Ref. Ares(2024)346330 - 16/01/2024

This project is funded by the European Union
G.A No 101016906

Project Number: 101016906
Start Date of Project: 2021/01/01
Duration: 48 months

Type of document D6.1 – V1.0

Human body prediction and human-robot motion planning

Dissemination level	PUBLIC
Submission Date	2024-01-16
Work Package	WP6
Task	T6.1, T6.2
Type	Report
Version	1.0
Author	UPC
Approved by	UNIROMA1, UNIROMA3

DISCLAIMER:
The sole responsibility for the content of this deliverable rests with the authors. It does not necessarily reflect the opinion of the European Union. Neither the Directorate-General for Communications Networks, Content and Technology, Artificial Intelligence and Digital Industry nor the European Commission are responsible for any use that may be made of the information contained therein.

This project is funded by the European Union G.A No 101016906

Executive Summary

In this deliverable, we present several Human-Robot Interaction (HRI) techniques that can be applied to agriculture tasks.

In the first part, we explain two techniques for predicting human-body pose motion in the future, which can be used for safety and cooperative tasks. These techniques have been tested with datasets collected in the lab and at the Corsira Agricultural Cooperative Society (Aprilia, Italy) field. The first technique predicts future 2D human motion trajectory, based on the previous trajectory, the possible destinations and the fixed or moving objects of the environment. This technique has been tested with standard datasets and with real-life experiments in the lab. The second technique predicts 3D human skeleton poses (position and orientation in 3D), based on the previous poses' trajectories, the task context and the desired human intention. This technique has been tested with a handover task in the lab, and with data collected of harvesting operations in the Corsira field.

In the second part, we explain a new general model for doing human-robot cooperative tasks and a cooperative human-robot task for harvesting bunches in a vineyard field using a multiagent planner. The general model is denominated PIA (Perception-Intention-Action) model, which includes perception and intention modules, a situation awareness module (which is used to understand current progress of the cooperative task and makes prediction of the future movements of the human/robot), and a decision-making module (which includes negotiation, role assignment and multiagent planning). The output of this process is the cooperative plan agreed between the robot and the human. This general model has been applied to several human-robot cooperative tasks, and specifically it has been applied to a handover task between a robot and a human. It has also been applied to bunches harvesting task in a vineyard field, where a human and a robot negotiate a multiagent plan. This system includes mapping of the vineyard bunches including quality detection, shared trajectory planning and negotiation to distribute where the human and the robot have to harvest. In this system we have considered that the human will harvest the bunches where the quality is not so good, and the robot will harvest the good bunches, although other criteria may be used.

Most of the methods presented in these tasks have been published in top international conferences and journals. Two of the methods explained in this deliverable, the PIA model applied to handover task and the multiagent planning are going to be sent to a top international conference and later on to journals.

Table of Content

1 Introduction 6

2 Human-Body prediction 6

2.1 Prediction of the 2D human motion trajectory 7

2.1.1 SoFGAN architecture 8

2.1.2 SoFGAN results 8

2.2 3D prediction of human skeleton body poses 10

2.2.1 CANOPIES dataset of human-body prediction: Handover and harvest datasets 10

2.2.2 System architecture 11

2.2.3 Experimental results of human-body prediction and intentionality in handover task 12

2.2.4 Experimental results of human-body prediction and intentionality in harvesting task 14

3 Human-Robot Motion Planning in Agreement 15

3.1 The PIA (Perception-Intention-Action) model 15

3.2 The PIA model applied to a handover task 16

3.3 Multiagent planning for human-robot teams 17

3.4 Mapping and collaborative planning in agreement 21

4 Conclusions 27

5 References 28

Abbreviations and Acronyms

CANOPIES	A Collaborative Paradigm for Human Workers and Multi-Robot Teams in Precision Agriculture Systems
UPC	Universitat Politècnica de Catalunya
IRI	Institut de Robòtica i Informàtica Industrial
RGB	Red Green Blue
RGB-D	Red Green Blue – Depth
CNN	Convolutional Neural Network
HRI	Human-Robot Interaction
2D	Two dimensions
3D	Three dimensions
GAN	Generative Adversarial Network
CVAE	Conditional Variational Autoencoder
SFM	Social Force Model
SoFGAN	Social Force Generative Adversarial Network
mADE	minimum Average Displacement Error
mFDE	minimum Final Displacement Error
ROS	Robot Operating System
LiDAR	Light Detection and Ranging
RViz	ROS 3D Robot Visualizer
Mediapipe	It is an open-source framework for building pipelines to perform computer vision inference over arbitrary sensory data such as video or audio.
REE	Robot End Effector
PIA	Perception-Intention-Action model
MCTS	Monte Carlo Tree Search planner
SLAM	Simultaneous Localization and Mapping
SRS	Social Reward Sources

RRT	Rapidly Exploring Random Tree
RTK	Real Time Kinematic GPS
April tag	April tags is a fiducial system that can be handled by SMODE for tracking and calibration.
HRT	Human Robot Team

This project is funded by the European Union G.A No 101016906

1 Introduction

In this deliverable we present the results of Task 6.1 “Human-Body Prediction” and Task 6.2 “Human-Robot Motion Planning in Agreement”.

The objective of Task 6.1 is to predict the human-body poses in the future to be used for safety and cooperative applications in farmer tasks. In this way, a robot will be able to know in advance where will be the future positions of different human body parts (head, body, arms, legs), when the human is working close to a robot.

The objective of Task 6.2 is to build a shared plan between a robot and a human for doing cooperative tasks. In this way, the robot and the human will be able to propose a plan, negotiate it, and finally agree with the final cooperative plan.

In the following sections, we briefly explain the methods and the tests that were done with datasets collected in the lab and others collected in the Corsira Agricultural Cooperative Society (Aprilia, Italy) vineyard field.

The organization of the document is as follows. In Section 2, we explain the methods that we have developed for human-body prediction, the first one for predicting future 2D human motion trajectory, and the second for predicting 3D human skeleton poses. In Section 3, we explain a new model for cooperative tasks between humans and robot, denominated PIA (Perception-Intention-Action) model, and an application of it for a handover task, and finally we explain a multiagent planning, between a robot and a human, applied to a harvesting task.

2 Human-Body prediction

The objective of this section is to describe different techniques that allow the robot to predict the future movements of the humans that is close to it, for safety and collaborative farmer tasks. In safety, we have to know where the human body will be in the future, to avoid collisions. In a collaborative harvesting task, we need to know where the human body will be in the future, to anticipate robot body motion and also to anticipate what the robot has to do depending on the human actions.

To do so, we have proposed two models:

- A first one for predicting future 2D human motion trajectory
- A second one for predicting future 3D human skeleton poses

The first method makes a prediction of the future human motion trajectory in a 2D space, that means we predict where the human will be in the next seconds.

In the second method, we predict where all the 3D human skeleton poses will be in the near future, that means the motion of the head, arms, legs, and body. Moreover, in this case we can also predict

the underlying intention of the human in this predicted sequence to let the robot understand, for example, if the human will cooperate with it or not.

2.1 Prediction of the 2D human motion trajectory

The objective of this module is to predict the 2D human motion trajectory with an observation horizon of next 4.8 seconds. The method that we have developed [1] combines a Generative Adversarial Network (GAN) and Social Force Model (SFM). The method is denominated SoFGAN and it is explained in detail in the article [1].

The SoFGAN uses a CVAE (Conditional Variational Autoencoder) to predict possible destinations, includes the environment features (fixed and moving objects) through the SFM (Social Force Model) and generates the future trajectory using a GAN (Generative Adversarial Network), by taking into account the past trajectory, the possible destinations and the environmental features. Fig. 1 shows these elements.

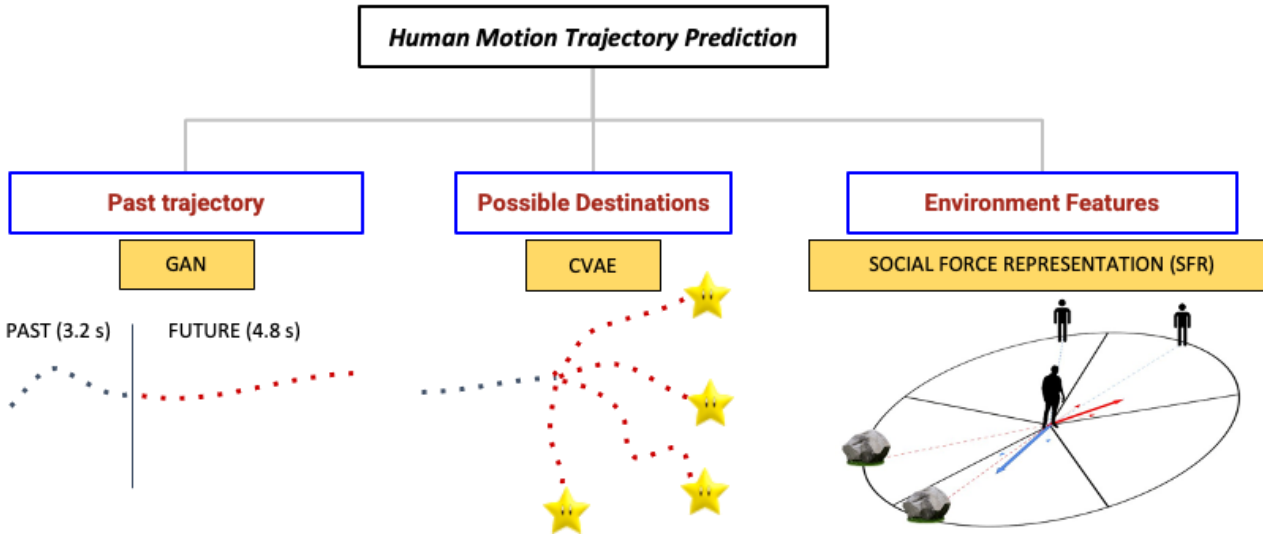


Fig.1 The human motion prediction requires: i) possible destinations (generated by the CVAE module); ii) past and future trajectory (generated by the GAN module); and iii) environment features (generated by the SFM module)

2.1.1 SoFGAN architecture

The SoFGAN architecture can be seen in Fig.2.

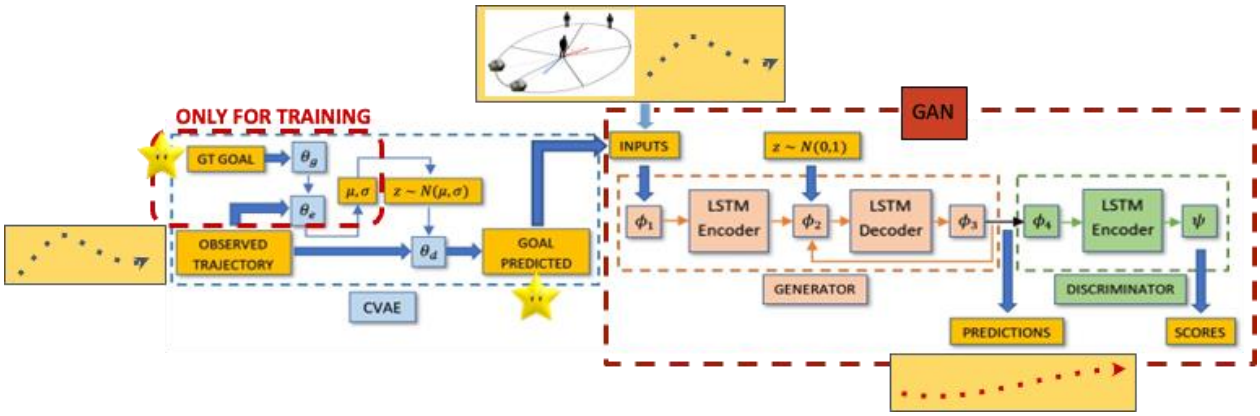


Fig.2 Social Force GAN with CVAE architecture. The GAN generator provides predictions using the predicted goals given by the CVAE module.

The model architecture has three main modules: SFM (Social Force Model), CVAE (Conditional Variational Autoencoder) architecture and GAN (Generative Adversarial Network). SFM is used to obtain environmental representation, as a part of the inputs for the Social Force GAN model. CVAE is used to reduce the main source of prediction errors, the last position prediction, and for this reason is used to estimate the goal trajectories. GAN is used to make human motion trajectory prediction using the estimated goal obtained in the CVAE module and constrained forces of the environment obtained by the SFM module. The output is a distribution function which has to be sampled, and any of the samples is a trajectory solution.

2.1.2 SoFGAN results

This model has been tested with standard datasets (see [1] for more details) in human motion prediction, and the results are quite good, predicting better than most of the techniques, although it is not better than Y-NET or NSP (see [1] for more details). However, the predictor works very well with low computational cost and without the need to use GPU's.

The tests with the standard datasets can be seen in Table 1, where the values in the table are the mADE and mFDE metrics (see article [8] for more details).

Model	ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
SGAN [4]	0.81/1.52	0.72/1.61	0.34/0.69	0.42/0.84	0.60/1.26	0.58/1.18
SoPhie [19]	0.70/1.43	0.76/1.67	0.30/0.63	0.38/0.78	0.54/1.24	0.54/1.15

Next [9]	0.73/1.65	0.30/0.59	0.38/0.81	0.31/0.68	0.60/1.27	0.46/1.00
CVM-20 [21]	0.96/2.09	0.29/0.54	0.52/1.03	0.34/0.70	0.61/1.26	0.54/1.12
PECNet [12]	0.54/0.87	0.18/0.24	0.22/0.39	0.17/0.30	0.35/0.60	0.29/0.48
Trajectron ++ [20]	0.57/1.05	0.16/0.26	0.22/0.41	0.16/0.31	0.28/0.56	0.28/0.52
AgentFormer [22]	0.45/0.75	0.14/0.22	0.18/0.30	0.14/0.24	0.25/0.45	0.23/0.39
Y-NET [11]	0.28/0.33	0.10/0.14	0.17/0.27	0.13/0.22	0.24/0.41	0.18/0.27
NSP [23]	**0.25/0.24**	**0.09/0.13**	**0.16/0.27**	**0.12/0.20**	**0.21/0.38**	**0.17/0.24**
SoFGAN	0.44/0.68	0.13/0.18	0.20/0.35	0.17/0.31	0.25/0.46	0.24/0.40
fSoFGAN	0.48/0.79	0.15/0.22	0.22/0.42	0.20/0.38	0.28/0.53	0.27/0.47

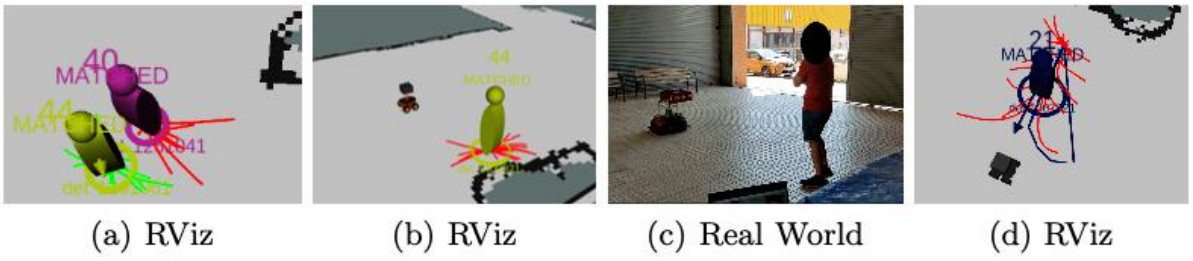
Table 1 mADE/mFDE results. Comparison between SoFGAN and other models for 20 samples

The SOFGAN has also been applied to real-life experiments using the robot Helena in the IRI`s outdoor lab spaces (these experiments are in lab conditions).

Through a ROS implementation in the Helena robot, real outdoor experiments have been performed with people. Helena is a transporter robot with a RS- LiDAR-16 and a Pioneer P3-DX. Although the robot Helena motion prediction is not computed, robot Helena is taken into account as an agent to compute the environmental forces. Similar experiments have been done with the IVO robot [2].

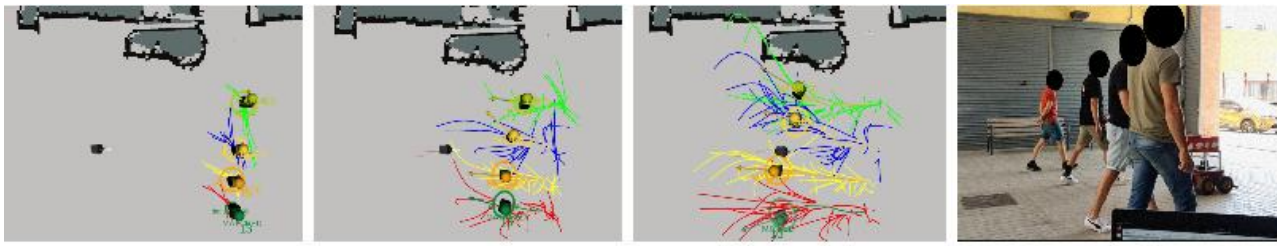
The first experiment is to evaluate the predictions when the robot is not moving. In Fig. 3, from left to right, in the first image the interaction between 2 people in a conversation is shown. The model does not sample trajectories between them. In the second image, the effect of the obstacle behind the person causes most of predictions to appear in different directions. In the last image, the person walks towards Helena. In this case, the predictions take into account the robot and try to avoid collisions.

The second experiment, shown in Fig. 4, evaluates the interaction between 4 people and the interaction between the group and the robot. From left to right, in the first image, the group is not moving, and the predictions are short and avoid collisions between them. In the second image, the group moves, and the predictions try to avoid collisions with the robot. In the third image, the people change their velocity directions due to the robot and the blue and green predictions close to obstacles try to avoid a collision.



(a) RViz (b) RViz (c) Real World (d) RViz

Fig.3 Predictions when the robot is not moving. The predictions are the red and green lines in the ground. The visualization is obtained using RViz



(a) RViz: t (b) RViz: t+1 (c) RViz: t + 2 (d) Real World

Fig.4 Predictions during an encounter between 4 people and the robot. The predictions are the colored lines in the ground. (a) shows the result at time t, (b) at time t+1 and (c) at time t+2.

2.2 3D prediction of human skeleton body poses

The objective of this module is to predict the future 3D human skeleton poses, based on the previous human skeleton poses and in contextual information. This model has been applied to predict 3D human skeleton motion and human intention in a handover and a harvesting task. The method that is explained in this section has been submitted for publication in a scientific journal [3], however different parts of the model have been already published in international conferences [4], [5] and [6].

The whole pipeline assumes that the human is doing the harvesting operations, and the robot head camera is recording the human.

Then, the Mediapipe pose estimator is used in order to obtain the 2D human joint positions in the image frame. After that, the depth map from the RGB-D camera is used to reconstruct the 3D skeleton of the human.

These skeletons are stored in an array and passed to the model, an attention based neural network and graph convolutional networks.

2.2.1 CANOPIES dataset of human-body prediction: Handover and harvest datasets

Two different datasets have been used to train the neural networks: a handover and a harvest dataset. The handover dataset was recorded in lab conditions and the harvest dataset was recorded in the experimental field of the CANOPIES project, that is the “Cooperativa Agricola Corsira” located in the municipality of Aprilia, Lazio, Italy.

The handover dataset consists of sequences recorded in an indoors environment where several humans delivered an object to an anthropomorphic robot. The human was the deliverer and the robot the receiver. During the data collection, the human had to repeat the action several times featuring four different intentions: “collaborative” (where the human acted in order to successfully finish the handover), “gesture” (where the human did some movement not related with the handover task, such as checking the smartphone, waving the hand to greet the robot, …), “neutral” (where the human stops at some point during the handover and doesn’t continue with the task) and “adversarial” (where the human acts against the handover task, such as moving the hand away from the robot).

The other dataset was collected in the real experimental field in Aprilia, with the project’s robot. During the recordings, the robot camera was looking at the human harvesting bunches. The human collects a bunch, and then he/she can drop the bunch in a box placed in front of the robot (which is considered “collaborative” intention) or do some gesture not related with the task, like drinking water or checking the smartphone (“gesture” intention) or move away from the robot (“non-collaborative” intention).

Also, the second dataset was expanded with sequences where the robot used the prediction model to identify the intention of the human. Once the human intention was identified, the robot responded by saying a sentence to the human. When the human wanted to drop the bunch, the robot said “Please, drop the bunch in the box”. When the human made some gesture, the robot said, “Do you want to continue with the harvesting?”. Finally, when the human moved away from the robot, the robot said, “You are leaving the harvesting area”.

The idea behind this expansion is to also capture the reactions that robot feedback can have over humans, since we assumed that the act of speaking with humans would change the predicted motion. This assumption makes more natural the HRI operations.

2.2.2 System architecture

The model (see Fig.5) first processes the input sequence with a discrete cosine transform function to work in the frequency domain (see Laplaza et. al 2024 [3]). Then, by using one dimensional convolutional neural network the causality in the sequence is modeled. In this model, the “task context” can be used to add information that can help to improve the predictions. For instance, a task context in the harvesting operation would be the position of the box in front of the robot, since the human will probably move the arm towards the box. This additional input is treated the same way as the input array. The model allows to add as many tasks’ context as desired.

Then, the embedding is used to feed an attention module, which provides the attention scoring to decide which part of the input is more relevant. The attention scoring from all the possible inputs is leveraged in the fusion module. The fusion module also allows to select a specific intention for the predicted skeleton sequence.

Finally, the attention embedding is fed into a graph convolutional network, where each skeleton joint is considered as a node in a graph and the limbs are the edges of the node. This graph shows the human skeleton topology and provides the output predicted skeletons, which are converted again to cartesian space with an inverse discrete cosine transform. An auxiliary node is used to predict the intention displayed during the predicted sequence with a final one-dimensional convolution network.

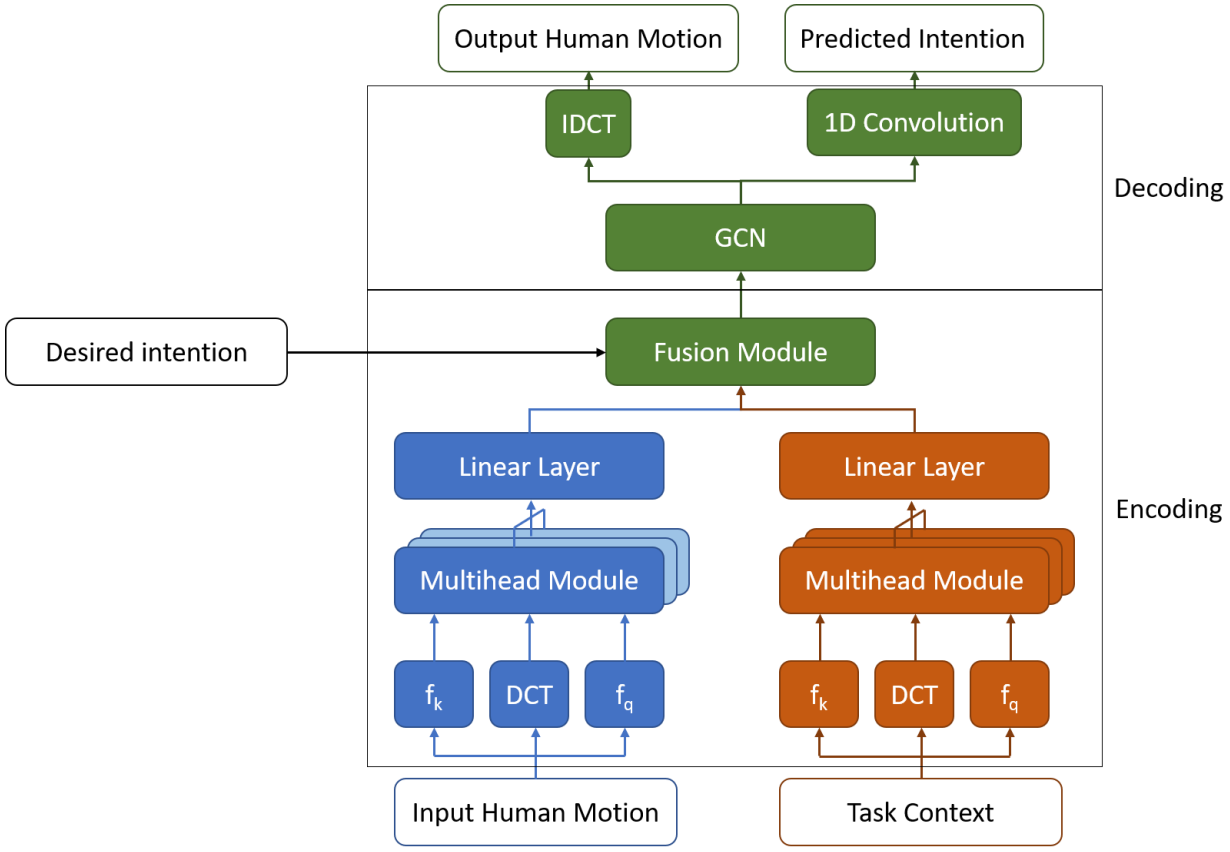


Fig. 5. Model's architecture. The inputs for the model can be seen on the bottom of the figure, while the outputs are on the top. The optional intention selector input can be seen on the left side of the figure. The blue modules process the past human skeletons. The orange modules process the contextual information related to the task. The green modules are in charge of fusing all the embedded information and generating the future output sequence and detecting the intention of the prediction (Source: Laplaza et al. 2024 [3])

2.2.3 Experimental results of human-body prediction and intentionality in handover task

The handover dataset was used to train and evaluate the neural network model. Two contextual pieces of information from the task were included in the study: the position of the Robot End Effector (REE) since the human goal was to place an object in that part of the robot and the position of some obstacles that were placed in the scene. Also, the "desired intention" input is implemented, which allows the model to be trained knowing beforehand the intention of the human. The desired intention can be obtained by an intention prediction module and used in the execution time to improve the prediction of the 3D human motion skeleton. The Table 2 shows the ablation study of all

these features. Note that the intention accuracy is not provided in the first row, since the first row corresponds to the original model, which did not provide this output. The meaning of the table columns is the following: the first column is the square mean error of all the joint poses; the second column is the % of results which mean square error is lower than 0.35m; the third column is the % of results which mean square error is lower than 0.4m; the fourth column is the mean square error of the right hand; and the sixth column is the % of the intention accuracy.

Model	Body L_2 (m)	<0.35 m (%)	<0.4 m (%)	Right Hand L_2 (m)	Intention Accuracy (%)
○ **REE** ○ **Obstacles** ○ **Intention**	0.403	34.13	37.14	0.188	-
• **REE** ○ **Obstacles** ○ **Intention**	0.378	**41.65**	45.78	0.174	56.45
○ **REE** • **Obstacles** ○ **Intention**	0.444	41.31	44.87	0.187	62.02
○ **REE** ○ **Obstacles** • **Intention**	0.453	30.81	36.15	0.173	86.29
• **REE** • **Obstacles** ○ **Intention**	0.381	41.60	**47.38**	0.172	74.16
• **REE** ○ **Obstacles** • **Intention**	0.375	34.84	38.86	0.162	85.44
○ **REE** • **Obstacles** • **Intention**	0.387	40.22	43.73	0.170	88.69
• **REE** • **Obstacles** • **Intention**	**0.355**	32.15	35.73	**0.151**	**88.90**

Table 2. Table containing the results from the handover model. The first column is the square mean error of all the joint poses; the second column is the % of results which mean square error is lower than 0.35m; the third column is the % of results which mean square error is lower than 0.4m; the fourth column is the mean square error of the right hand; and the sixth column is the % of the intention accuracy.

The conclusion from these results shows that the model which uses all of the three features is the one obtaining the best results in almost every metric except for the second and third metrics (see Table 2), being the REE model and the REE + obstacles models the ones obtaining the best results for

these two metrics, respectively. A possible hypothesis for this is that adding the intention embedding into the model causes some kind of noise for the model making it less accurate in these metrics.

2.2.4 Experimental results of human-body prediction and intentionality in harvesting task

The harvest dataset was also used to train and evaluate the neural network model. The only contextual information from the task was the bunch/cluster position. The "desired intention" input is also considered, allowing the model to train knowing beforehand the intention of the human. The following table (Table 3) shows the ablation study of all these features. We examine the model's accuracy during the initial stages of the harvesting process when a person approaches the clusters and severs the peduncle. This early phase is particularly crucial because it requires more manual dexterity than subsequent phases of the operation. Note that while the "intention" feature from the previous table was related to the percentage of predicted frames that were classified correctly, in this case we evaluate the percentage of times that the "desired intention" module produces sequences that follow this desired intention. Thus, when the 'intention' feature is not used, the "intention accuracy" is not provided.

Model	Body L_2 (m)	<0.35 m (%)	<0.4 m (%)	Right Hand L_2 (m)	Intention Accuracy (%)
○ **Bunch** ○ **Intention** ○ **Phase**	0.399	82	96	0.365	-
• **Bunch** ○ **Intention** ○ **Phase**	0.383	58	66	0.268	-
○ **Bunch** • **Intention** ○ **Phase**	0.358	72	72	0.363	90.2
○ **Bunch** ○ **Intention** • **Phase**	0.209	**100**	**100**	0.142	-
• **Bunch** ○ **Intention** • **Phase**	**0.208**	**100**	**100**	0.143	-
○ **Bunch** • **Intention** • **Phase**	0.21	**100**	**100**	**0.141**	**83.33**
• **Bunch** • **Intention** • **Phase**	0.208	**100**	**100**	0.145	**83.33**

Table 3. Table containing the results of the harvesting task (the metrics are explained in Table 2)

Similarly with the previous case, we can see that by adding the intention feature, the model internal noise increases and thus the accuracy is compromised. Nonetheless, we believe that being able to predict and generate different sequences with different intentions displayed by the human is valuable enough to compensate for the decrease in accuracy.

3 Human-Robot Motion Planning in Agreement

The objective of this section is to describe a new general model for human-robot interaction tasks, and to describe an application of a multiagent planning for human robot teams applied to harvesting.

In the first part, we will briefly describe a new model, denominated PIA model which takes into account the human intention and the perception of how the cooperative task is being developed, and how using this information, the robot can anticipate what the human is going to do and what the robot has to do. We also include a real-life experiment of a handover task between a robot and a human, where the main steps of the PIA model are illustrated. Moreover, the handover task is a typical task that appears in many agricultural operations, for example when the human delivers a cut vineyard grape in the robotic box.

In the second part, we will explain a multiagent method for human- robot team planning, applied to harvest a field of table grapes in a collaborative manner. First, we explain the multiagent plan method that we are using, and then we explain the use case of collaborative planning for harvesting between a human and a robot.

3.1 The PIA (Perception-Intention-Action) model

A new human-robot model for cooperative tasks, denominated PIA model [7], has been designed. The model is shown in Fig.7, where there are two agents who cooperate in a specific task. Each one of the agents (for example a robot and a human), perceive the environment and the other agents and bystanders and at the same time gets the implicit or explicit intention of the other agent. These two inputs are used to understand the current situation of the progress of the cooperative task and make a prediction of how the task is going to continue in the future. This last module is denominated Situation awareness of the agent. The outputs of this module, the output of the Intention module and the knowledge about the task are the inputs for the Decision-Making module, which includes the negotiation among the agents, the role task assignment, and the cooperative plan. If the agents agree on the plan, then the plan is executed.

The PIA model has been used in human-robot collaborative object transportation, and the article [8] analyzes the differences of using implicit intention versus explicit intention in this task and the article [9] analyzes that using a force predictor the effectiveness of the task can be improved. For the communication of the human intention, we have worked on body gesture recognition [4] and hand

gesture recognition [10]. We have also worked in communication-aware multi-robot coordination with intermittent connectivity [11] to be applied to this model.

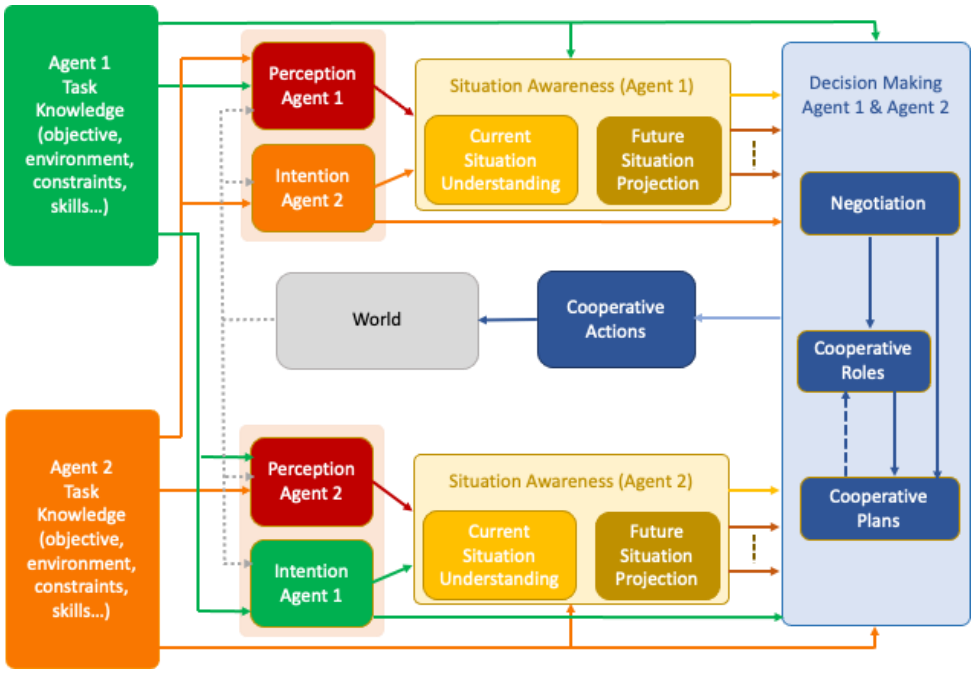


Fig.7 The PIA (Perception-Intention-Action) model

3.2 The PIA model applied to a handover task

We have also applied the PIA model in a typical handover operation, for example the robot gives a tool or box to the farmer. In this case, the robot tries to understand the human intention of receiving the tool or box, predicts the human hand and 3D skeleton motion in the near future, plans the robot arm motion to approach its gripper to the human hand meeting pose, and execute the actions. In all these robot actions, the robot is detecting if the human intention is going to change, in order to change its actions. This model has been tested in a real-life human-robot handover task using the results of Task 6.1.

Fig.8 shows a real-life experiment using the PIA model, where the robot has to deliver a box to a human agent. We reiterate that this represents an illustrative example which can be relevant in agricultural settings by exchanging different objects. The sequence of steps composing the task is the following (see Fig.8):

- The robot arrives to the entrance of a building and looks for a human to deliver a box. It detects a human (Fig.8-1)
- The robot perceives the human intention using the Intention module and detects that the human is willing to receive the box. Then makes a prediction where the human will be after some seconds (white avatar in Fig.8-2). In case that the robot detects that the human is not willing then the robot plan will change.

- Using the prediction of the human motion, the robot makes a plan to anticipate the delivery of the box and starts moving its arm and gripper to approach to the human delivery point in the future (Fig.8-3).
- The human moves to the expected delivery position and picks up the box (Fig.8-4). Then the robot releases the box and moves back the arm. In case that the robot detects that the human intention changes, it will do another action, for example stop moving the arm and asking the human what he wants to do.

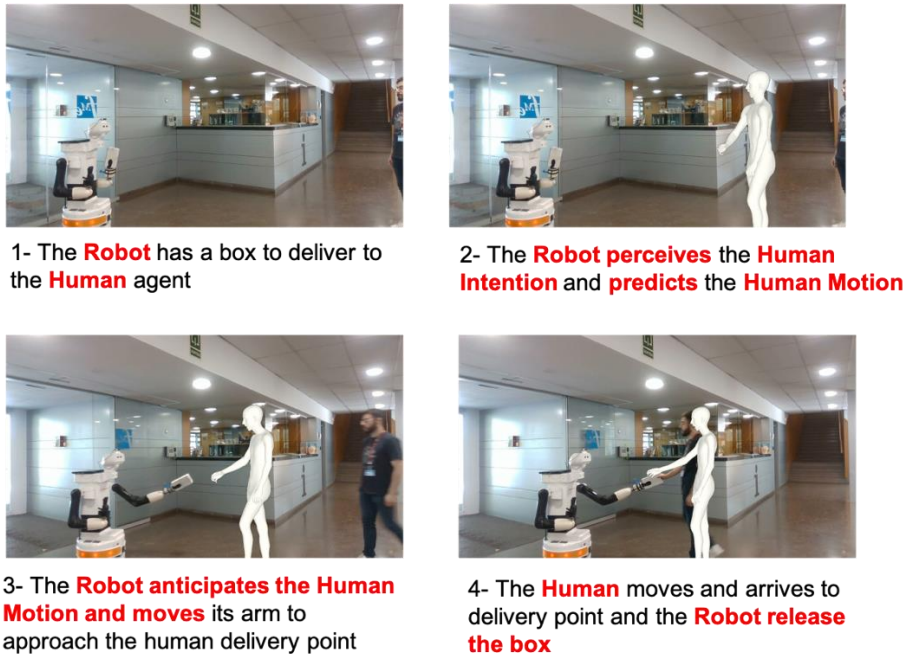


Fig.8 A handover task using the PIA model

We are planning to apply this PIA model in safety and harvesting operations in the field.

3.3 Multiagent planning for human-robot teams

In this use case we are using a multi-agent harvest plan method developed by our team, Dalmasso et. al [12], that uses the multi-agent Monte Carlo Tree Search (MCTS) planner, which generates team motion plans over the Social Reward Sources (SRS) world representation (see Dalmasso et al. [13] for more details).

A social reward source $\psi \in \Psi$ is a generative model that defines a reward function along all the search space. Reward sources act as building blocks of a subjective world representation, which models influences conditioning robot and human behavior. We define it as subjective because each source of reward may provide a different feedback to different agents, even under the same conditions. For example, rewards may vary due to risk, capabilities, effort, personal preferences, or eagerness to do a task or to follow someone's instructions. For instance, in the collaborative harvest planning example

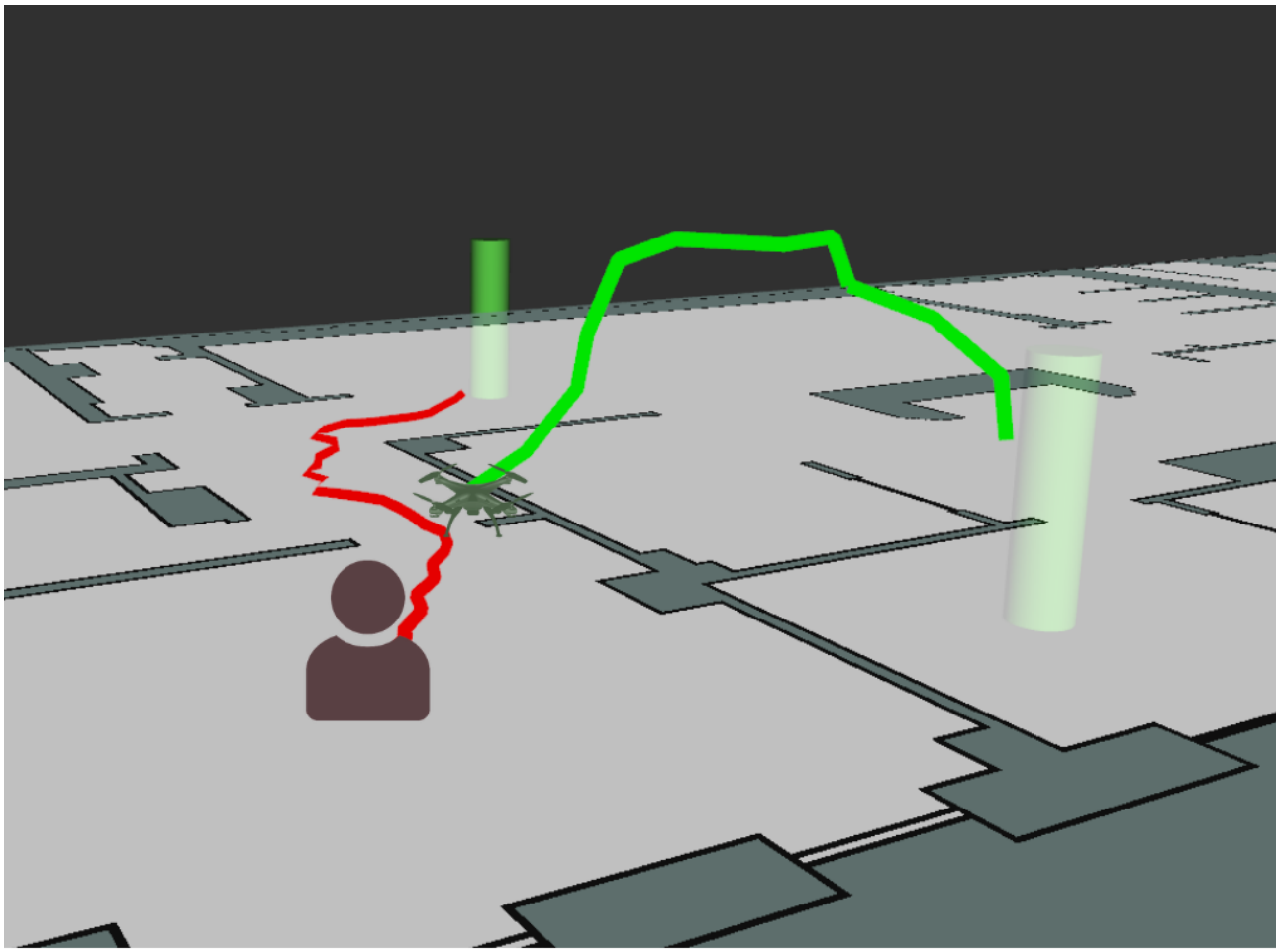
This project is funded by the European Union G.A No 101016906

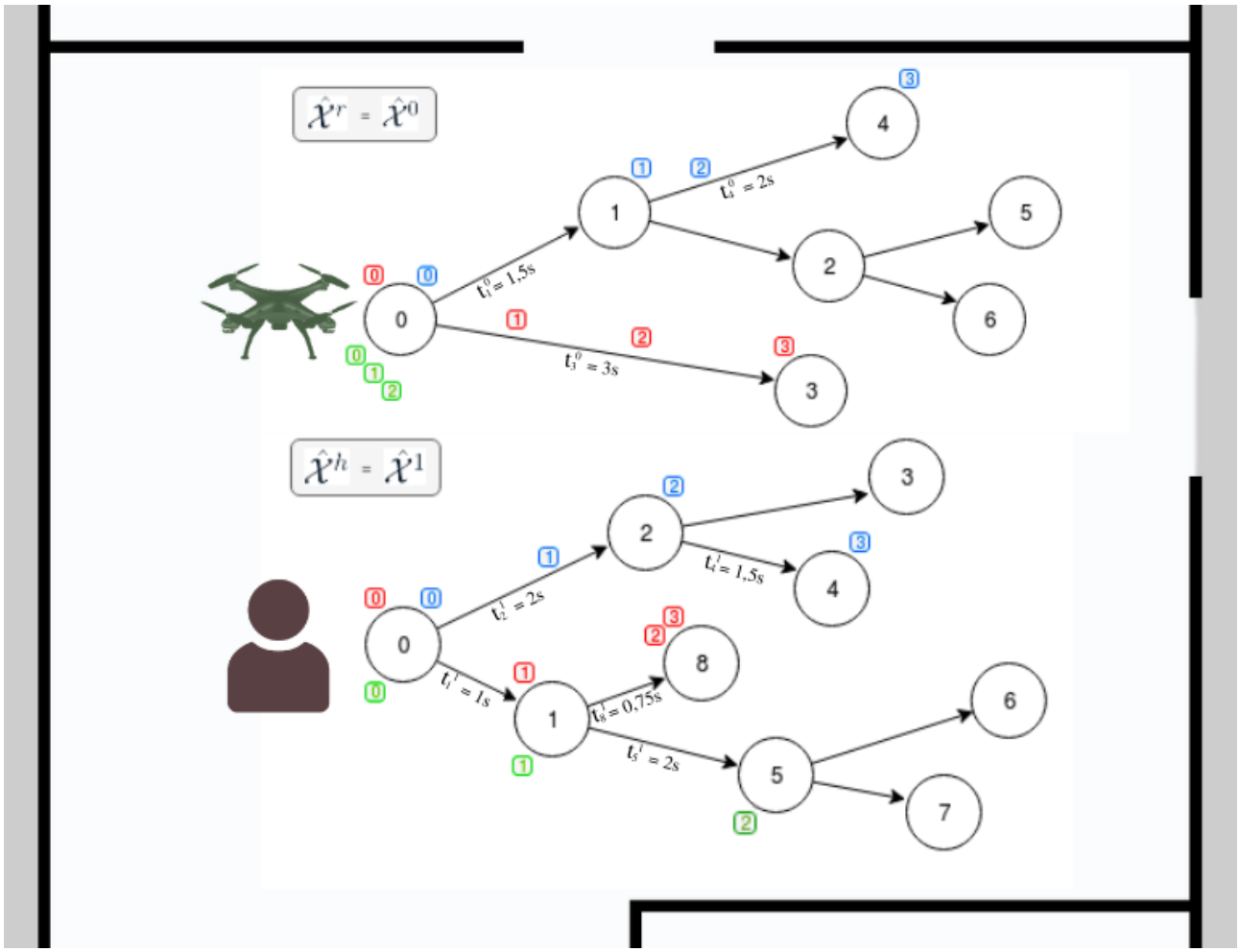
presented in section 3.4.2, each grape is connected to a source of reward modeling the positive outcome of its harvesting, which may be different due to quality or affordance.

We consider that all sources ψ have a boundary β. This boundary is a closed surface in the search space and the reward output (reward/time) of the source is defined by two functions corresponding to internal and external reward ratio functions.

This planner generates plans for teams of multiple agents using MCTS over a search space of possible individual paths, which is generated by expanding individual rapid random trees (RRTs) for each agent. States are identified as a list of RRT nodes encoding movement action towards a goal state and a list of reach times, rt, encoding the remaining action time (Fig.9) (see [13] for more details).

This project is funded by the European Union
G.A No 101016906




$\hat{\mathcal{X}}^r = \hat{\mathcal{X}}^0$
$t_1^0 = 1.5s$
$t_2^0 = 2s$
$t_3^0 = 3s$
$\hat{\mathcal{X}}^h = \hat{\mathcal{X}}^1$
$t_1^1 = 1s$
$t_2^1 = 2s$
$t_4^1 = 1.5s$
$t_5^1 = 2s$
$t_8^1 = 0.75s$

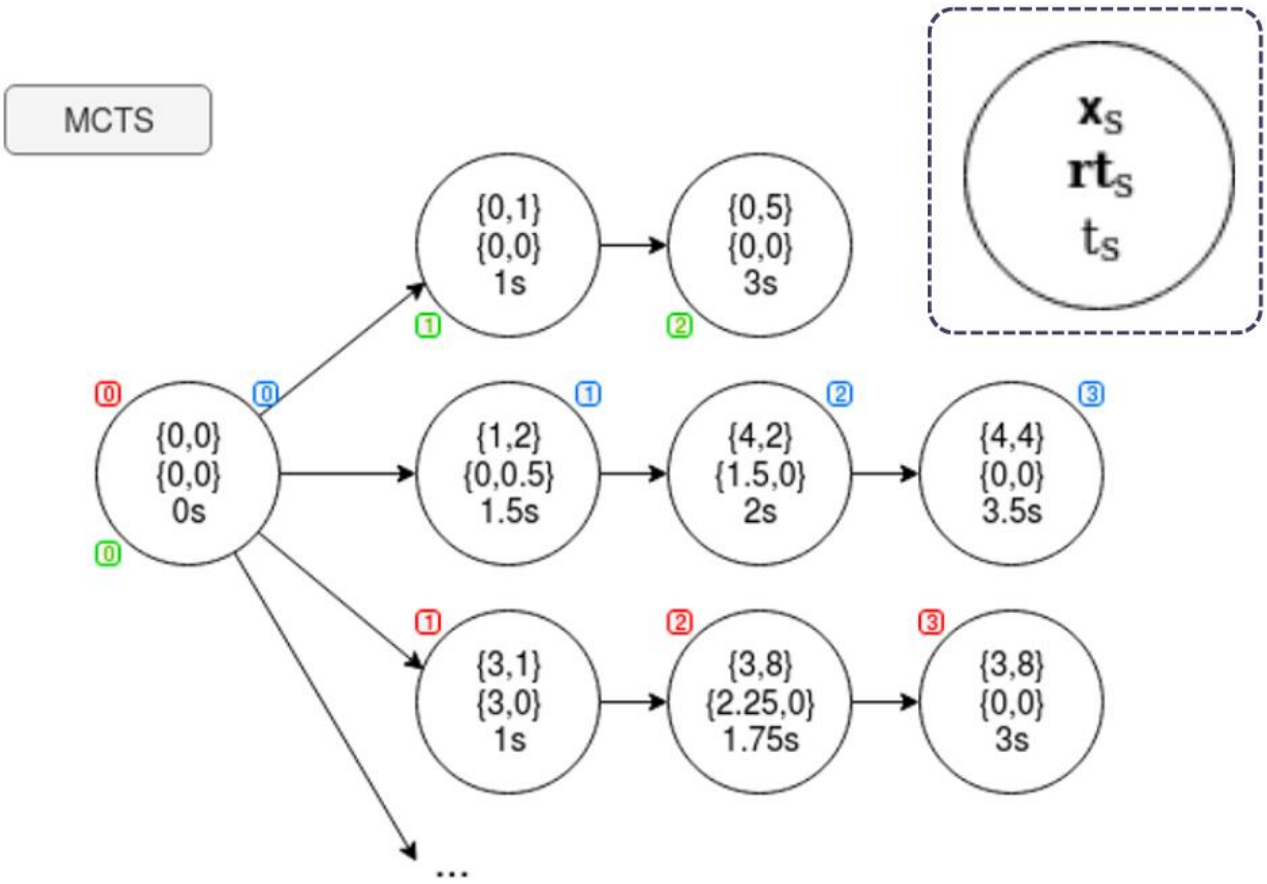


Fig.9. From top to bottom: a) Sample execution of the MCTS planner for to agents moving in 3 and 2 dimensions, respectively. The green cylinders represent sources of reward of cylindric boundaries collectable by both agents. b) Scheme of the search space of each actor. Each node of the RRTs is identified by its id and stores the action time of being reached from its parent. c) Sample scheme of a possible MCTS expansion over the search space presented in image b. (Source: Dalmasso et al [16]).

3.4 Mapping and collaborative planning in agreement

In this section, we are going to explain a use case of a collaborative planning between a robot and a human to harvest some vineyard trees. We have used the multi-agent Monte Carlo Tree Search (MCTS) planner described before.

The objective is to build harvest plans for human-robot teams in the field. The method has several steps:

- First create a map of the grape bunches, including the quality of the bunches, using only the CANOPIES robot. To do this operation, the robot moves through one lane of the vineyard field, acquires the images of the bunches, and obtains their localization and the quality of the grape bunches.

- Second, apply the multiagent plan to generate the paths that the robot and the human have to harvest. The robot will only harvest the bunches that have good quality, are bigger than a specific size, etc., and the human will harvest the rest of the bunches. The multiagent planner creates both path plans, for the robot and the human, and presents the plan to the human.
- Third, negotiate the plan between the robot and the human, where the human can accept the robot plan or modify it. Once the negotiation is finished, the multiagent planner shows the path that the robot and the human must follow. This negotiation is done using a computer interface.
- Fourth, apply the plan.

In the next sections we will describe the different steps.

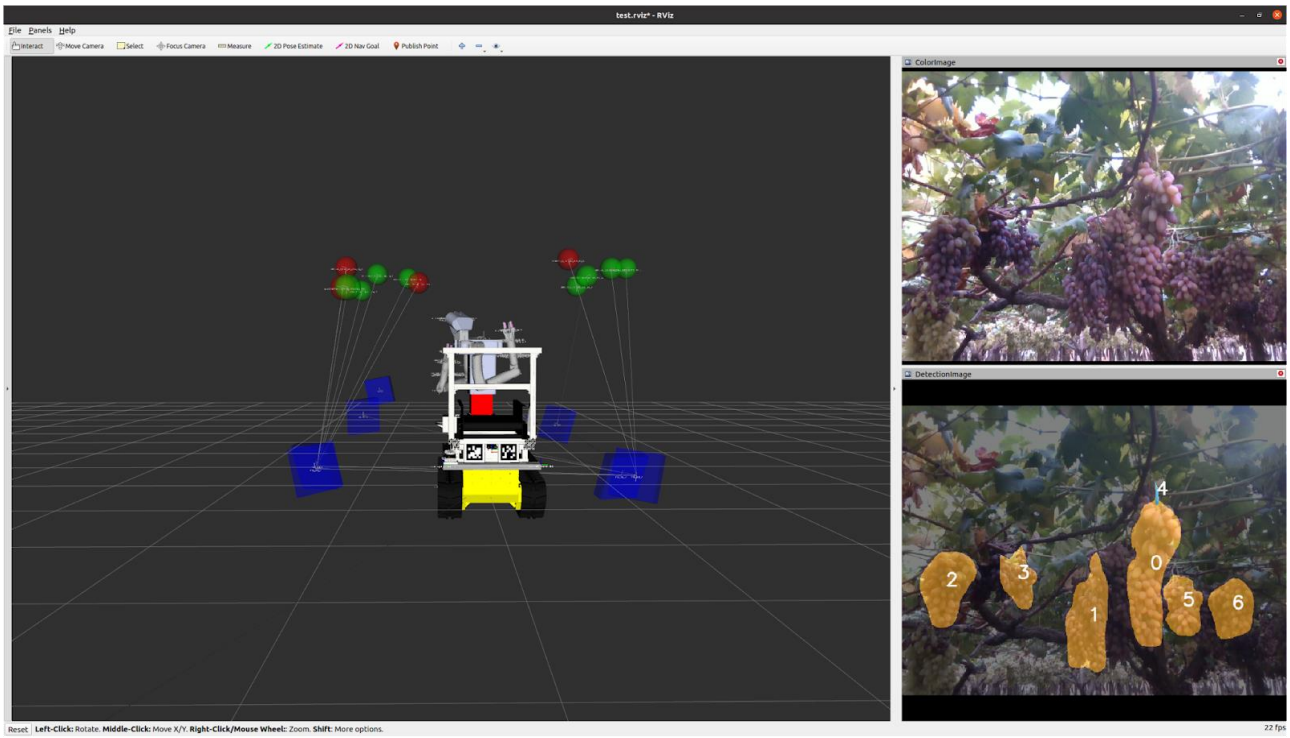
3.4.1 Mapping

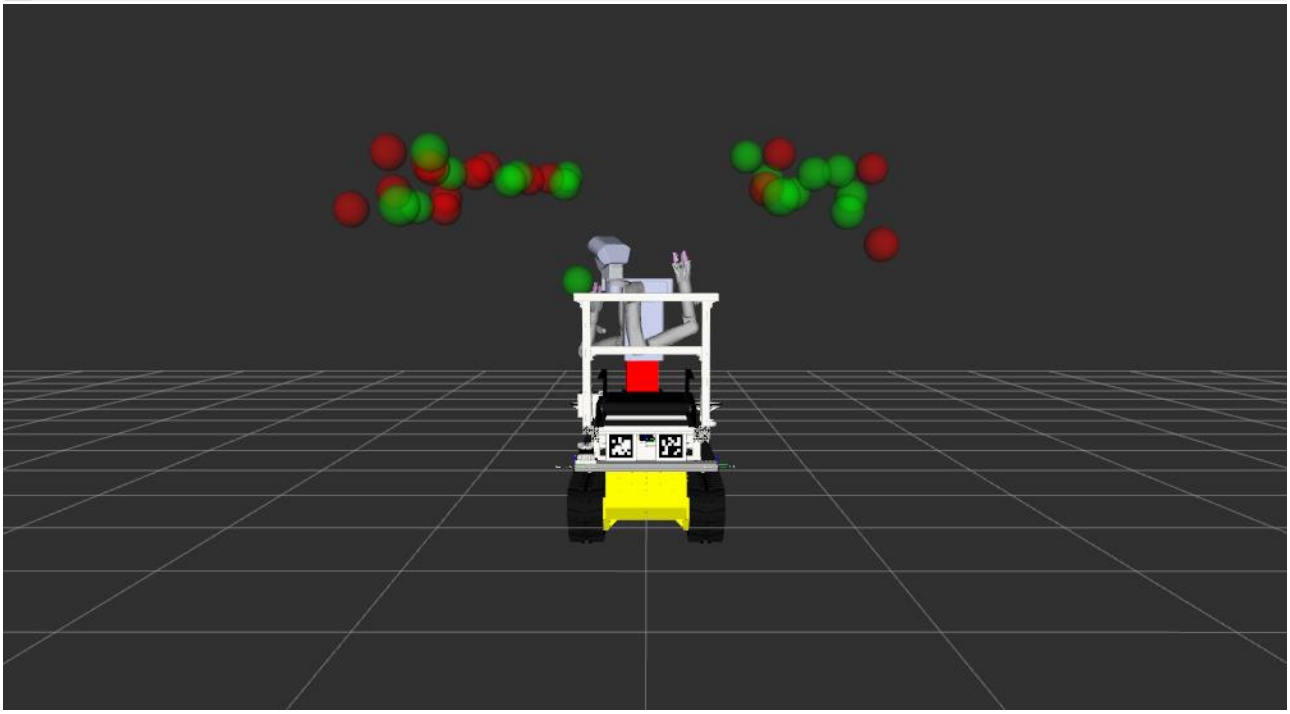
We aim to build a grape map to store grape positions and quality features and enable the robots to generate and reason over harvest plans. To do so, we make use our previously presented grape bunch detection system [14] and assign to the grapes a quality estimation value. The quality information is generated by the Quality Estimation Module (QEM), a ROS node described in Deliverable 4.2 that is able to estimate four quality parameters upon user request.

The position of the bunches is computed with respect to the robot which is located using RTK and SLAM techniques and also April tag (visual marks). More details on the SLAM architecture can be found in Deliverable 9.1.

To construct the map, we navigate a designated part of the field and repeatedly launch both systems to detect bunches and estimate their quality. Quality data is merged with the new one obtained from the new observations. Fig.10 shows the robot localized in the field and the images of the detected bunches. Fig.10 left (top image), shows the robot and the localization of the bunches. Fig.10 right (top image), shows the RGB image and the RGB image with the detected bunches. Fig.10 (bottom images) shows the robot and the localization of some of the bunches.

This project is funded by the European Union
G.A No 101016906





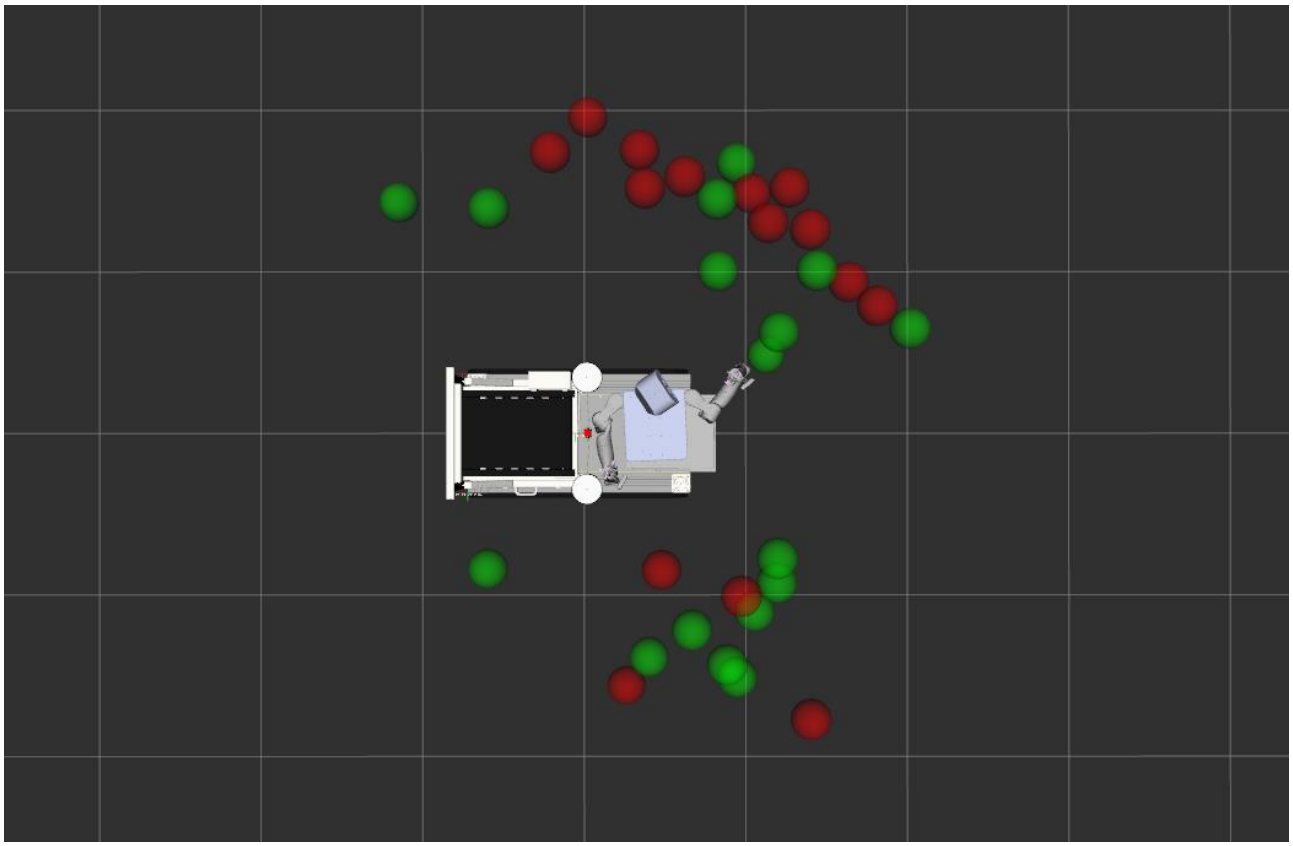


Fig.10. Mapping over Rosbag files recorded at the field. From left to right: Top image: Detection and mapping in the local map. Bottom images: Detection and mapping using the canopies project localization. In all images, grape color identifies estimated quality as a random binary variable q = {0,1}.

3.4.2 Negotiation

Consider a human-robot team (HRT) of agents Λ = {h,r} confronted with the Collaborative task T of recollecting the quality adequate grape bunches from a delimited area of a field. We may consider the task as a set of subtasks γ, each linked to the action of checking and, if necessary, recollecting each bunch.

$$T = \{\gamma_1, \dots, \gamma_n\}$$

Each grape in the field may be recollected by either the human or the robot. Still, the robot only considers himself capable of making the harvest decision for grapes with high quality value, information which is stored in the grape map. Given that restriction, each sub task has a target set $\tau \subseteq \Lambda$ describing which agents are capable of fulfilling it. Generally, we will use the following notation:

$$T^{\Lambda} = \{\gamma_1^{\tau_1}, \dots, \gamma_n^{\tau_n}\}$$

To plan over such subtasks, we model them as sources of reward using the SRS mode. Such sources are defined to have a cylindrical boundary related to the agents' reachability of the grape. A depiction of the model can be seen in Fig.11.

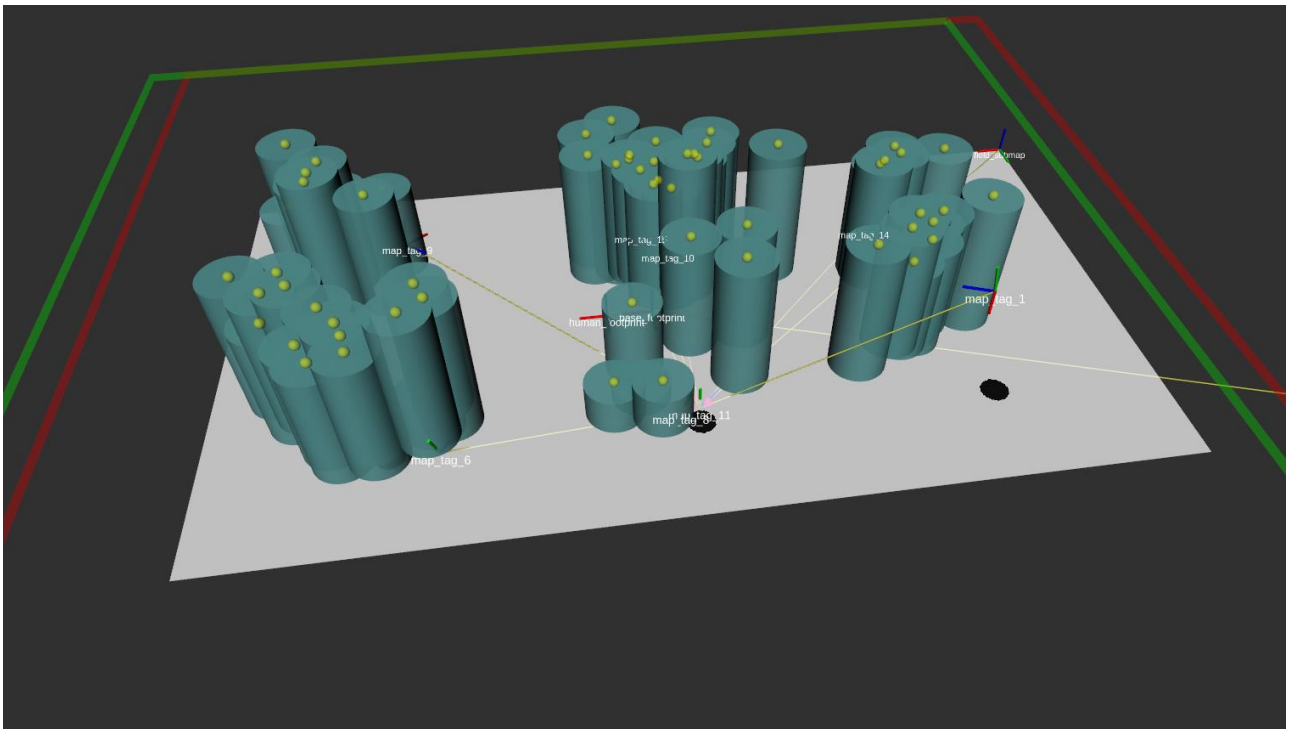


Fig.11. Social Reward Source (SRS) model representation of the harvest task on a 3x2 tree section of the field. Each grape is represented as a source of cylindrical boundary and a step reward model. The radius of the cylinders corresponds to the expected reachability distance of the agents. Yellow dots correspond to the peduncle position of the represented grape bunches.

Through negotiation, the HRT will engage in a generation process of a collaborative plan to tackle this task. We aim to modify and refine the plans calculated by the robot through human input. The negotiation will be carried out in two steps.

3.4.2.1 Negotiation assignment

First, the HRT agrees on a subtask distribution between the team members. The robot uses the previously introduced multi-agent planner to extract a feasible assignment of the subtasks. Then, it is proposed to the human through a visual interface, and it makes any changes to the assignment plan as they feel necessary.

Formally, the assignment proposal can be described as

$$\Gamma = \{\Gamma_1^{\{h\}}, \Gamma_2^{\{r\}}\}$$

where

$$\Gamma_j = \{\gamma_{i_j(1)}, \dots, \gamma_{i_j(n_j)}\}\, \Gamma_j^{\tau_j} \quad \text{with} \quad i_j(k_j) \in \{1 \dots n\}, \forall k_j \in \{1 \dots n_j\}$$

$$\Gamma_j^{\tau_j} = \{\gamma_{i_j(1)}^{\tau_j}, \dots, \gamma_{i_j(n_j)}^{\tau_j}\} \quad \text{with} \quad \tau_j \subseteq \tau_i \in \{1 \dots n\}, \forall i \in \{i | \gamma_i \in \Gamma_j\}$$

$$\bigcup_{j \in N} \Gamma_j = T, \qquad \bigcap_{j \in N} \Gamma_j = \emptyset$$

and will be represented to the human through a visual interface as the following example (Fig.12):

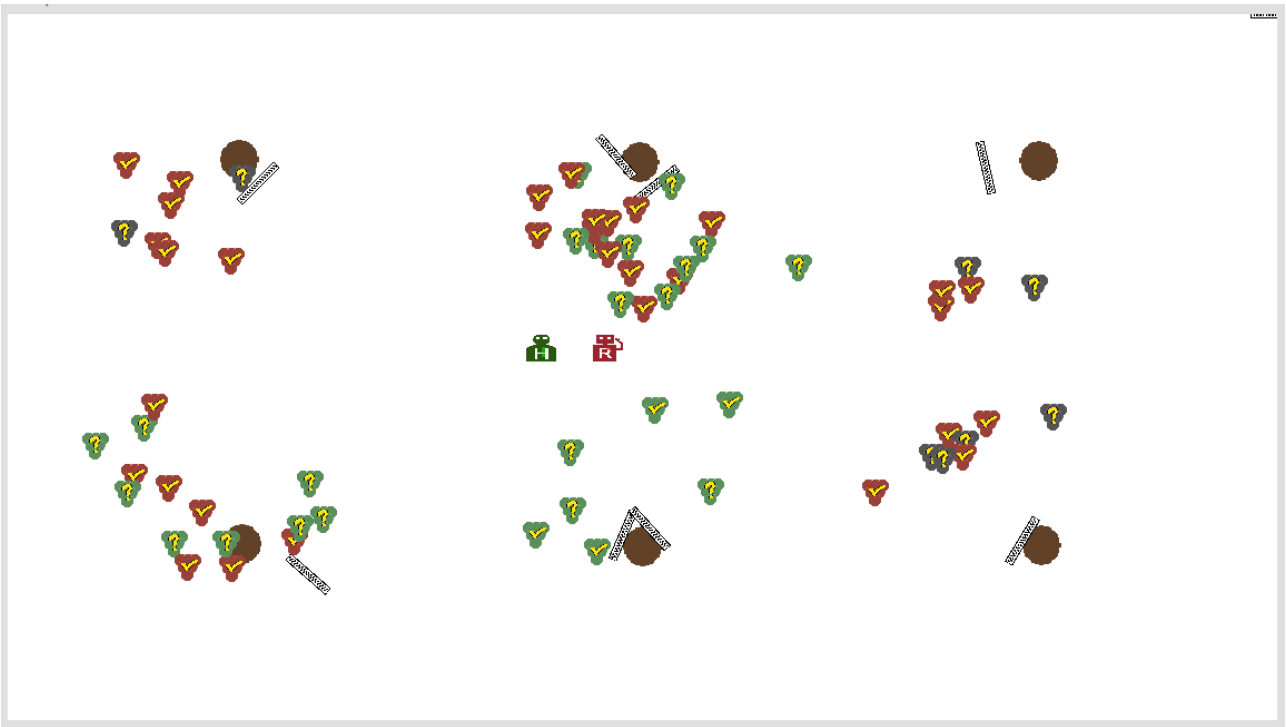


Fig.12. Assignation proposal representation in robot to human communication for the harvest task in a 3x2 tree section of the field. Coloring of the grapes signifies agent assignation of the grape harvesting subtask, whilst ticks and question marks denote the grape quality knowledge of the robot. The assignation plan presented by the robot will only assign grapes to the robot if they are over a certain quality threshold. Brown circles are the vineyard trees, and the gray rectangles are April Tags assigned to the vineyard trees.

3.4.2.2 Navigation assignment

Once the HRT has agreed on an assignation plan, the robot proposes to it a navigation plan. The robot uses the previously introduced multi-agent planner over the new assignment to extract a collaborative motion plan capable of fulfilling it.

Formally, the navigation proposal can be described as

$$x = \{x^{a_1}, \dots, x^{a_m}\}$$

where each element is defined as an ordered sequence of basic movement actions

$$x^a = \{x_1^a, \dots, x_k^a\}$$

Basic movement actions are defined by their goal, encoding the action of moving to it. In this scenario, actions are represented by 2-dimensional points (x,y). Again, it is proposed to the human through a visual interface. Then, they can accept the proposed plan or request a new one. This process can be repeated until satisfaction, as done in Dalmasso et al. [12]. One example of this visualization can be seen in Fig.13.

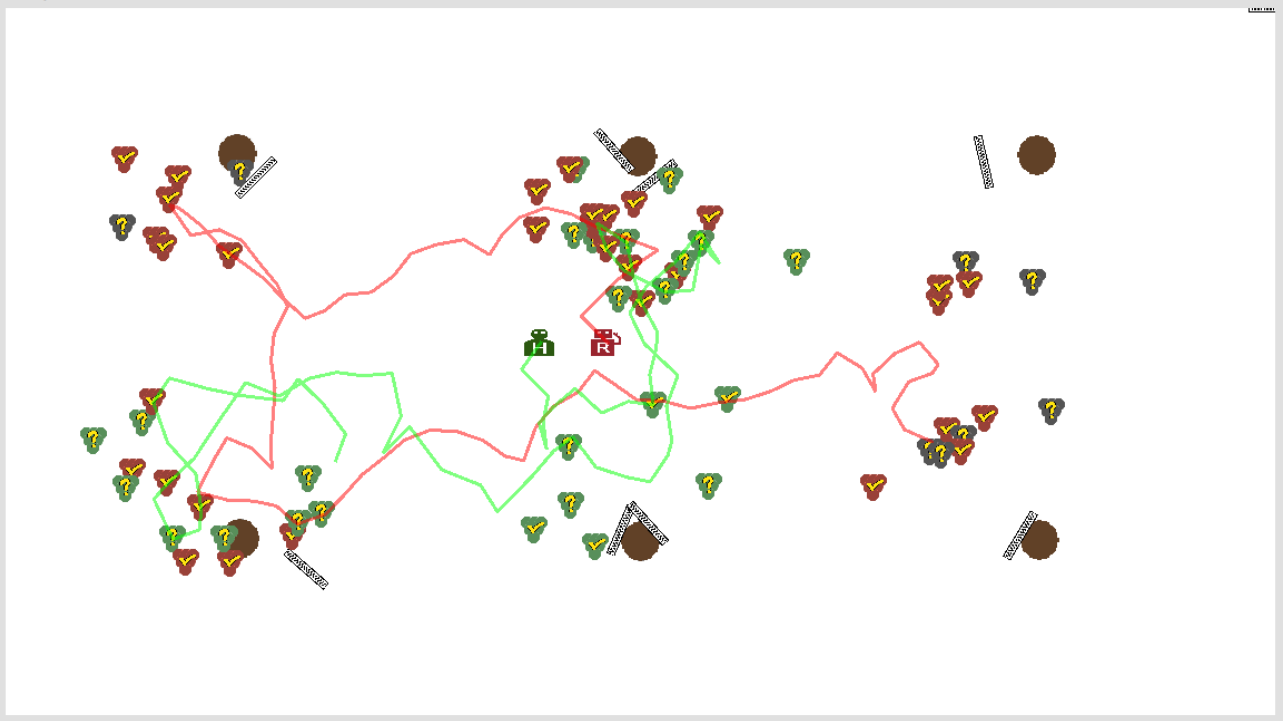


Fig.13. Navigation proposal representation in robot-to-human communication for the harvest task in a 3x2 tree section of the field. Coloring of the grapes signifies agent assignation of the grape harvesting subtask, whilst ticks and question marks denote the grape quality knowledge of the robot. The coloring of the path indicates the represented agent.

4 Conclusions

In this deliverable, we have described different techniques that have been developed for Task 6.1 “Human-Body Prediction” and Task 6.2 “Human-Robot Motion Planning in Agreement”.

In the task 6.1, new human body prediction techniques have been developed and they have been tested in standard datasets and datasets collected in the Corsira Vineryard field in Italy.

In task 6.2, a new multiagent cooperative task model has been presented denominated PIA (Perception-Intention-Action) model and a multiagent planner between a human and a robot has been developed to build a shared planning for harvesting.

The majority of the methods presented in these tasks have been published in top international conferences and journals. Two of the methods explained in this deliverable, the PIA model applied to handover task and the multiagent planning is going to be sent to a top international conference and later on to journals.

5 References

[1] O. Gil and A. Sanfeliu (2023). Human motion trajectory prediction using the Social Force Model for real-time and low computational cost applications, 6th Iberian ROBOT Conference 2023 (ROBOT23), Coimbra, Portugal, pp. 1-12

[2] J. Laplaza, N.A. Rodríguez, J.E. Domínguez, F. Herrero, S. Hernández, A. López, A. Sanfeliu and A. Garrell Zulueta (2022). IVO Robot: A new social robot for Human-Robot collaboration, 2022 ACM/IEEE International Conference on Human-Robot Interaction (RO-MAN22), 2022, Sapporo, Japan, pp. 860–864.

[3] J. Laplaza, F. Moreno-Noguer and A. Sanfeliu (2024), Enhancing robotic collaborative tasks through contextual human motion prediction and intention inference, International Journal of Social Robotics (In review process, 2nd review, confidential copy available at the link[1]).

[4] J. Laplaza, R. Romero, A. Sanfeliu and A. Garrell Zulueta (2023). Body gesture recognition to control a social mobile robot, 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI'23), 2023, Stockholm, in Companion of the HRI'23, pp. 456-460.

[5] J. Laplaza, A. Garrell Zulueta, F. Moreno-Noguer and A. Sanfeliu (2022). Context and intention for 3D human motion prediction: Experimentation and user study in handover tasks, 31st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN22), 2022, Napoli, Italy, pp. 630-635.

[6] [J. Laplaza, F. Moreno-Noguer, A. Sanfeliu (2022). Context Attention: Human Motion Prediction Using Context Information and Deep Learning Attention Models. 5th Iberian Robotics Conference (ROBOT22), 2022, Zaragoza (1) 2022, pp.102-112

[7] J.E. Domínguez, N.A. Rodríguez and A. Sanfeliu (2023). Perception-intention-action cycle as a human acceptable way for improving human-robot collaborative tasks, 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI'23), 2023, Stockholm, in Companion of the HRI'23, pp. 567-571.

[8] J.E. Dominguez, A. Sanfeliu (2023). Inference VS. Explicitness. Do We Really Need the Perfect Predictor? The Human-Robot Collaborative Object Transportation Case. 32nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN23), 2023, Korea.

[1] https://uniroma3.sharepoint.com/:b:/s/project-canopies/EebhyDZ6wNBGkO7gpaOke-wBL4pV2w5awCD_gaUZtGT0-Q?e=K0Sc85

[9] J.E.Dominguez-Vidal, A. Sanfeliu (2023). Improving Human-Robot Interaction Effectiveness in Human-Robot Collaborative Object Transportation using Force Prediction. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023), 1-5 October, Detroit, USA.

[10] M. Peral, A. Sanfeliu and A. Garrell Zulueta (2022). Efficient hand gesture recognition for human-robot interaction. IEEE Robotics and Automation Letters, 7(4): 10272-10279, 2022.

[11] M. Saboia Da Silva, L. Clark, V. Thangavelu, J. Edlund, K. Otsu, G.J. Correa, V.S. Varadharajan, A. Santamaria-Navarro, T. Touma, A. Bouman, H. Melikyan, T. Pailevanian, S. Kim, A. Archanian, T.S. Vaquero, G. Beltrame, N. Napp, G. Pessin and A. Agha-mohammadi (2022). ACHORD: Communication-aware multi-robot coordination with intermittent connectivity. IEEE Robotics and Automation Letters, 7(4): 10184-10191, 2022.

[12] Dalmasso, M., Garrell, A., Domínguez, J.E., Jiménez, P. and Sanfeliu, A., 2021, May. Human-robot collaborative multi-agent path planning using Monte Carlo tree search and social reward sources. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 10133-10138). IEEE.

[13] Dalmasso, M., Domínguez-Vidal, J.E., Torres-Rodríguez, I.J. et al. Shared Task Representation for Human-Robot Collaborative Navigation: The Collaborative Search Case. International Journal of Social Robotics (2023). https://doi.org/10.1007/s12369-023-01067-0

[14] G. Coll-Ribes, I. J. Torres-Rodriguez, Antoni Grau, Edmundo Guerra, Alberto Sanfeliu (2023). Accurate detection and depth estimation of table grapes and peduncles for robot harvesting, combining monocular depth estimation and CNN methods. Computer and Electronics in Agriculture. Vol. 215, December 2023, 108362.

[15] W.O. Chamorro, J. Solà and J. Andrade-Cetto (2022). Event-based line SLAM in real-time. IEEE Robotics and Automation Letters, 7(3): 8146-8153, 2022.