

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: INTRASOFT International



Funded by the Horizon 2020
Framework Programme of the
European Union

DELIVERABLE

D5.5 – Visual Scene Analysis for Safety Zones Detection-Initial version

Dissemination level	PU -Public
Type of Document	Report
Contractual date of delivery	31/03/2022
Deliverable Leader	THALES SIX GTS GRANCE
Status - version, date	Final – v1.0, 31/03/2022
WP / Task responsible	WP5/ THALES
Keywords:	Visual Scene Analysis, Safety zones, localization

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 first 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 provides also justifications to support the approach followed for the demonstrator development. In the final part a roadmap of the improvements that have to be implemented and the missing AI assets to develop will be presented.

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 in order 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:	THALES: Andreina Chietera, Jean-Emmanuel Haugeard
Reviewers:	IBER : Mihail Fontul UNIPI : Spyros Theodoropoulos
Approved by:	Charalampos Ipektsidis, John Soldatos (INTRA)

Document History			
Version	Date	Contributor(s)	Description
V0.1	03/03/2022	Thales	TOC
V0.2	15/03/2022	Thales	First Draft of the deliverable available
V0.3	25/03/2022	Thales	Second Draft of the deliverable available
V0.4	28/03/2022	UPRC	First review
V0.5	29/03/2022	IBER	Second review
V0.6	30/03/2022	Thales	Integration of comments
V1.0	31/03/2022	INTRA	QA and creation of the final submitted version

Table of Contents

EXECUTIVE SUMMARY	2
TABLE OF CONTENTS.....	4
TABLE OF FIGURES.....	5
LIST OF TABLES.....	6
DEFINITIONS, ACRONYMS AND ABBREVIATIONS	7
1 INTRODUCTION.....	8
2 MODERN COMPUTER VISION APPROACHES FOR SITUATION AWARENESS	11
2.1 DETECTION OF OBJECTS OF INTEREST IN VIDEO STREAMS	11
2.1.1 <i>Moving Object Detection Using Background Modelling.....</i>	<i>11</i>
2.1.2 <i>Objects detection and classification Using CNNs</i>	<i>14</i>
2.1.3 <i>Human Detection Based on Deep Learning</i>	<i>16</i>
2.2 OBJECT GEOLOCATION USING 3D CALIBRATION	17
3 FIRST IMPLEMENTATION OF THE STAR DEMONSTRATOR	21
3.1 INPUT DFKI PLATFORM:	21
3.2 OUTPUT:	23
3.2.1 <i>Heatmaps for the path planning implementation</i>	<i>23</i>
4 CONCLUSION.....	24
5 BIBLIOGRAPHY	25

Table of Figures

FIGURE 1 : AN OVERVIEW OF THE INTEGRATION FROM THE POINT OF VIEW OF THE ARCHITECTURE. 9

FIGURE 2 : VISUAL SCENE ANALYSIS COMPONENTS.....10

FIGURE 3: BASIC STEPS FOR BACKGROUND SUBTRACTION ALGORITHMS – THALES LABORATORY EXAMPLE12

FIGURE 4 : RESULTS OF DIFFERENT BACKGROUND SUBTRACTION ON CDNET DATASET13

FIGURE 5 : SUBTRACTION EVALUATION14

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.15

FIGURE 7: EXAMPLE OF HOG FEATURE ON THE STAR PROJECT IMAGE.....16

FIGURE 8: DEFORMABLE PART MODEL : MODEL FOR THE PERSON CATEGORY.....16

FIGURE 9: 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 10 : SKELETON DETECTION AND 3D LOCATION WITH KEY POINTS (FEET, HIP, SHOULDER) PROJECTION 18

FIGURE 11 : SKELETON DETECTION AND 3D POSITIONS ALONG THE WHITE LINES GROUNDTRUTH19

FIGURE 12 : COMPARISONS BETWEEN 3D POSITIONS ESTIMATED BY FEET PROJECTION, HIP PROJECTION, SHOULDER PROJECTION AND THE GROUNDTRUTH20

FIGURE 13 : DFKI PLATFORM – 3 WORKSTATIONS, ROBOTS AND 2 CAMERAS.....21

FIGURE 14 : THALES CALIBRATION TOOLS BASED ON VANISHING POINTS.21

FIGURE 15 : HUMAN 3D POSITION BASED ON POSE ESTIMATION MODEL AND CALIBRATION22

FIGURE 16 : BACKGROUND SUBTRACTION - MOVING OBJECT DETECTION BASED ON SUBSENSE22

FIGURE 17 : BACKGROUND SUBTRACTION - MOVING OBJECT DETECTION BASED ON GAUSSIAN MODEL. SOME FALSE ALARMS - SHINY REFLECTIONS IN LEFT WORKSTATION23

FIGURE 18 : HEATMAP EXAMPLE23

List of Tables

No table of figures entries found.

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AGV	Autonomous Ground Vehicles
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
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 required 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 analyze the nearby environment, in order to move safely and avoid obstacles. Unfortunately, this technology does not provide dynamical global view of the work floor. Thus, the cohabitation between humans and robots remains unoptimized 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 analyze the scene using the global point of view of the camera network already deployed in the factory.

Video analytics allows to exploit automatically the video streams in real time to detect anomalies and to raise immediately an alarm. To this end, the algorithms detect and track elements of interest (such as people, robot and new object occupying the scene) over the time, and alert the robots of the presence of any obstacles in the surrounding area. Where 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 a 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 and workers' positions in the work floor on the IoT Middleware in order to update the HDT picture.

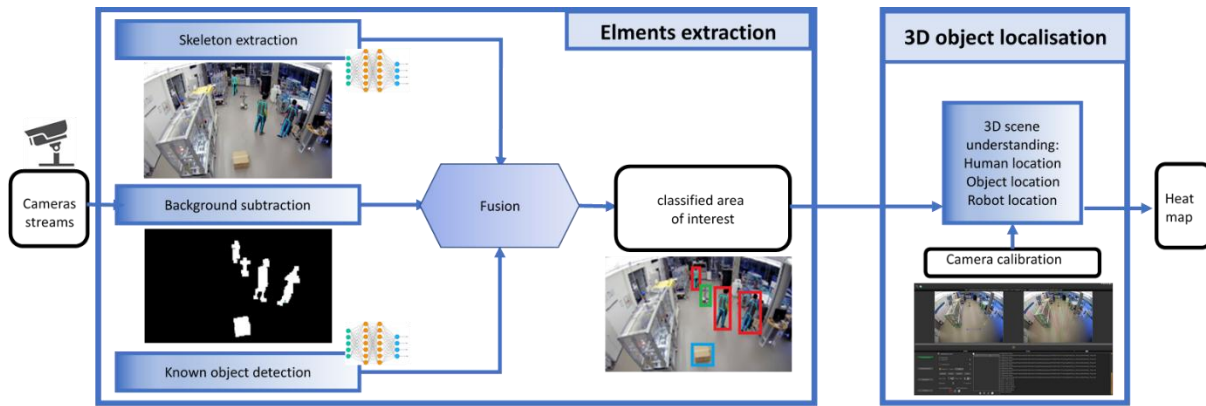


Figure 2 : Visual Scene Analysis Components

Particularly the elements extractor engine merges two deep learning algorithms, either for the skeleton reconstruction in order 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 in order to infer moving elements in the scene under observation.

The 3D object localization takes as input the results of the elements extractor in order to localize them using the Euclidian reference system. To archive this task a calibration software is developed in order 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 realize the first 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 optimization 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 recognize the postures and motion of workers, 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 modeling
 - Dynamic object detection via a convolutional neural network (CNN)
 - Skeleton extraction by human pose detection CNN
- 3D Object geo-localization 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 modeling of the observed scene (“background”) as a first step, then on the analysis of the differences between each image and the estimated background (cf 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 (color/texture) from the background model in order to detect the difference between the current image and the background model.

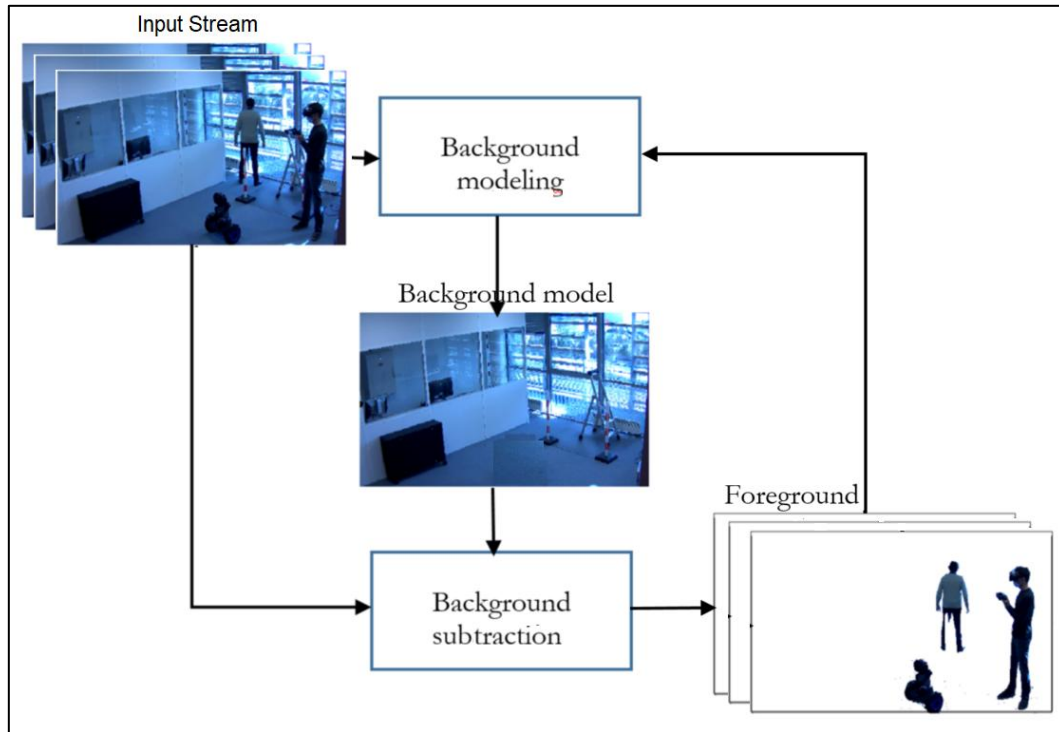


Figure 3: Basic Steps for Background Subtraction Algorithms – Thales laboratory example

A detailed survey of various background modeling 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 modeling of the background

Stochastic methods (GMM approaches) are more commonly used in the video applications.

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 analyze 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 ChageDetection 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

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

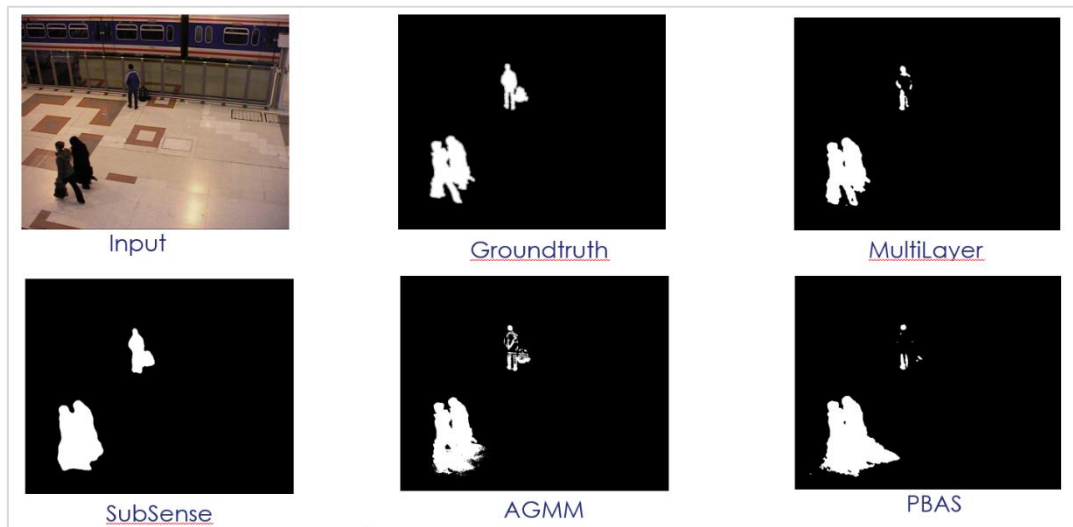


Figure 4 : Results of different background subtraction on CDNet dataset

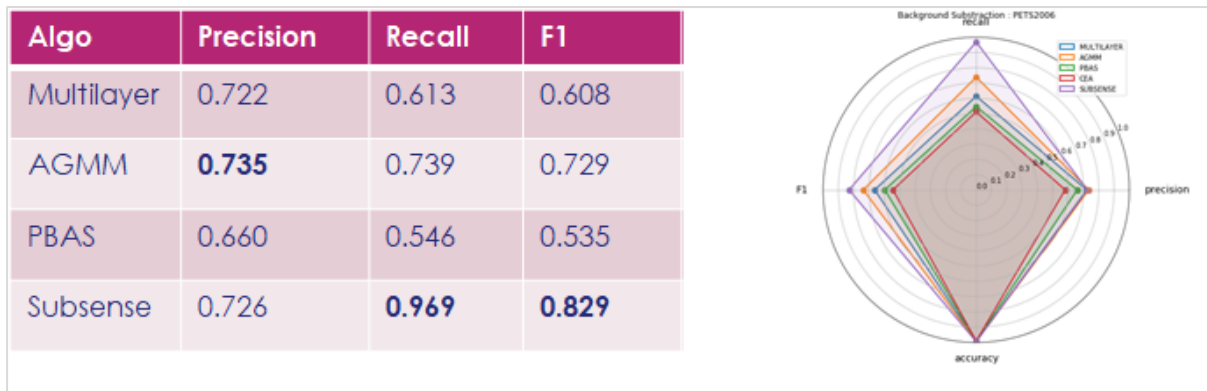


Figure 5 : Subtraction Evaluation

The best results (Figure 5) are obtained by approaches based Gaussian model (GMM) and the SuBSENSE approach. The SuBSENSE [9] method combines color 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]).

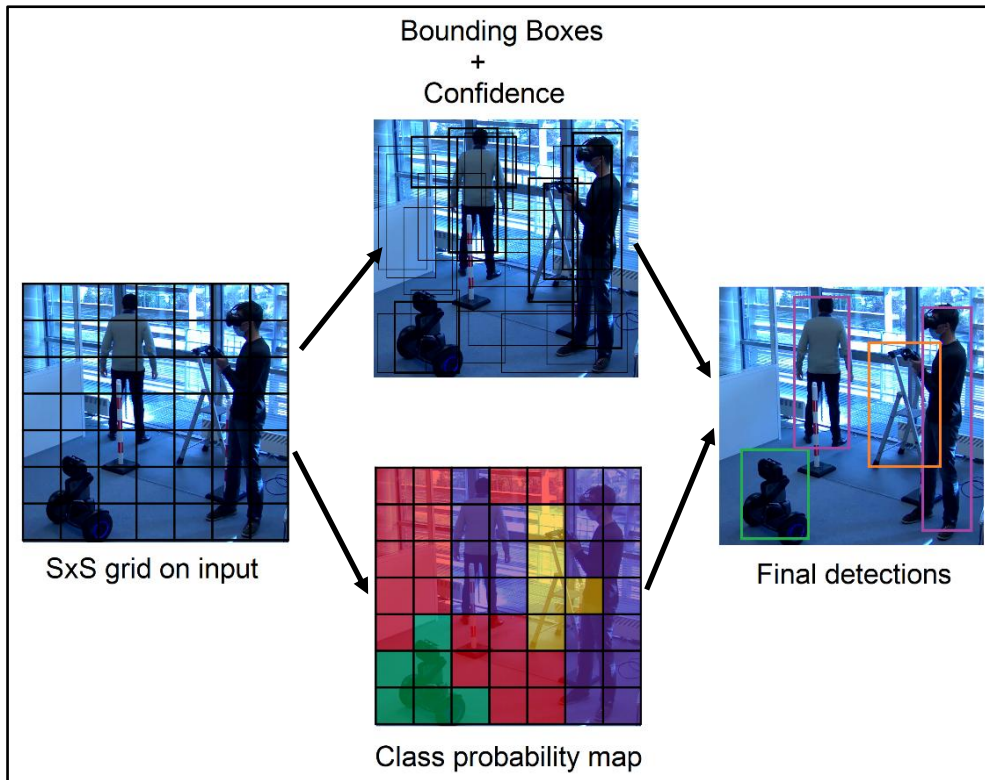


Figure 6: Object Classification using YOLO Algorithm in the context of Robot-Human Cohabitation. In this result, the final detection is robot, human and stool.

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, the vast majority 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.... In order 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 labeled data of sufficient size. With a large dataset, the optimization 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 labeled 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 and to tackle the scarcity of examples available for a given problem.

2.1.3 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 configuration of the shape (feature templates - Deformable Part Model (DPM) [20] Figure 8). These techniques relied on the use of local attributes (descriptors) such as Histogram of Oriented Gradients (HOG [21] Figure 7), 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 7: Example of HOG feature on the STAR project image

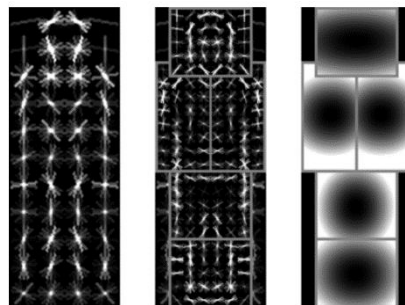


Figure 8: Deformable Part Model : model for the person category.

A technique for people detection, called OpenPose, was recently proposed by [22] and takes into account 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 9). 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.

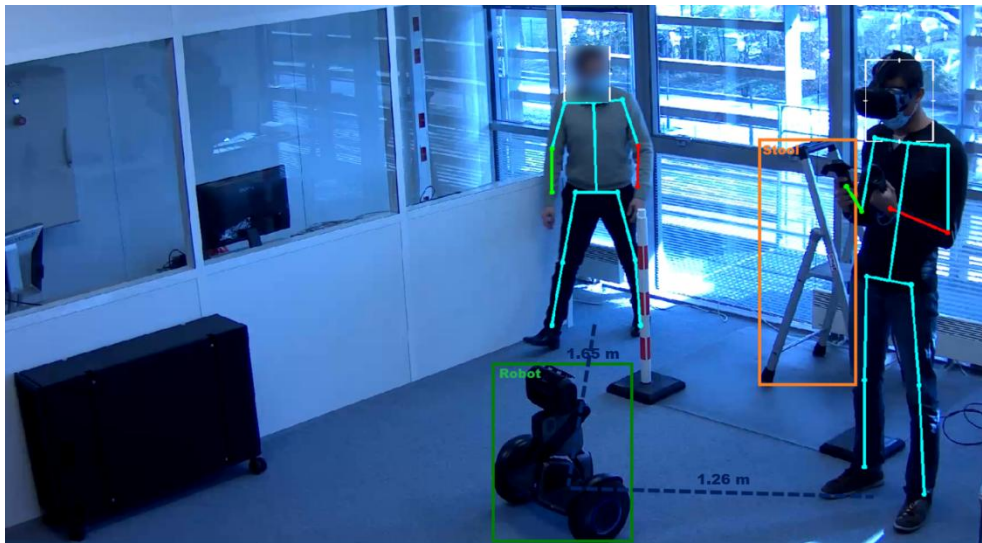


Figure 9: the Figure shows an example of the results obtained using different detectors: 1) moving object detection, (GMM subtraction) 2) object detector to identify the stool and the robot (Yolo), 3) human detector (OpenPose)

2.1.3.1 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 makes 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 [23] 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 localization 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 a large number of 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, in order 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 10) illustrates the principle and the results compared to the calculated vertical of each individual.



Figure 10 : 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 11 and Figure 12), we measured the drift between the estimated position with our approach and the groundtruth (the reality). As illustrated in the figure 11, people walk in different positions known (measured by a rangefinder). For example, in figure 11 (white lines), the person moves along axis $y=5\text{ m}$ and then $x=2.5\text{ m}$. 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 11 : Skeleton detection and 3D positions along the white lines groundTruth

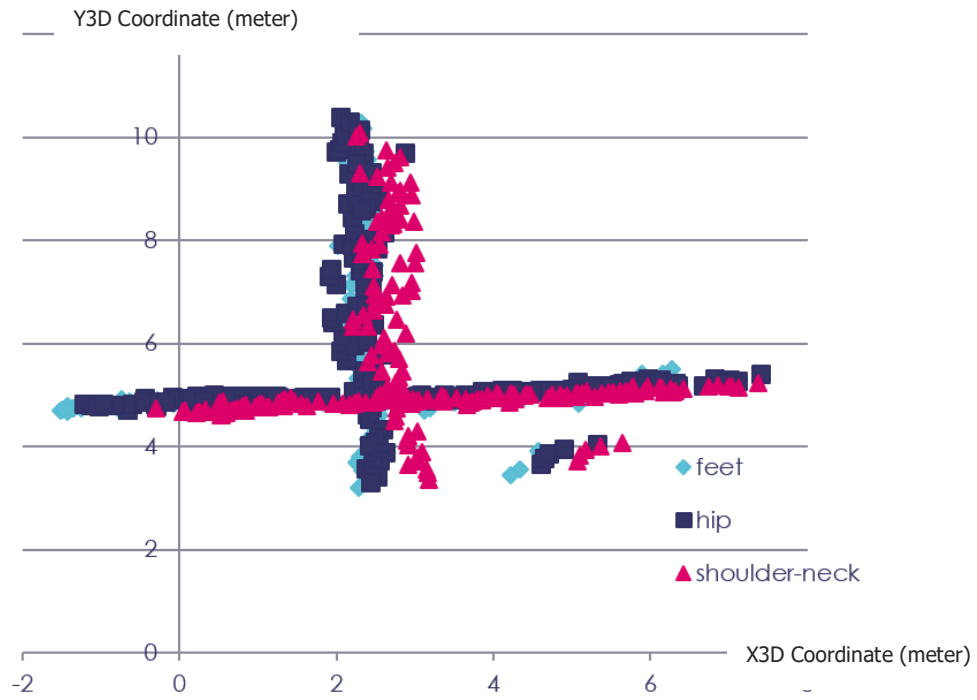


Figure 12 : comparisons between 3D positions estimated by feet projection, hip projection, shoulder projection and the groundTruth

3 First 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 13).



Figure 13 : DFKI platform – 3 workstations, robots and 2 cameras

Each camera has been calibrated using the Thales tool (Figure 14) developed in the framework of STAR, which allows fast calibration with just a few clicks (horizontal lines, vertical lines, yardstick).

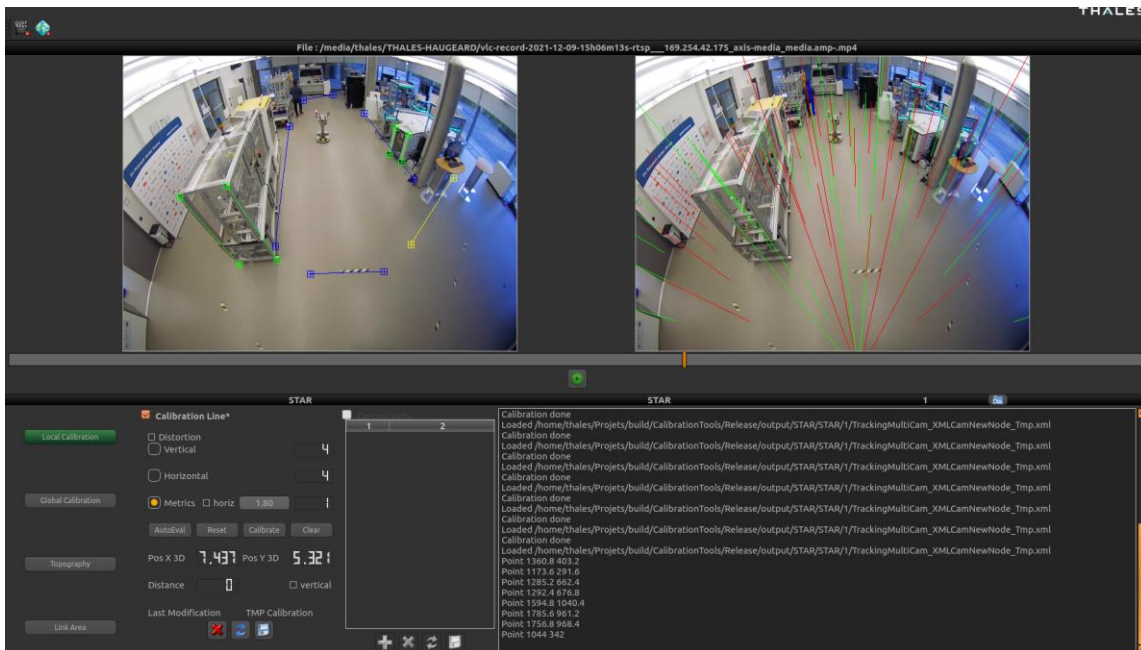


Figure 14 : Thales Calibration tools based on vanishing points.

The first results (Figure 15) 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.

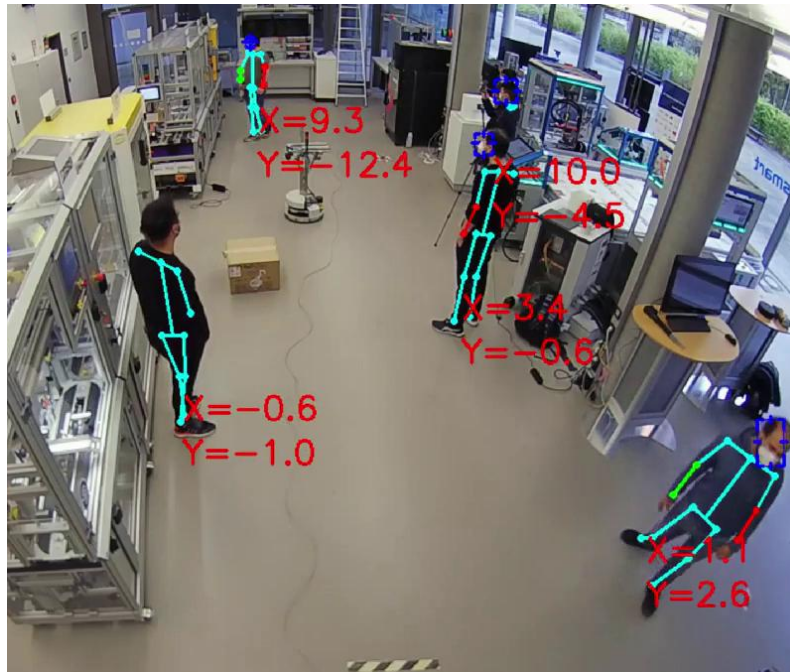


Figure 15 : Human 3D position based on pose estimation model and calibration

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 16 : Background subtraction - moving object detection based on SuBSENSE



Figure 17 : Background subtraction - moving object detection based on Gaussian model. Some false alarms - shiny reflections in left workstation

3.2 Output:

3.2.1 Heatmaps for the path planning implementation

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. We will represent the results of the implementation presented in the previous sections using anonymized heatmaps. This description represents the global environment, including human and object location as occupancy cells. The heatmaps will be used as input by Reinforcement Learning that will perform AMRs Fleet control (cf. Figure 18 and STAR D5.7: Reinforcement Learning Techniques for AMRs).

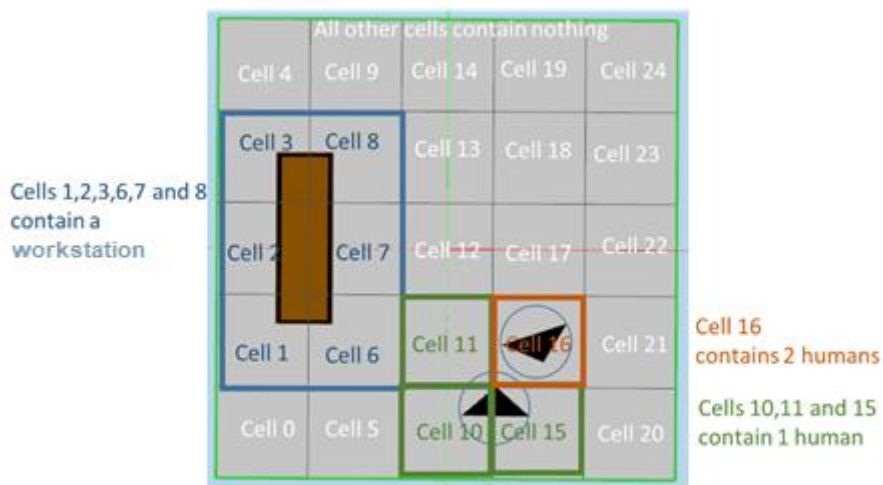


Figure 18 : Heatmap example

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 first prototype.

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

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

The final demonstrator will also be optimized to run on a low consumption processor in order to 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] 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.
- [21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [22] 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.
- [23] 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.