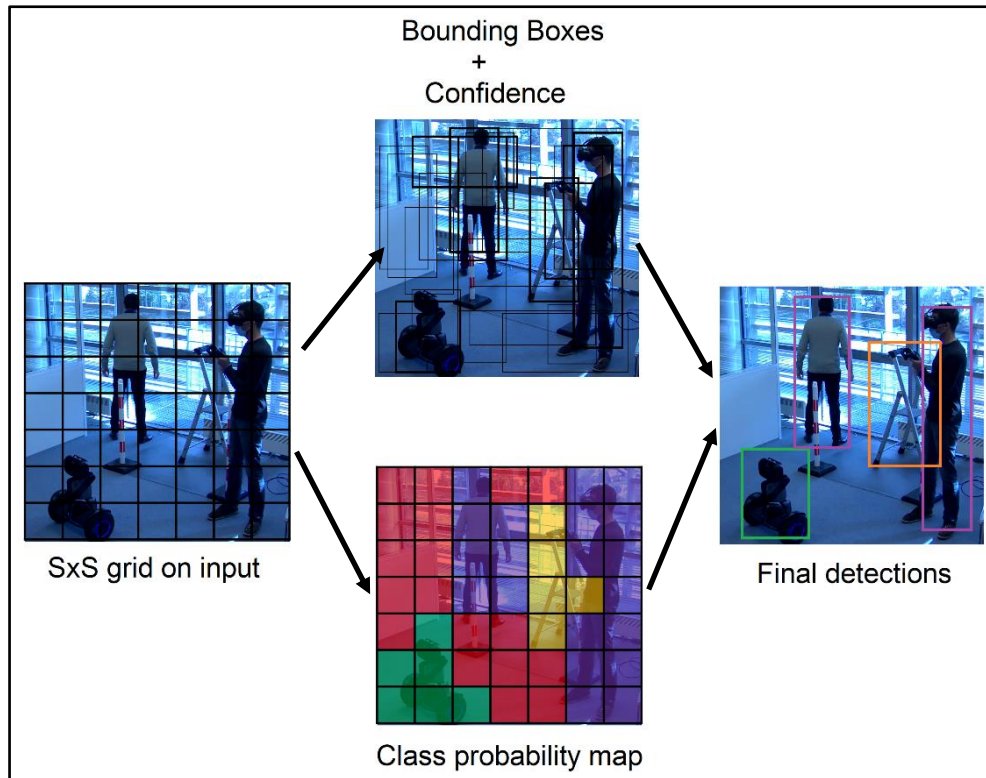


Figure 6: Object Classification using YOLO Algorithm in the context of Robot-Human Cohabitation. In this result, the final detection is robot, human and stool.

The performance of these models depends on their own architecture (meta-architecture), and from the backbone one. Among the CNNs most used in this role, there are two families of state-of-the-art models for classification: VGG16 and its derivatives [18] and ResNet models (Deep Residual Learning [19]).

As a result, most of these techniques sketches, for each object detected, a rectangle called a "bounding box" surrounding the object in the image. The main exception is Mask R-CNN, which additionally provides the "mask" as the shape of each object detected, consisting of all the pixels belonging to the object in the image.

2.1.2.1 The STAR approach and development Status

Currently, deep learning methods obtain state-of-the-art performance on topics as varied as facial recognition, vehicles tracking/identification.... To obtain quality performance on the task(s) that a neural network tries to "learn", it is then necessary to have at its disposal a corpus of labelled data of sufficient size. With a large dataset, the optimisation of the parameters can converge on a stable state and the new model could be generic and used on any other corpus reasonably different from that used in training. However, in our STAR use case, we do not have a lot of labelled data on new objects (robots, workstation...). In the next months, we are therefore going to set up approaches based on few-shot learning. The objective of "few shot" learning is to compensate for and to tackle the scarcity of examples available for a given problem.

2.1.3 Object Identification Based on Few-Shot and One-Class Approaches

In the context of human-machine cohabitation, all moving objects must be detected as potential obstacles, with the additional identification of some objects of interest (robots, workstation...).

To be as general as possible and class-agnostic in the detection, we adopt a two-step detection strategy: first, background subtraction detects and predicts bounding boxes for any type of moving object, then a classifier trained with a one-class strategy identify an object of interest against any other potential class. One-Class approach can distinguish one class against “the rest” of possible ones. The basic assumption is often to only use the class of interest during training: the inference phase then relies on a learned decision boundary (One-Class SVM...) or a fixed or learned similarity metric (Clustering, Feature extraction approach...). A meta-dataset representing the negative class may be required for supervised deep learning methods (One-Class CNN).

However, we only have few examples of targeted objects of interest at our disposal, which is not sufficient to train a robust and generic deep learning method. We then rely on “few-shot” approaches for the classification step: this is a supervised learning framework in which few annotated data is available for training. Data limitation is compensated by some prior knowledge:

- Prior knowledge of data mainly corresponds to data augmentation strategies;
- Prior knowledge of algorithms can ease initialisation and optimisation of models’ parameters via a “meta learning” learning approach to “learn how to learn”: the generality and diversity learned by a pre-trained model or feature extractor through other data or tasks is often leveraged, such as Resnet and VGG features for images;
- Prior knowledge of models is the main lever of few-shot learning for model design: the main approach consists in building a small “support”-dataset from the labelled data, keeping only a few samples for each class. The class prediction is then based on the comparison with these samples in a latent space, using extracted features and pre-defined (Matching Networks [20], Prototypical Networks [21]) or learned (Relation Networks [22]) metrics.

2.1.3.1 The STAR approach and development Status

Few shot learning approaches are often applied to multi-class classification with pre-selected classes. One-class tasks in a few shos context are very little explored in literature, and are often pre-trained via Meta learning on a large dataset.

We experimented with the use of generic features extractors (Histogram of Gradients, Resnet...) combined with one-class models able to identify an object of interest against any other one via a learned decision boundary (One-Class SVM with Gaussian kernel) or a similarity metric with learned centroids (K-Means Clustering with Cosine Similarity). Deep image features such as Resnet are more robust than shape features such as HoG. The different possible aspects of Robotino are handled by diversity in collected training samples (with/without tray, different viewpoints...) but is still limited by the scarcity of data. False positives also appear, especially regarding background subtraction’s output format containing noise or little object’s patches (hand, foot, piece of head, background patch...). This may be

addressed in STAR demonstrator by a filtering based on shape and/or based on tracking consistency.

Other experiments were done with a model designed for few-shot one-class classification: Meta-SVDD [23], which adapts the methodology of a Prototypical Network to a one class task. The model is trained from scratch on a multiclass dataset “to learn how to learn” relevant features via the classification performance of classical One-Class algorithms (SVDD...). We succeeded in reproducing the performance of this paper on CIFAR-FS (100 classes, 5-shot setting), but a lot of false positives are obtained in practice on background subtraction’s patches. This could be due to a lack of diversity or very low resolution of this meta-dataset causing a lack of expressiveness of the obtained features.

We also tried to better adapt to the specific patches’ output format of background subtraction by using some of them as negative class during the training of a One-Class CNN [24] in order to finetune Resnet features or learn features from scratch with a small architecture. One-Class CNN is trained as a Binary classifier (Binary Cross Entropy Loss) with additional Compactness Loss (Features’ variance among the class of interest) and other classic architecture choices to avoid overfitting (LayerNorm, Dropout...). However, the scarcity of data and lack of diversity of the negative patches we generated from available STAR videos lead to over-learning of the studied scene and quick loss of the initial generality of Resnet feature extractor. Our measures against overfitting on the classical One-Class approach are not sufficient to compensate for the lack of data.



Figure 7: Result of Robotino identification (green) using Resnet features and One-Class SVM applied to background subtraction’s detected moving objects (bounding boxes)

2.1.4 Human Detection Based on Deep Learning

Flexible object (e.g., a person's body) can take multiple appearances in the image. This characteristic makes the task of detection/classification more complex. From the 2010s, research laboratories worked on methods based on the shape of objects of interest merged with machine learning techniques to be able to take into account all the possible configurations of the shape (feature templates - Deformable Part Model (DPM) [25] Figure 9). These techniques relied on the use of local attributes (descriptors) such as Histogram of Oriented Gradients (HOG [26] Figure 8), and could be a stand-alone solution or could be applied in combination with a background subtraction method to decrease false negatives. This learning-based approach has seen significant improvements with the advent of Convolutional Neural

Network CNNs, and their adaptation to object detection. The main problem with the techniques proposed in the factory context is the lack of robustness when partial occlusion occurred. Especially in a production line, the occlusion affects the people’s detection making the task more complex.



Figure 8: Example of HOG feature on the STAR project image

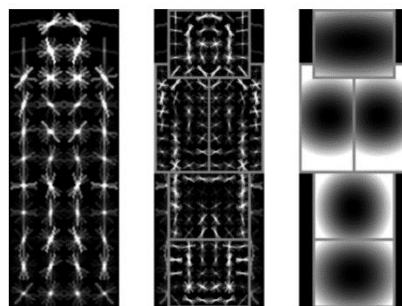


Figure 9: Deformable Part Model: model for the person category.

A technique for people detection, called OpenPose, was recently proposed by [27] and considers both the variability of the shapes observed (due to the fact that people are articulated objects) and the presence of partial occlusions. OpenPose is based on a CNN architecture and makes it possible to detect different characteristic points of the human body (joints, eyes, mouth, nose, ears, hands, feet) and, jointly, to group these points in a graph forming a skeleton representation (cf. Figure 10). More specifically, the skeleton detection algorithms allow to track human poses by detecting and estimating the position of the characteristic points defining human postures. The approach creates heat maps for joint extraction and extracts affinity fields considering all the detected joints in order to infer the link between them and, consequently, allow the detection of human limbs. The algorithm can simultaneously process different observation scales. It should also be noted that it can detect people both by their silhouette when it is clearly visible and by their head, which is more rarely masked.

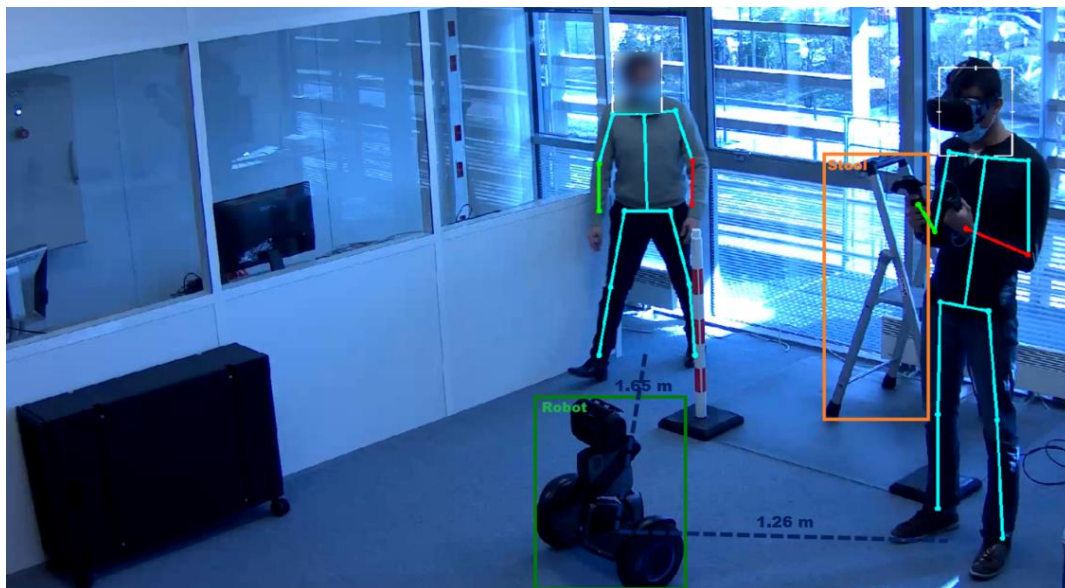


Figure 10: the Figure shows an example of the results obtained using different detectors: 1) moving object detection, (GMM subtraction) 2) object detector to identify the stool and the robot (Yolo), 3) human detector (OpenPose)

2.1.4.1 The STAR approach and development Status

In the STAR project, we chose to implement and improve the solution based on Human Pose estimation like OpenPose. This real-time approach is accurate and robust to occlusion. With this approach, we don't need to see the whole person to detect them. In addition, this approach also manages to estimate the posture of workers and follow his gesture to improve future functionalities.

2.2 Object Geolocation Using 3D Calibration

Once people have been detected by the previous algorithms in the 2D images, these people must be located in the real 3D infrastructure. More exactly, the 3D position corresponding to the intersection of the main axis of the person with the plane of the ground is estimated. This absolute location by video analysis of all the people present in a video stream requires a calibration phase. This is the geometric calibration of the camera, to associate each pixel of the image with absolute Cartesian coordinates, assuming that these pixels are on the same plane (the ground in our case). Thus, with calibration parameters, the 2D position in the image of an object will provide its absolute 3D position.

Several calibration methods [28] to determine the intrinsic and extrinsic parameters of the camera were tested: a fully manual method, a semi-automatic and a fully automatic. The more automated the calibration, the lower the accuracy. Moreover, the geometric distortions are estimated only by the manual method, by presenting the system with a checkerboard pattern.

One of the conclusions of our experiments is that the distortions (mainly radial) of the camera optics are difficult to estimate. If they are neglected or incorrectly estimated, the localisation accuracy is strongly degraded in wide-angle (wide field) cameras. This is not so obvious when the field of view is tighter. To be effective, the narrow-field camera must then adopt a more plunging point of view so as to avoid excessive occultation which, combined with the tight

field, would induce excessive location inaccuracies. There is therefore a compromise between installing many cameras with a narrow field of view, having little distortion, and installing fewer cameras with a wide field, but whose distortions must be finely estimated if we want to avoid degraded performance in the borders of the field.

Once this calibration has been carried out, it is also necessary to estimate the height of the key points of the body, to project it correctly on the ground.

The main hypothesis taken is that the individual is 1m75 tall. As a result, the feet are supposed to be on the ground, the hip at a height of 88cm, the shoulders and the neck at a height of 1m52. Each skeleton detection is so projected on the ground, according to the measurements indicated above. The position on the ground makes it possible to go up to the effective 3D position. The image below (Figure 11) illustrates the principle and the results compared to the calculated vertical of each individual.



Figure 11: Skeleton detection and 3D location with key points (feet, hip, shoulder) projection

In order to evaluate the accuracy of our approach (and our hypothesis), we have carried out several tests with around twenty people (people of different heights). During these tests (Figure 12 and Figure 13), we measured the drift between the estimated position with our approach and the groundtruth (the reality). As illustrated in the figure 12, people walk in different positions known (measured by a rangefinder). For example, in figure 12 (white lines), the person moves along axis $y=5$ m and then $x=2.5$ m. In these tests, the position is estimated in 3 different ways: feet projection on the ground, hip projection and shoulder projection. Given that the camera calibration is based on the ground plane estimation, the estimated positions are more accurate with points of interest close to the ground (feet).

The geolocation precision depends on:

- the number of persons
- their height compared to the hypothesis
- their distance to the camera

- camera angle and distortion.

The drift measured during our tests with these different configurations is between 15 to 60 cm. In order to be more robust, and to correct the drift, a temporal smoothing of the positions is carried out and we can average with other measurements from other sensors (other cameras...)



Figure 12: Skeleton detection and 3D positions along the white lines groundTruth

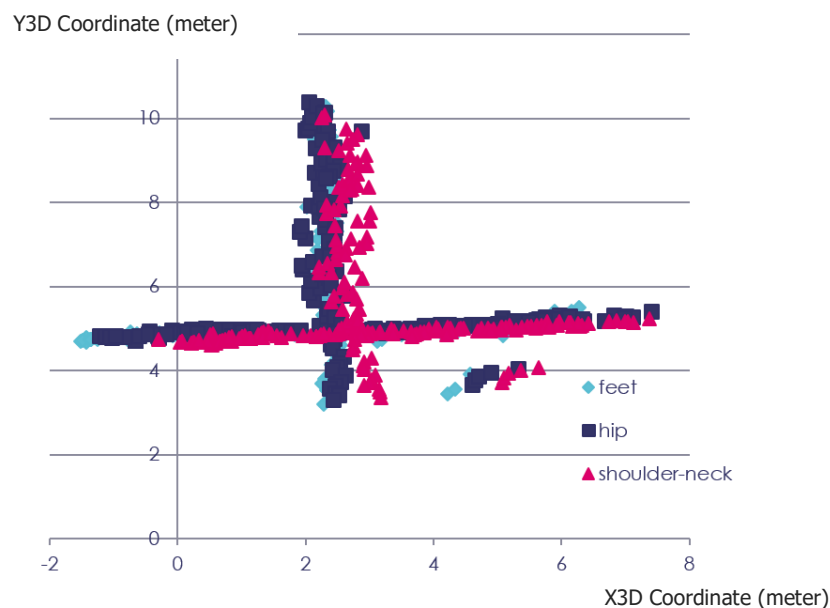


Figure 13: comparisons between 3D positions estimated by feet projection, hip projection, shoulder projection and the groundTruth

2.3 Object Tracking Using 3D Geolocalisation

Once people and other moving objects of interest (robots, obstacles...) have been detected and geolocalised in the real 3D infrastructure by the algorithms of previous sections, a tracking can additionally be performed to assign them a consistent id over time. Being able to model

and predict their trajectory will also ease the planning of robots' trajectories within the infrastructure.

The tracking of previously detected objects ("Tracking by detection") is often performed directly from 2D RGB images in literature. SORT [29] algorithm and its derivate perform tracking via a Hungarian Assignment, using a distance criterion based on bounding boxes' overlapping (IoU, Euclidean distance between centroids...). Kalman Filters are then used for motion estimation of each track before performing the assignation with new detected points. This allows next state's prediction regarding estimated motion, and strengthens the tracking against fast motions or occlusions. Tracking methods including Deep Learning now achieve state of the art performance on this task by adding efficient appearance descriptors that strengthen the tracking (DeepSORT...).

In the STAR use case, people and other moving objects are geolocalised with 3D coordinates corresponding to their position in the real 3D infrastructure. Locations from the detections of several cameras are merged to obtain a better estimation of localisation and to be able to better cover the area of interest. This fusion is done through a Hungarian Assignment using Euclidean distance as cost matrix.

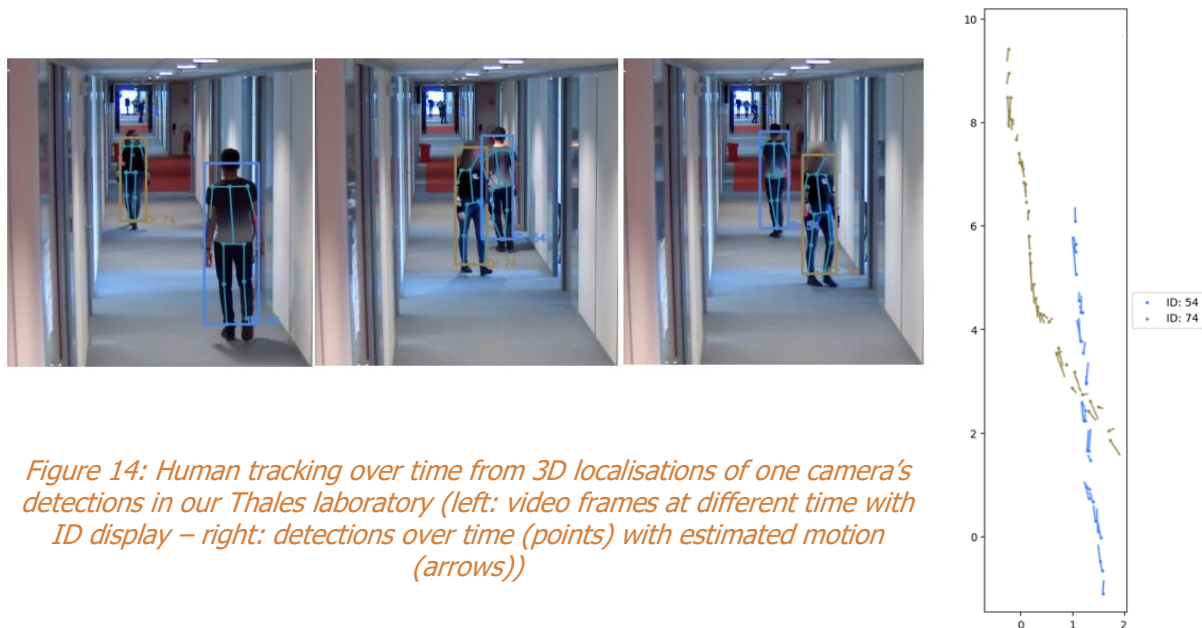


Figure 14: Human tracking over time from 3D localisations of one camera's detections in our Thales laboratory (left: video frames at different time with ID display – right: detections over time (points) with estimated motion (arrows))

We perform tracking in the 3D space after this fusion in order to benefit from the resulting localisations' smoothing. We adopt a similar methodology to the SORT algorithm on these 3D points: at each time step, the tracks' new position is estimated using Kalman Filters, before assigning them to the new detections of the current frame and updating Kalman Filters' parameters. Some additional rules have been designed for track management, such as creation (new track considered as consistent after a chosen number of consecutive frames), suppression (track suppression after being lost for a chosen number of consecutive frames) or priority (tracks considered as consistent by creation rule matched in priority with new detections) rules. In particular, the priority rule showed a diminution of id swaps or changes, in particular in the case of false or double parasitic detections.

3 Implementation of the STAR demonstrator

3.1 Input DFKI platform:

The DFKI platform is a manufacturer-independent demonstration and research platform. This platform is equipped with 3 workstations, several robots and 2 cameras (cf. Figure 15).



Figure 15: DFKI platform – 3 workstations, robots and 2 cameras

Each camera has been calibrated using the Thales tool (Figure 16) developed in the framework of STAR, which allows fast calibration with just a few clicks (horizontal lines, vertical lines, yardstick).

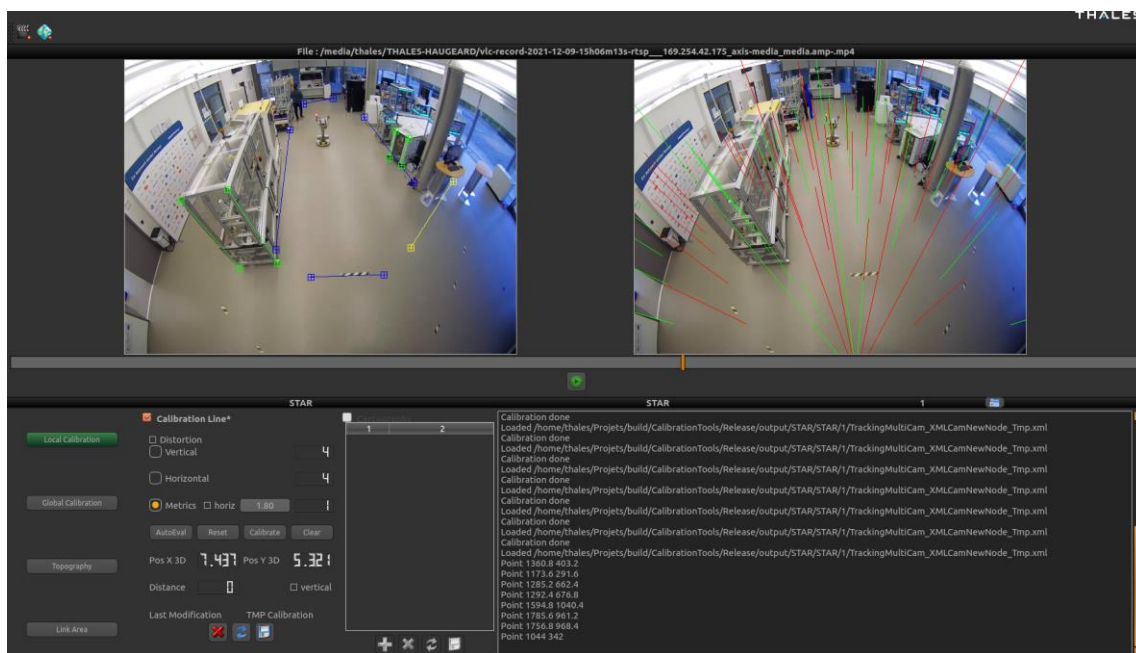


Figure 16: Thales Calibration tools based on vanishing points.

The first results (Figure 17) of human detection and 3D position are promising. However, when people are too hidden, the person detections become difficult. In order to tackle this problem, we need to add people tracking for each camera and merge these from multi-cameras results.

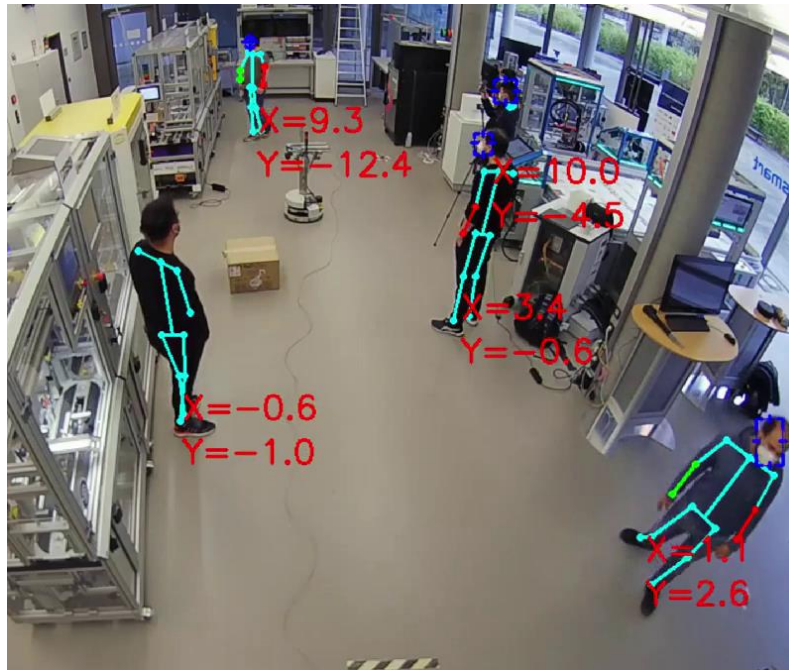


Figure 17: Human 3D position based on pose estimation model and calibration

In addition, in order to detect moving objects present in the scene, we have implemented two methods: GMM and SuBSENSE. These methods are able to learn a dynamic background static model on a video footprint and they are updated using the last frames. Thanks to these approaches the STAR system for safety zone detection detects accurately moving objects present in the scene. GMM has some false alarms like shiny reflections in the workstation and shadows on the ground. SuBSENSE has fewer false alarms but is more time consuming.



Figure 18: Background subtraction - moving object detection based on SuBSENSE



Figure 19: Background subtraction - moving object detection based on Gaussian model. Some false alarms - shiny reflections in left workstation

Identification of an object of interest, such as Robotino in our implementation, from a few numbers of available training examples is performed via deep features' extraction (Resnet) combined with a One-Class learning method (One-Class SVM or K-Means Clustering with Cosine Similarity). Using pre-trained features extractor compensates the lack of data to obtain generic representations, while One-Class learning strategy allows an identification of the object of interest against any possible class. This approach can be reproduced for other objects of interest, and seems the most generalisable under the joint constraints of data scarcity and other potential object classes' agnosticism. However, this identification of objects of interest has shown limitations in areas where they are too hidden or occluded. Merging multiple viewpoints and adding tracking over time should add robustness (see Figure 20 and Figure 21).

Finally, the objects and people thus detected are geolocated in real 3D infrastructure, using our 3D calibration and merging locations from multiple viewpoints. This allows a complete geolocalisation of all objects and people in the area of interest, with a matching of unique states in the overlapping areas (see Figure 20). Note that a quite precise calibration must also be made so that the locations sufficiently correspond between the different cameras in the entire overlapping zones.

A tracking is then performed over time in the 3D space adapting SORT methodology. This tracking includes Kalman Filtering for motion estimation, which allow future position's prediction for each track. In practice, creation, deletion and priority rules have been implemented as described in previous sections. It demonstrated more consistency and robustness in our experiments, in particular to compensate for object identification's limitations. However, detection of people and objects is still difficult in obstructed areas generally covered by a single camera, such as the outer part of the room in our setting.



Figure 20: Example of human detection, 3D geolocalisation and tracking with merge from 2 cameras.



Figure 21: Example of human and object detection, 3D geolocalisation, tracking (done separately for objects and humans) and robot's identification with merge from 2 cameras in which the robot is moving.

3.2 Output

3.2.1 Heatmaps for the path planning implementation

The main output is an “average spatial heatmap” representing a probabilistic occupancy of the production lines based on fixed RGB cameras deployed in the factory. We represent the results of the implementation presented in the previous sections using anonymised heatmaps. This description represents the global environment, including human and object location as occupancy cells. The results from object/human detection and localisation allow to update the heatmap representation. Indeed, these heatmaps are simplified and anonymised representations of the occupancy of the work floor in real time. It serves to feed the reinforcement learning module and forecast Robotino trajectories in order to avoid potential crowded areas of the factory and thus avoid collisions. The heatmaps are then sent to the HDT module and will be used as input by Reinforcement Learning module that will perform AMRs Fleet control (cf. Figure 22 and STAR D5.8: Reinforcement Learning Techniques for AMRs)

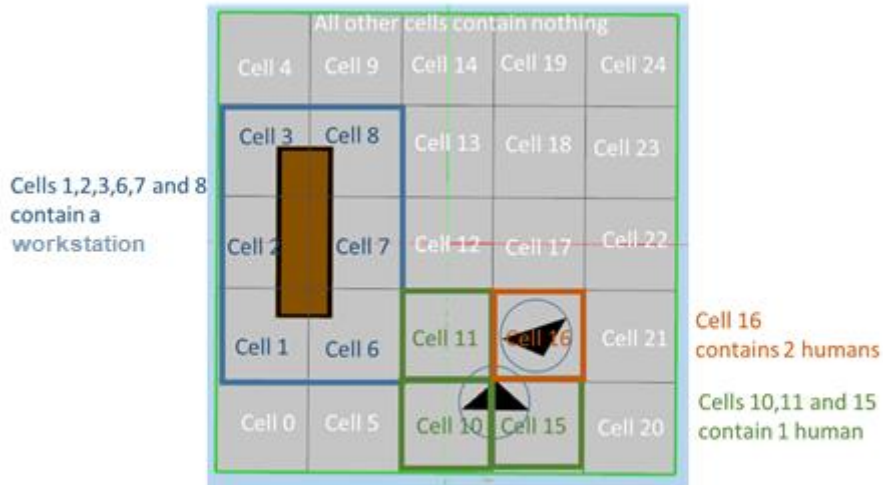


Figure 22: Heatmap example

To support the Privacy by Design principle, the Safety Zones Detection System publish on the HDT system only computation results as “spatial heatmaps” containing information on objects and workers’ positions in the work floor on the IoT Middleware in order to update the HDT picture (see Figure 23).

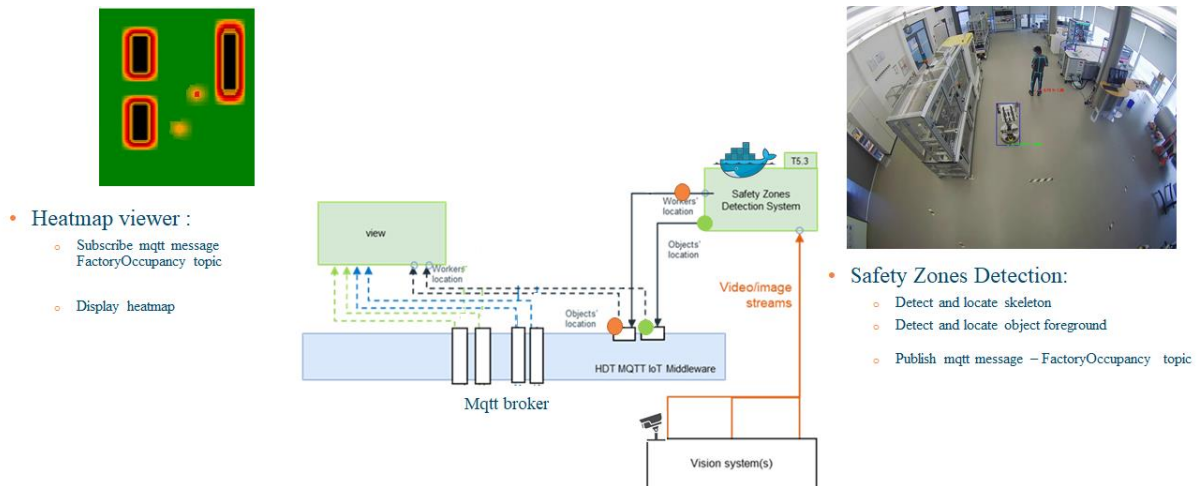


Figure 23: Safety Zones detection detects and locates human and object and publishes the position on mqtt broker present in the HDT system

4 Conclusion

This deliverable aims to present the global picture in which the Safety Zones Detection System is installed, namely the goal of this development, the main interaction forecasted with the HDT system and the other HDT components.

Starting from the section 2, different Computer Vision approaches are described. A particular attention is brought to the methodologies implemented in the context of the STAR project and the motivation of the choice we decide to follow for this final prototype.

Section 3 offers the description of our demonstrator implemented thanks to the technologies described in section 2.

This final version of the demonstrator presents a state of the art on the few shot learning approach we forecasted for the STAR development. The prototype is also completed with the implementation of the methodology we decide to follow to reach the object classification tasks.

Currently, the final demonstrator runs on servers with GPUs (NVIDIA GeForce RTX 3080 Ti, model optimisation based on TensorRT) in a local network, but a possible evolution would be some optimisation to run on a low consumption processor in order to install the solution as close as possible to the sensors (cameras) and robust the security and preserve privacy and avoid sharing personal data locally in the infrastructure.

5 Bibliography

- [1] Y. Xu, J. Dong, B. Zhang and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Transactions on Intelligence Technology*, pp. 43-60, 2016.
- [2] K. Toyama, J. Krumm, B. Brumitt and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Proceedings of the seventh IEEE international conference on computer vision*, vol. 1, pp. 255-261, 1999.
- [3] C. Ridder, O. Munkelt and H. Kirchner, "Adaptive background estimation and foreground detection using kalman-filtering," *Proceedings of International Conference on recent Advances in Mechatronics*, pp. 193-199, 1995.
- [4] S. Indupalli, M. A. Ali and B. Boufama, "A novel clustering-based method for adaptive background segmentation," *The 3rd Canadian Conference on Computer and Robot Vision*, 2006.
- [5] K. Kim, T. H. Chalidabhongse, D. Harwood and L. Davis, "Real-time foreground--background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172-195, 2005.
- [6] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *arXiv preprint arXiv:1302.1539*, 1997.
- [7] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," vol. 2, pp. 246-252, 1999.
- [8] A. Elgammal, D. Harwood and L. Davis, "Non-parametric model for background subtraction," *European conference on computer vision*, pp. 751-767, 2000.
- [9] P.-L. a. B. G.-A. a. B. R. St-Charles, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, 2014.
- [10] L. a. Z. F. Jiao, F. Liu, S. Yang, L. Li, Z. Feng and R. Qu, "A Survey of Deep Learning-Based Object Detection," *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 7, 2019.
- [11] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580-587, 2014.
- [12] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision*, pp. 1440-1448, 2015.
- [13] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, pp. 1137-1149, 6.

- [14] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [15] J. Redmon, S. Divvala, G. R. and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.
- [16] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018.
- [17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, "SSD: Single shot multibox detector," *European Conference on Computer Vision (ECCV)*, pp. 21-37, 2016.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.
- [19] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.
- [20] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra and others, "Matching Networks for One Shot Learning.," *Advances in neural information processing systems*, 2016.
- [21] J. Snell, K. Swersky and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, 2017.
- [22] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
- [23] G. Dahia and M. Segundo , *Meta Learning for Few-Shot One-class Classification.*, 2020.
- [24] P. Perera and V. Patel, "Learning Deep Features for One-Class Classification," *IEEE Transactions on Image Processing*, 2018.
- [25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627-1645, 2009.
- [26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [27] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [28] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[29] Z. Ge, A. Bewley, L. Ott, F. Ramos and B. Upcroft, *Simple online and realtime tracking.*, 2016.