**Deliverable D3.2**

**Dissemination Level (PU)**

---

**PHC-19-2014: Advancing active and healthy ageing with ICT: service robotics within assisted living environments**

**Project Title:**

Robotic Assistant for MCI Patients at home

---

**RAMCIP**

Grant Agreement No: 643433
Research and Innovation Action (RIA)

---

**Deliverable**

<table>
<thead>
<tr>
<th>Deliverable No.</th>
<th>D3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workpackage No.</td>
<td>WP3</td>
</tr>
<tr>
<td>Task No.</td>
<td>T3.1 T3.2</td>
</tr>
<tr>
<td>Task Title</td>
<td>Monitoring and Modelling Human Activity at Home</td>
</tr>
<tr>
<td>Human activity</td>
<td>Monitoring and monitoring the home environment</td>
</tr>
<tr>
<td></td>
<td>Human monitoring</td>
</tr>
<tr>
<td>Lead beneficiary</td>
<td>CERTH</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU – Public</td>
</tr>
<tr>
<td>Nature of Deliverable</td>
<td>Other</td>
</tr>
<tr>
<td>Delivery date</td>
<td>10 November 2016</td>
</tr>
<tr>
<td>Status</td>
<td>F: final</td>
</tr>
<tr>
<td>File Name:</td>
<td>RAMCIP Deliverable 3.2.doc</td>
</tr>
<tr>
<td>Project start date, duration</td>
<td>01 January 2015, 36 Months</td>
</tr>
</tbody>
</table>

---

This project has received funding from the European Union’s Horizon 2020 Research and innovation programme under Grant Agreement nº643433

---

October 2016 1 CERTH
# Authors List

**Leading Author** (Editor)

<table>
<thead>
<tr>
<th>Surname</th>
<th>Initials</th>
<th>Beneficiary Name</th>
<th>Contact email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malassiotis</td>
<td>S.</td>
<td>CERTH</td>
<td><a href="mailto:malasiot@iti.gr">malasiot@iti.gr</a></td>
</tr>
</tbody>
</table>

**Co-authors** (in alphabetic order)

<table>
<thead>
<tr>
<th>#</th>
<th>Surname</th>
<th>Initials</th>
<th>Beneficiary Name</th>
<th>Contact email</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agriomallos</td>
<td>I.</td>
<td>CERTH</td>
<td><a href="mailto:jagrio@iti.gr">jagrio@iti.gr</a></td>
</tr>
<tr>
<td>2</td>
<td>Apostolidis</td>
<td>A.</td>
<td>CERTH</td>
<td><a href="mailto:arapostol@iti.gr">arapostol@iti.gr</a></td>
</tr>
<tr>
<td>3</td>
<td>Charalampous</td>
<td>K.</td>
<td>CERTH</td>
<td><a href="mailto:charalampousk@iti.gr">charalampousk@iti.gr</a></td>
</tr>
<tr>
<td>4</td>
<td>Kargakos</td>
<td>A.</td>
<td>CERTH</td>
<td><a href="mailto:akargakos@iti.gr">akargakos@iti.gr</a></td>
</tr>
<tr>
<td>5</td>
<td>Kiatos</td>
<td>M.</td>
<td>CERTH</td>
<td><a href="mailto:kiatosm@iti.gr">kiatosm@iti.gr</a></td>
</tr>
<tr>
<td>6</td>
<td>Kostavelis</td>
<td>I.</td>
<td>CERTH</td>
<td><a href="mailto:gkostave@iti.gr">gkostave@iti.gr</a></td>
</tr>
<tr>
<td>7</td>
<td>Kyritsis</td>
<td>K.</td>
<td>CERTH</td>
<td><a href="mailto:kokirits@iti.gr">kokirits@iti.gr</a></td>
</tr>
<tr>
<td>8</td>
<td>Rontsis</td>
<td>N.</td>
<td>CERTH</td>
<td><a href="mailto:nrontsis@iti.gr">nrontsis@iti.gr</a></td>
</tr>
<tr>
<td>9</td>
<td>Sigalas</td>
<td>E.</td>
<td>CERTH</td>
<td><a href="mailto:esigalas@iti.gr">esigalas@iti.gr</a></td>
</tr>
<tr>
<td>10</td>
<td>Vasileiadis</td>
<td>M.</td>
<td>CERTH</td>
<td><a href="mailto:mavasile@iti.gr">mavasile@iti.gr</a></td>
</tr>
</tbody>
</table>

# Reviewers List

**List of Reviewers** (in alphabetic order)

<table>
<thead>
<tr>
<th>#</th>
<th>Surname</th>
<th>Initials</th>
<th>Beneficiary Name</th>
<th>Contact email</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Endo</td>
<td>S.</td>
<td>TUM</td>
<td><a href="mailto:endos@itr.ei.tum.de">endos@itr.ei.tum.de</a></td>
</tr>
<tr>
<td>2</td>
<td>Stanczyk</td>
<td>B.</td>
<td>ACCREA</td>
<td><a href="mailto:b.stanczyk@accrea.com">b.stanczyk@accrea.com</a></td>
</tr>
<tr>
<td>Version</td>
<td>Date</td>
<td>Status</td>
<td>Modifications made by</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>--------------------------------------------------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>10/10/2016</td>
<td>1st draft</td>
<td>CERTH</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>25/10/2016</td>
<td>2nd draft, circulated for internal peer-review</td>
<td>CERTH</td>
<td></td>
</tr>
<tr>
<td>3.0</td>
<td>10/11/2016</td>
<td>Final version, following the peer-review comments</td>
<td>CERTH</td>
<td></td>
</tr>
</tbody>
</table>
## List of definitions & abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGR</td>
<td>Background Removal</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>DART</td>
<td>Dense Articulated Real-Time Tracking</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
</tr>
<tr>
<td>DTW</td>
<td>Dynamic Time Warping</td>
</tr>
<tr>
<td>FOV</td>
<td>Field of View</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HOG</td>
<td>Histogram of Oriented Gradients</td>
</tr>
<tr>
<td>HSV</td>
<td>Hue Saturation Value</td>
</tr>
<tr>
<td>ICP</td>
<td>Iterative Closest Point</td>
</tr>
<tr>
<td>LFW</td>
<td>Labelled Faces in the Wild</td>
</tr>
<tr>
<td>MSCR</td>
<td>Most Stable Color Region</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>POV</td>
<td>Point Of View</td>
</tr>
<tr>
<td>RGB</td>
<td>Red Green Blue</td>
</tr>
<tr>
<td>RGB-D</td>
<td>Red Green Blue and Depth</td>
</tr>
<tr>
<td>RHSP</td>
<td>Recurring High-Structured Patches</td>
</tr>
<tr>
<td>SDF</td>
<td>Signed Distance Function</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-Invariant Feature Transform</td>
</tr>
<tr>
<td>subUC</td>
<td>Sub Use Case</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
Executive Summary

This deliverable presents the results of the Task T3.2 "Human activity monitoring" and the advances in the Task T3.1 "Modelling and monitoring the home environment", focusing on the research and development of methods for monitoring the human activity, and the 3D reconstruction and tracking of the RAMCIP robot operational (home) environment and its objects respectively, which have been carried out in the context of the European Union (EU) HORIZON 2020 Programme (H2020) Research and Innovation Action RAMCIP. The deliverable has been developed in the scope of WP3 of the RAMCIP project, responsible for developing the RAMCIP methods for "Monitoring and Modelling Human Activity at Home". Its primary aim is to describe the methods that have been established through the T3.2 efforts, focusing on the development of the RAMCIP robot's core human sensing infrastructure, towards human detection, body pose estimation and action recognition. Moreover, the present deliverable provides an update on the advances that have been achieved in Task T3.1 since the Deliverable D3.1, in regard to the home environment monitoring and the objects reconstruction, recognition and tracking modules.

In order to enable the desired human-oriented perception capabilities on the RAMCIP robot, the T3.2 efforts on human tracking and action recognition focused on four main research subjects, trying to cover the following requirements:

(a) The first, basic pre-requisite is for the robot to be able to detect multiple humans present in its surrounding environment, as this knowledge is essential for a wide range of operational robotic tasks, such as path planning, safe locomotion, human-robot interaction, activity monitoring etc. Due to its critical nature, human detection should be robust and efficient under any environment conditions, regardless of human/robot position, ambient lighting or processing load.

(b) Person identification should be performed on the detected humans, in order for the RAMCIP robot to be able to recognize its main target user, and also alter its behaviour based on the identity of the humans around it in accordance to specific needs of the target use cases (i.e. identify someone as a potentially threatening intruder or as a known user's peer).

(c) After detecting and recognizing its target user, the RAMCIP robot should be able to extract information about her/his body pose/gestures, through detection techniques capable of handling occlusions and noisy data that are highly prone to appear in realistic environments.

(d) Finally, towards achieving high-level human activity and behavior monitoring, the RAMCIP robot should be enabled through the T3.2 outcomes to recognize low-level human actions, by taking into consideration the body posture information in relation to home objects (e.g. open the fridge).

In this line, after a brief introduction (Section 1) and a summary of related works which were investigated in the scope of this task (Section 2), accompanied from a summary of the key advances beyond the state-of-art that were introduced from the present research efforts, the four core sections of the present deliverable describe the results that were obtained in the scope of Human detection (Section 3), Person identification (Section 4), Human body pose tracking (Section 5) and Action recognition (Section 6). The updates in regard to the RAMCIP framework for the monitoring of the home environment are presented in Section 7, while Section 8 sums up the experimental results of all the implemented methods.

More specifically, Section 3 presents the implemented methods for enabling human detection capabilities on the RAMCIP robot. The proposed methods provide human detection either when the humans are inside or out of the robot camera’s FOV, through vision-based or lasercan-based approaches respectively.
The vision-based approaches, which provide a detailed human silhouette mask, include (a) NiTE's built-in movement-triggered Background/Foreground segmentation implementation, along with (b) a custom segmentation algorithm based on the recreation and removal of the scene background utilizing the offline acquired 3D metric map of the current area, which removes the movement prerequisite of NiTE's implementation and can work even when the human in the foreground is stationary. Additionally, an upper-body-detection-based algorithm is proposed, providing an alternative to the BGR methods, while the special case of fallen human detection is also taken into consideration.

Alongside to the vision-based human detection algorithms, a laserscan-based human leg detector provides the approximate position relative to the robot of all the humans around it, who are not inside the camera's FOV. The latter has been integrated with the vision-based human detection framework, leading to a pipeline that can detect multiple humans in the house environment and track their position in space, even when they are not in the robot's FOV.

In Section 4, the methods for the identification of the detected humans are presented. To this end two core implementations are considered. First, a face-based identification approach built upon State-of-Art algorithms, which when provided with a clear close-to-frontal view of a person's face, can robustly identify her/him among a large number of people. Second, a custom person identification algorithm has been developed, which uses soft biometric features and clothing colour information to re-identify a person within a small number of people, without requiring, in contrast to the face-based identification approach, a close-up view of the person in question.

Having detected and identified its target user through the methods described in Sections 3 & 4, the RAMCIP robot should be capable of tracking the body pose of its target end user. To this end, the methods described in Section 5 enable the robot to achieve robust and accurate human pose tracking in a realistic home environment. For the initial pose estimation, the NiTE's built-in human tracker, along with the further human detection solutions developed in T3.2 (as described in Section 3) are utilized, while in order to increase the accuracy and consistency of the extracted skeleton, the core part of our human pose tracking method follows a generative, articulated full body model fitting approach. To this end, we drew upon a State-of-Art articulated human tracker, using in our case a human template customized to fit the target user. The State-of-art method has been extended through the T3.2 efforts through a series of complementary features which are added, leading to an implementation that is better able to handle the challenges that realistic home environments present in the scope of human pose tracking (occlusions, noisy data, partial human views etc.). Through the above and by integrating the tracking approach to single-shot pose estimation, we have developed a hybrid, integrated human pose tracking module, capable to better suit the specificities of the RAMCIP target use cases and application contexts.

By extracting the positions of the joints of the user, the RAMCIP robot is then provided with the necessary data in order to perform low-level action/gesture recognition, as described in Section 6, which constitutes the basic building block towards human activity recognition. A State-of-Art action recognition approach, based on the movement of the individual joints, has been adopted and extended herein. By establishing modifications on the state-of-art algorithm and adding information about the home object that the human interacts with during these actions, the RAMCIP action recognition module is capable to operate in real time, in realistic, continuous video streams that are captured through the robot's camera, by elegantly solving the stream segmentation issue that is inherent in contemporary action recognition approaches and poses a major challenge to their practical applicability in realistic contexts.

Section 7 presents the final functional versions of the core methods towards the reconstruction and tracking of objects in the home environment, which have been
developed from Task T3.1, including updates and advances that have been achieved since the development of the corresponding initial versions described in the Deliverable D3.1, submitted on M16. In this scope, T3.1 has developed a change detection –oriented module capable to continuously track objects with which the user interacts during her/his monitored activities, as well as to cover the detection of fallen/misplaced objects as necessitated by corresponding RAMCIP subUCs. Alongside, T3.1 has further elaborated the small household objects reconstruction toolkit, finalizing and further testing the core method that will be used from the RAMCIP robot during the trials.

The experimental results obtained for the methods developed in T3.2, along with the datasets that were used for this purpose, are presented in Section 8.

The above research efforts and results described in the present deliverable have led to the development of the corresponding human activity monitoring –oriented software modules which are being integrated in the RAMCIP robot. More specifically, the majority of methods described in the present deliverable will already be included on the RAMCIP robot V1, as their use is considered necessary for the use cases of the RAMCIP preliminary tests planned to be established by November 2016.
# Table of Contents

List of definitions & abbreviations ........................................................................... 4
Executive Summary ........................................................................................................ 5
Table of Contents ........................................................................................................... 8
List of figures ................................................................................................................ 10
List of tables .................................................................................................................. 12

1. Introduction .............................................................................................................. 13
   1.1 Scope of the Deliverable ....................................................................................... 13
   1.2 Relation to other Deliverables ........................................................................... 13
   1.3 Deliverable structure ......................................................................................... 14

2. Related work .............................................................................................................. 15
   2.1 Human detection ................................................................................................. 15
   2.2 Human body pose tracking ............................................................................... 15
   2.3 Person identification ......................................................................................... 17
   2.4 Action recognition ............................................................................................. 19
   2.5 RAMCIP advances ......................................................................................... 20

3. Human detection ...................................................................................................... 22
   3.1 Introduction ....................................................................................................... 22
   3.2 Background/foreground segmentation in static scene with moving foreground .................................................................................. 22
   3.3 Background/foreground segmentation based on house map ............................ 22
      3.3.1 Fallen human detection .............................................................................. 23
   3.4 Human detection in a single depth image ............................................................ 24
   3.5 Orientation estimation ....................................................................................... 25
   3.6 Human detection during robot locomotion ......................................................... 26
   3.7 The integrated RAMCIP human detection pipeline ........................................... 27

4. Person identification ................................................................................................ 28
   4.1 Introduction ....................................................................................................... 28
   4.2 Face-based person identification ....................................................................... 28
   4.3 Body-based person identification ..................................................................... 28
   4.4 The integrated RAMCIP person identification pipeline .................................... 30

5. Human pose tracking .............................................................................................. 32
   5.1 Introduction ....................................................................................................... 32
   5.2 Single-shot pose estimation .............................................................................. 32
      5.2.1 NiTE human pose estimation .................................................................... 32
   5.3 Articulated full body model –based pose tracking ............................................ 34
      5.3.1 DART-based human pose tracking ............................................................ 34
   5.4 The integrated RAMCIP human pose tracking pipeline .................................. 42

6. Action Recognition ................................................................................................ 43
   6.1 Introduction ....................................................................................................... 43
   6.2 EigenJoints-based action recognition ................................................................. 43
      6.2.1 Action recognition in realistic input streams ............................................... 44

7. Home Environment Monitoring ............................................................................. 46
   7.1 Dynamic Update of the Hierarchical Semantic Map ......................................... 46
      7.1.1 Change detection during home environment monitoring .......................... 46
7.2 Monitoring of small objects during human activities ......................... 47
7.3 Detection of fallen objects during human activities monitoring .......... 50
7.4 Object reconstruction methods and improvements .......................... 52

8. Experimental Evaluation .................................................................... 53
  8.1 Introduction ...................................................................................... 53
  8.2 Datasets ........................................................................................... 53
    8.2.1 Public Datasets ........................................................................... 53
    8.2.2 CERTH Datasets ....................................................................... 55
    8.2.3 LUM activities dataset .............................................................. 58
  8.3 Human Detection ............................................................................. 62
    8.3.1 Orientation Estimation ............................................................... 65
  8.4 Person Identification ......................................................................... 65
    8.4.1 Face-based Identification ........................................................... 65
    8.4.2 Body-based identification .......................................................... 68
  8.5 Human pose tracking ...................................................................... 70
    8.5.1 DART-based human pose tracking .............................................. 70
  8.6 Action Recognition .......................................................................... 75
  8.7 Home environment monitoring ....................................................... 79
    8.7.1 Evaluation of 3D object reconstruction method ........................... 79
    8.7.2 Evaluation of large object state tracking ..................................... 82

9. Conclusions ....................................................................................... 85

References ............................................................................................. 87
List of figures

Figure 1: The CERTH BGR algorithm: Top left - current robot view, top right - reconstructed default background view, bottom left - octree change detection results, bottom right - final filtered foreground ........................................ 23
Figure 2: The integrated RAMCIP human detection pipeline ......................... 27
Figure 3: Selected body areas for color-based feature extraction: Green - torso, Red: Legs. The arms and lower legs are not used because subjects often wear short-sleeved clothes or shorts, which may result in the skin color being most dominant under specific viewpoints ....................................................... 30
Figure 4: The integrated RAMCIP person identification pipeline ..................... 31
Figure 5: The skeleton joints estimated by the NiTE human tracking implementation .................................................................................................................. 33
Figure 6: Body part segmentation of the human body, during pixel classification 33
Figure 7: The kinematic tree of the articulated human template, including the joints and the rigid body parts ................................................................. 35
Figure 8: Human template customization. Left: [g=1,s=0], Right [g=1.2, s=0.5] ................................................................................................................................. 36
Figure 9: Reciprocal error correction. Left: λ=0, Right λ=0.5 ......................... 38
Figure 10: Joint visibility correction in occluded view. Left: without correction, both legs converge to the visible left leg, Right: With correction, the right occluded leg remains at the default position and is not rendered. Moreover, the natural leg pose anchors the lower body on the floor plane, thus improving the fit of the upper body as well ......................................................... 39
Figure 11: Leg collision fix ............................................................................. 40
Figure 12: The initialization pipeline of the RAMCIP human pose tracker ....... 41
Figure 13: Examples of successful (left) and failed (right) human pose tracking 41
Figure 14: The integrated RAMCIP human pose estimation pipeline .......... 42
Figure 15: The eigenJoints algorithm data processing pipeline [73] .......... 43
Figure 16: Calculation of the monitoring workspace, a) upper left: the RGB reference image, b) upper right: the retrieved supporting surface from the hierarchical semantic map, c) lower left: the detected plane and d) lower right: the workspace defined Convex Hull, with the remaining clusters of interest ...... 47
Figure 17: Sequential parsing of calculated clusters and inference of the object detection and recognition algorithm ......................................................... 48
Figure 18: Human excluded from the scene utilizing Nite's background removal algorithm, i.e. scene analyzer; removing user from background ....................... 49
Figure 19: User picking a cup from the monitored workspace ...................... 49
Figure 20: User placing the cup back to the monitored workspace ............... 50
Figure 21: Sequence of snapshots from robot platform during monitoring for fallen object. The first row contains reference frames exhibiting a user with a fallen cup. The second row of images corresponds to the detection of the user and his/her exclusion of the scene using the background/foreground segmentation human detection module. The third row of images describes the detected workspace along with the defined clusters. The last row presents the output of the detection algorithm operating over all the defined clusters ..................... 51
Figure 22: Sample frames from the EVAL dataset ......................................... 53
Figure 23: Sample sequences of the actions in the MSR Action3D dataset. Top - Draw tick, bottom - Tennis serve .............................................................. 55
Figure 24: Sample frames from the CERTH face dataset ............................ 56
Figure 25: Sample sequences from the CERTH action dataset. a) 1 - Standing to sitting, b) 12 - Handle pillbox sitting, c) 14 - Eating sitting, d) 15 - Open cupboard door from mid handle .................................................. 58
Figure 26: Rendering of the top-down view of the LUM room .................... 59
Figure 27: Sample RGB and Depth frames from the LUM activities dataset ...... 62
Figure 28: Sample scenes where the single shot human detection algorithm failed ............................................................................................................. 63
Figure 29: Sample scenes where the single shot human detection algorithm successfully detected the human and extracted the silhouette mask .......... 64
Figure 30: Percentage of samples for which the deviation from ground truth lies within the range of n bins.................................................................................. 65
Figure 31: Sample frames from the human body-based identification evaluation on the CERTH action-dataset: Left-training template, right-identified frames .... 70
Figure 32: Sample frames from the human pose tracking evaluation on the EVAL dataset ........................................................................................................ 72
Figure 33: Sample frames of successful human pose tracking in the CERTH action dataset. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NiTE estimator ........................................ 73
Figure 34: Sample frames of tracking in realistic conditions where the CERTH tracker outperforms the NiTE estimator. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NiTE estimator. a) Lower body occlusion while standing, b) Turn around, c) Leg crossing, d) Sitting, e) Lower body occlusion while seated, f) heavy body occlusion near the edges of the FOV .................................................. 74
Figure 35: Sample frames of failed human pose tracking in the CERTH action dataset, due to the human’s interaction with a large object. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NiTE estimator ........................................................................ 75
Figure 36: Confusion matrix of the RAMCIP action recognition evaluation on the CERTH action dataset .................................................................................. 78
Figure 37: The cupboard model (red), with all its articulated models in different colours. All cupboard doors were labelled as “closed”............................. 82
Figure 38: The cupboard model (red) is registered to the scene, and the left door of the cupboard (pink) is found to be open. From the other two doors of the cupboard, only the rightmost one was detected to be closed. The middle door was occluded by the user and its state couldn’t be detected ....................... 82
Figure 39: A series of instances where the user interacted with the fridge. The parts of the fridge (base, upper door, lower door) are illustrated with different colours. The fridge model was registered to the scene, and the states of its doors were recognized in each case. In the first row, the lower fridge door was recognized as “Open” and the upper door as “Closed”. In the last three rows, the upper door was recognized again as “Closed” and the state of the lower door was correctly recognized as “Open” during the human interaction with partial occlusions. ........................................................................ 83
List of tables

Table 1: Geometric features used in the laser scan-based leg detector ............... 26
Table 2: List of features used for body-based person identification .................. 29
Table 3: Joints and markers tracked by the Vicon motion capture system, in the EVAL dataset ........................................................................................................ 54
Table 4: Actions recorded in the MSR Action3D Dataset ............................... 54
Table 5: Recording scenarios in the CERTH face dataset ............................... 56
Table 6: List of action sequences recorded in the CERTH action dataset ....... 57
Table 7: Activity scenarios recorded in the LUM activities dataset ............... 59
Table 8: Experimental evaluation of the RAMCIP human detection and silhouette extraction implementation, on a random subset of the LUM activities dataset .... 63
Table 9: Experimental evaluation of the RAMCIP body orientation estimation implementation, on a randomly selected subset of the CERTH action dataset .... 65
Table 10: Face detection rates for the initial randomly selected training set ...... 66
Table 11: Face detection and identification rates for the randomly selected test set ................................................................................................................................................ 66
Table 12: Face detection and identification rates for the randomly selected test set, with the use of the memory module ......................................................... 67
Table 13: Allocation of samples where face detection failed, per recording scenario ......................................................................................................................... 68
Table 14: Identification accuracy for each feature set, on 6-subject subset of the CERTH action dataset ........................................................................................................ 69
Table 15: Identification accuracy for each subject, on 6-subject subset of the CERTH action dataset ........................................................................................................ 69
Table 16: Experimental evaluation of the DART-based RAMCIP human pose tracking implementation, on the EVAL dataset ............................................. 71
Table 17: Experimental results for non-cross subject testing of the eigenJoints-based RAMCIP action recognition implementation on the MSR Action3D dataset 75
Table 18: Experimental results for cross subject testing of the eigenJoints-based action recognition implementation on the MSR Action3D dataset ............... 76
Table 19: Comparison between the original eigenJoints approach and the RAMCIP extension ............................................................................................................. 77
Table 20: Experimental evaluation of the RAMCIP action recognition pipeline on the CERTH action dataset ............................................................................. 77
Table 21: Summary of the experimental evaluation of the 3D reconstruction method. The deviation in the Translation is measured in mm, and the deviations from the three rotational axes dRx, dRy, dRz are in degrees. For the symmetric objects only the orientation of the one principal axis is measured. The detection method is assessed for various Distances (0.9, 1.4 and 1.8)m between the RGBD sensor and the object .................................................................................. 81
1. Introduction

1.1 Scope of the Deliverable

The scope of this deliverable is to present the core methods that have been researched and developed in the context of T3.2, towards establishing the RAMCIP robot's human sensing infrastructure with respect to human detection, identification, pose tracking and action recognition. Moreover, it describes the evolution of the home environment modelling and monitoring framework methods developed in T3.1, which were initially presented in the deliverable D3.1.

The specific methods reported herein first concern the detection of humans present in the vicinity of the robotic platform. Computer-vision techniques are employed for the detection of people who are inside the robot camera’s FOV, while a secondary, non-vision, laser-based method is also employed in order to allow approximate human detection in the areas outside of the camera’s FOV. The second aim of the T3.2 methods concerns the identification of the detected humans, including both the initial identification of people who enter the robot’s operational space for the first time, and the re-identification of previously identified people who re-enter the working area.

Human detection and identification allow the RAMCIP robot to recognize and locate its target user, thus proceeding to human activity monitoring. To this end two further core modules have been researched and developed in the context of Task 3.2 of the RAMCIP project: (a) the human body pose tracking module, which provides information about the position of the skeletal joints of the tracked person, and (b) the action recognition module, which, by using the extracted skeleton data, detects sequences of discrete body postures and gestures, and clusters them into low-level activities, taking into account objects that the user interacts with.

Moreover, the updates and advancements concerning the methods developed for Task T3.1, which are described in the deliverable, introduce further refinements to the already developed methods, so as to seamlessly operate, within a fully functional ROS-based computational framework for modelling and monitoring the home environment of the RAMCIP robot user.

1.2 Relation to other Deliverables

This deliverable presents the results of the Task T3.2 of the RAMCIP project, and the final updates on the work corresponding to the Task 3.1, thus connecting the current deliverable with the deliverable D3.1.

Moreover, both tasks T3.2 and T3.1, which are responsible for establishing the RAMCIP robot's computational infrastructure for monitoring the user’s activities and the state of her/his domestic environment respectively, take closely into account the use cases and requirements defined in the deliverable D2.1, as well as the technical specifications and overall architecture of the RAMCIP robotic system, defined in the project deliverables D2.2 and D2.3.

Furthermore, the methods of the present deliverable will provide input for the RAMCIP methods of the rest of the tasks of WP3: Fine grained body motion analysis in T3.3, User behavior modelling and monitoring in T3.4, Modelling physical and cognitive user skills in T3.5, contributing to deliverables D3.3, D3.4 and D3.5 respectively. Alongside, a relation exists among the T3.2 efforts on human activity monitoring and the T3.6 efforts on developing the robot’s cognitive functions, as well as the WP4 human-robot communication methods.

Finally, the advancements on the home environment modelling framework of T3.1, formulate a basic prerequisite for the RAMCIP robot's methods for safe navigation and robotic manipulations towards objects reaching and grasping,
being developed in tasks T5.1 and T5.2, reported in the deliverables D5.1 and D5.2 respectively.

### 1.3 Deliverable structure

The deliverable is structured and organized in the following sections/chapters:

**Chapter 1** provides the introduction of this deliverable and outlines its scope as well as its relation to other deliverables.

**Chapter 2** presents the related work and exhibits the SoA methodologies for each of the software components that have been developed in the context of task T3.2. The related work of each software component is presented individually for improved readability. At the end of this chapter, the key advances of the T3.2 efforts are summarized in order to help the reader to better position the conducted work with respect to the state-of-the-art methodologies.

**Chapter 3** presents the methodologies developed for human detection and silhouette extraction. Three vision-based approaches have been investigated and are described, for detecting humans inside the camera's FOV: a) NiTE's built-in movement-triggered BGR algorithm, b) a novel BGR method based on the dynamic recreation of the scene background from the offline acquired 3D metric map of the current area, along with a custom implementation targeting specifically the detection of fallen persons and c) a single-shot upper-body detector for human detection from depth images. Additionally, a body orientation estimation algorithm is presented, as well as a laserscan-based human leg detector which can provide the approximate position of humans in the robot’s operating space, when they are not inside the camera’s FOV.

**Chapter 4** presents the developed methods for person identification, utilizing a face-based identification approach, which when provided with a clear close-to-frontal view of a person’s face, can robustly identify her/him among a large number of people, and a custom identification algorithm which uses soft biometric features and clothing colour information to re-identify a person during the robot’s operation, within a small number of people.

**Chapter 5** presents the methods developed for tracking the human body pose, including, NiTE’s built-in human tracker, which is used for the initial pose estimation and a custom DART-based articulated tracker developed by CERTH, which is used to track the pose of the target user. Moreover, a series of complementary tracking features are also described, which, combined with the NiTE pose estimator and the DART-based pose tracker developed herein, compose a hybrid body pose tracking solution able to handle the challenges that a realistic home environment may present in the scope of human pose tracking (occlusions, noisy data, partial human views etc.)

**Chapter 6** describes the methodology behind the developed skeleton joints-based action recognition methods, while also extending the methodology by incorporating information about home objects that the human may use during these actions.

**Section 7** presents the final functional versions of the core methods towards the reconstruction and tracking of objects in the home environment, which were developed for Task T3.1, including all the updates and advances that have been achieved since the development of the corresponding initial versions described in Deliverable D3.1.

**Section 8** provides a summary of the preliminary results obtained from the evaluation of the methods described in the chapters above. The datasets used and created for the Task T3.2 are also presented.

Finally, **Chapter 9** provides the conclusions about the methods presented in the previous chapters, discusses the results and summarizes the findings.
2. Related work

In the present chapter, state-of-art approaches that fall in the scope of the research efforts of the present deliverable are described. Specifically, the state of the art survey described below focuses on four key research topics that are essential towards achieving efficient human activity monitoring; human detection and body pose tracking (Sections 2.1 and 2.2 respectively), person identification (Section 2.3) and action recognition (Section 2.4). Finally, the key RAMCIP advances that are introduced from the research efforts reported in the present deliverable are summarized in Section 2.5.

2.1 Human detection

Human detection refers to the localization of all human subjects that are present in an image or video sequence. This problem corresponds to determining regions, typically the smallest rectangular bounding boxes, in the image or video sequence that enclose humans. Human detection is a challenging problem because of the large variations in visual appearance, due to various viewpoints, different clothes and changes of illumination of the target people. Moreover, possible occlusions and complex backgrounds further add to the complexity of the problem.

Generally, a human detector mainly has two components: a feature extraction algorithm that encodes the input data as a feature vector, and a detection model that locates the target human bodies according to the extracted feature vector.

A good feature extraction algorithm provides robust invariance to the large variations of human bodies while extracting enough information for detection. Dalal & Triggs [1] suggest the Histograms of Oriented Gradients (HOG) features that are robust to significant changes in image illumination and colour as well as small changes in image contour locations and directions. The HOG features have proven effective for the detection of human and other shape-based object categories. Zhang et al. [2] and Wang et al. [3] show that HOG-LBP features, a combination of HOG and Local Binary Patterns [4], under some circumstances, can further improve the detection performance. This fact implies that the HOG features also have self-bias, and thus could be improved by combining with the other kinds of features. Other popular features for detection include the Scale-Invariant Feature Transform (SIFT) [5], Haar-like features [6], Wavelet features [7], Shape-Context features [8], and so on.

For the detection model, Felzenszwalb [9] introduced the Deformable Part-based Model (DP-M), which describes an object as a root block surrounded with several movable parts, and thus can alleviate the problems of appearance variations and occlusion. Girshick et al. [10] has proposed the person Grammar Model which extends DPM from simple star-structure to general hierarchical structure, while in [11] a R-CNN model is constructed, based deep CNN features which have already been used for general object detection. An extensive review of the state-of-the-art of human detection can be found in [12].

2.2 Human body pose tracking

Human pose estimation and tracking refers to the process of detecting and extracting the positions of the joints of the human body from, either single or sequences of, RGB and Depth images or 3D point-clouds, in order to reconstruct the skeletal structure and provide information about body motion.

State of the art human pose tracking algorithms tend to fall into two categories. Discriminative approaches use large training datasets and machine learning techniques in order to map the extracted features from the input data to body parts and poses. Generative approaches, on the other hand, try to match the input data to articulated body templates by minimizing an objective function, utilizing various optimization techniques. There are also hybrid approaches which
combine discriminative and generative techniques towards pose estimation. While initial implementations relied mainly on RGB data, the recent development of low-cost high-accuracy RGB-D sensors, such as the Microsoft Kinect and the ASUS Xtion, has pushed the research community towards approaches that utilize the sparse partial-view depth/3D data that these sensors offer.

Discriminative approaches have been successfully used for human pose estimation, utilizing both RGB and Depth images [13]. They rely on large datasets in order to directly train the conditional probability of a body part within an image. As a result, they provide robust human pose estimation from a single frame, without requiring any prior knowledge. On the other hand, discriminative approaches require extremely large and diverse datasets in order to train recognition models, which are hard to acquire. Moreover, they often fail to efficiently handle occlusions, leading to lower detection accuracy, especially from challenging viewpoints, such as side views or views obstructed by obstacles (i.e. human sitting behind a desk).

Bourdev & Malik [14] and Wang et al. [15] have introduced the concept of Poselets, which refer to pieces of human poses that are tightly clustered in both appearance and configuration spaces, for human body part detection from RGB images. They extract HOG features and use linear SVMs to train classifiers, employing a two-step training process, in order to detect poselets for each body part in images, through multi-scale scanning. Andriluka et al. [16] use a Pictorial Structure Model for human pose estimation. The human body is represented through a kinematic tree, while the appearance of body parts is modelled using densely sampled shape context descriptors and discriminatively trained AdaBoost classifiers. Toshev & Szegedy [17] propose a method for human pose estimation based on Deep Neural networks. They formulate pose estimation as a joint regression problem and use a 7-layered generic convolutional DNN to regress the location of each body joint within the image.

Recently, the research focus has shifted towards the utilization of depth data, for body part detection, taking advantage of the fact that depth images, contrary to RGB images, are not affected by changes in lighting conditions and can provide a 3D representation of the scene. Plagemann et al. [18] propose a novel interest point detector suitable for mesh and depth data. The interest points, called Accumulative Geodesic Extrema, are computed by incrementally maximizing geodesic distances on the surface of the 3D mesh. Small depth image patches surrounding these points are then used as local descriptors in order to train a boosted classifier. Shotton et al. [19] use randomized decision trees and forests for body part detection and treat the body part segmentation as a per-pixel classification task. Each pixel in the depth image is evaluated separately and a weighted mean shift-based approach is used to estimate the body pose based on the inferred body part probability on each pixel. Similarly, Pons-Moll et al. [20] also utilize randomized decision forests, but propose an alternative training approach by employing the Metric Space Information Gain training objective.

Generative approaches try to estimate the human pose by fitting a human body model to the observed data. The body model is usually comprised of a 3D mesh that represents the shape of the human and an articulated skeleton which is made up by the skeletal joints and the fixed length bones. The pose of the body model is described by a pose vector which includes the relative rotation of each joint and the global translation of the whole model. As a result, the fitting process involves the estimation of the optimal values of the pose vector that minimize an objective function. This is achieved by calculating the partial derivatives of the objective function with respect to the pose vector and using optimization algorithms to estimate the optimal solution.

Generative approaches offer higher detection accuracy and better occlusion handling compared to discriminative techniques, mainly due to the use of a life-like human template. While they don’t require any prior training, generative
approaches do need a rough initial pose of the body in order for the optimization algorithm to converge, as they essentially track the human pose from an initial known position. Moreover, if there are large body movements between consecutive observations, the optimization process can get trapped in local minima and fail to recover properly, thus making it necessary to achieve a high framerate in order ensure smooth body motion between frames.

Ganapathi et al. [21] use a Dynamic Bayesian Network to model human motion and introduce an enhanced ICP-based model which utilizes free space constraints, termed Ray-Constrained ICP Model. A generalization of Signed Distance Functions for articulated objects is introduced by Schmidt et al. [22]. Objects are represented by a symmetric version of the articulated SDF and gradient based optimization is used to estimate the pose. Ye & Yang [23] relate the observed data to the body template using a Gaussian Mixture Model, while the Expectation Maximization algorithm is used for simultaneous pose estimation and body template shape adaptation. In [24] a Generalized Sum-of-Gaussians model for human shape modelling is presented. The observed data are represented by isotropic Gaussians through octree partitioning, and multivariate Gaussian kernel correlation is employed as a similarity measure between the body template and the observation.

Combinations of generative and discriminative approaches have also been proposed, in order to further increase the pose estimation accuracy. In these hybrid approaches, pose tracking is performed through generative techniques, while the discriminative algorithms are usually used to initialize the pose tracker, and recover it from failure when it gets trapped in local minima.

Wei et al. [25] use a gradient-based optimizer to track the human pose, and combine it with a randomized decision trees-based body part detector used only for initialization and tracking failure recovery. Ganapathi et al. [26] also utilize a gradient-based optimizer in conjunction with the body part detector from [18], however, in contrast to [25] the body part proposals are taken into consideration in every frame. Baak et al. [27] and Ye et al. [28] initialize the generative tracker by finding the pose most similar to the observation, from a database of pre-rendered mesh models, using a matching process based on [18] and on PCA of normalized depth images respectively.

2.3 Person identification

Historically, the human face has been successfully used for person identification, as it provides highly distinctive features among different people, which remain consistent for long periods of time. In recent years automated face verification and identification has also witnessed significant progress.

The standard approach for recognizing a face after it has been localized and cropped from an input image is to compute a low dimensional facial representation or signature. A lot of successful face representation techniques have been already proposed [29][30][31], yet all state-of-the-art results have been obtained via deep CNN. CNN learning may be based on a classification loss function where the output of the net is the recognized identity [32][33], or instead by learning a similarity/dissimilarity metric either by means of a Siamese architecture [34][35] or via a triplet loss layer [36]. The best results on the LFW [18] benchmark are obtained by ensembles of deep nets learned on different parts of the face [37][38][39] and recently by training a triple loss CNN on millions of face images [36]. The embedding based methods learn a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. This low dimensional face representation is obtained as the output of the top hidden layer [34][38][32][35][40]. Once this space has been produced, tasks such as face recognition and verification can be performed using standard techniques e.g.
using the L2-distance [29][35][36] or cosine similarity [38][40] over the descriptors.

The face recognition problem can become challenging due to the great variability in head rotation and tilt, lighting intensity and angle, facial expression, aging, etc. Data driven methods above exhibit some robustness by exploiting large datasets of images captured under uncontrolled conditions where these factors may be present. However even with this approach it is quickly becoming impractical to cover completely the aforementioned variability. When 3D data is available, it may be used to alleviate for example pose and illumination variability [41] or use 3D geometry as an additional modality [42].

A great amount of research effort has also been put towards person re-identification [43] based on colour cues from clothing appearance. Most of the existing descriptors are based on a multiple part-multiple component (MPMC) representation: they subdivide the body into several parts to deal with its non-rigid nature and represent each part as a set of components using various kinds of local or global features. SDALF [44] subdivides the body into left and right torso and legs. Three kinds of features are extracted from each part: 1) colour histograms in the HSV colour space, 2) MSCRs and 3) RHSPs. To extract the MSCR and RHSP, several image patches are randomly sampled and then clustered to find the most significant ones. In [45], the body is subdivided into head, torso, and legs as in [44] and each part is described using weighted Gaussian colour histograms features, pyramid of histograms of orientation gradients, and Haralick features. In [46], the body is subdivided into upper and lower parts: each part is represented using the MPEG7 dominant colour descriptor, and a learning algorithm is used to find the most discriminative appearance model. MCMimpl [47] subdivides the body into torso and legs, randomly extracts from each part of rectangular possibly overlapping patches, and represents them with HSV colour histograms. In [48], dense colour histograms in several colour spaces and different texture features are extracted from four body parts (upper and lower torso, upper and lower legs). A nonlinear warp function between features from two cameras is learnt, to deal with large changes in appearance between them, due to different lighting conditions and poses, occlusion, and background clutter. In [49], the body is subdivided into six horizontal strips, and clothing appearance is modelled in terms of colour and texture features as a function of the pose. A classifier is trained offline to identify the most discriminative features, and subject-discriminative features are further learnt online. In [50], a spatial pyramid is built by dividing an image into overlapping horizontal stripes of 16-pixels height, with colour histograms and histograms of oriented gradients being computed for each strip. A ranking method specific to re-identification, based on sparse discriminative classifiers, is also proposed in [50]. Other methods exploit more refined subdivisions. In [51], a body part detector is used to find fifteen non-overlapping square cells, corresponding to stable regions of the silhouette, which are represented by a covariance (COV) descriptor in terms of colour gradients. Colour histogram equalization is performed to improve robustness to changing lighting conditions. Descriptor generation and matching are performed through a pyramid matching kernel. The subdivision of [52] is based on decomposable triangulated graphs, and each part is described by colour and shape features. Pictorial structures are used in [53] to detect chest, head, thighs, and legs, which are described by HSV histograms and MSCR patches as in [44].

Other approaches treat the body as a whole and represent it using various kinds of features: haar-like features [46], SIFT-like interest points [52][54][55], texture (Schmid and Gabor filters) and colour (histograms in different colour spaces) [56], global colour descriptors (histograms, spatiograms, and colour/path length) [57], 4-D multicoulor height histograms and transform-normalized RGB
(illumination-invariant) [58] and biologically-inspired (BI) features and COV descriptors capturing shape, location, and colour information [59].

In addition to colour cues, anthropometric features [60] have also been employed towards person re-identification, especially since novel RGB-D sensors [19] have made it viable to extract such features. In [61] a series of anthropometric measures are extracted from the skeleton, using front and rear poses, along with three geodesic distances estimated from the 3D mesh provided by the Kinect RGB-D camera. In [62], a specific setting is considered in which the cameras are installed on the floor after an entrance door. The proposed anthropometric measures (extracted from a sequence of frames) are the individual’s average blob height, area and projected volume to the floor, and blob speed. Two kinds of descriptors are considered in [63], exploiting Kinect sensors: 13 anthropometric measures extracted from the body joints and a point cloud model of human body. Finally, Gianaria et al. [64][65] extract anthropometric measures from RGB-D for person identification through gait recognition.

2.4 Action recognition

Human action recognition has been an active area of research for the past several decades due to its applications in surveillance, video games, robotics, etc. In the past few decades, several approaches have been proposed for recognizing human actions from monocular RGB video sequences [66]. Unfortunately, the limitations of the monocular RGB data (sensitivity to illumination, various viewpoints etc.), as well as the lack of 3D space information, have hindered the recognition potential of such methods. However, the recent advent of cost-effective depth sensors has alleviated some of these problems, as RGB-D sensors can provide 3D depth data of the scene, which is robust to illumination changes and offers more useful information to recover 3D human skeletons. As a result, these advances have resulted in a renewed interest in skeleton-based human action recognition [67].

Existing skeleton-based human action recognition approaches can be broadly grouped into two main categories: joint-based approaches and body part-based approaches. Joint-based approaches consider the human skeleton as a set of points, whereas body part-based approaches consider human skeleton as a connected set of rigid segments. Approaches that use joint angles can also be classified as part-based approaches since joint angles measure the geometry between (directly) connected pairs of body parts.

Joint-based approaches consider the human skeleton simply as a set of points. These approaches try to model the motion of either individual joints or combinations of joints using various features. In [68] human skeletons are represented by their 3D joint locations, and the joint trajectories are modelled using a temporal hierarchy of covariance descriptors, while a similar representation is also used with Hidden Markov models (HMMs) in [69]. Sheikh et al. [70] use a set of 13 joint trajectories in a 4-D XYZ-T space to represent a human action, and their affine projections are compared using a subspace angles-based view-invariant similarity measure. In [71] the human skeleton is represented by the pairwise relative positions of the joints, and the temporal evolutions of this representation are modelled using a hierarchy of Fourier coefficients. An actionlet-based approach is utilized, in which discriminative joint combinations are selected using a multiple kernel learning approach. Yang & Tian [72][73] represent the human skeleton using relative joint positions, temporal joint displacement and offset of the joints with respect to the initial frame. Action classification is performed using the Naive-Bayes nearest neighbour rule in a lower dimensional space constructed using principal component analysis (PCA).Zhu et al. [74] utilize a similar skeletal representation in conjunction with random forests. A view invariant human skeleton representation is proposed in [75] by quantizing the 3D joint locations into histograms based on their
orientations with respect to a coordinate system fixed at the hip center. The temporal evolutions of this view-invariant representation are modelled using HMMs

Body part-based approaches consider the human skeleton as a connected set of rigid segments. In [76] the human body is divided into five different parts, with the human actions being represented by the motion parameters of individual body parts like horizontal and vertical translations, in-plane rotations, etc. Principal component analysis is used to represent an action as a linear combination of a set of action basis, and classification is performed by comparing the PCA coefficients. Chaudhry et al. [77] divide the human skeleton hierarchically into smaller parts, with each part is represented using certain bio-inspired shape features. The temporal evolutions of these bio-inspired features are then modelled using linear dynamical systems. 3D joint angles are used in [78], with DTW distance used for comparing their temporal evolutions. Ofli et al. [79] propose the automatic selection of a few informative skeletal joints at each time instance, based on highly interpretable measures such as mean or variance of the joint angles, maximum angular velocity of the joints, etc. Human actions are then represented as sequences of these informative joints, and compared using the Levenshtein distance. Skeletal sequences are represented in [80] using pairwise affinities between joint angle trajectories, and then classified using linear SVM.

2.5 RAMCIP advances

Within the context of the WP3 and specifically of Task 3.2 of the RAMCIP project, emphasis was given to the development of implementations that will perform robustly in real home environments, as the accuracy and capacity of the human monitoring modules can be severely affected by the conditions in the working area of the RAMCIP robot. To this end, state-of-the-art algorithms were implemented and tested both on publicly available datasets captured in controlled laboratory environments, as well as on datasets recorded in more realistic environments, under conditions similar to those that the RAMCIP robot is expected encounter when deployed. Specifically, four research topics were elaborated towards achieving robust operation in realistic environments: a) human detection, b) person identification, c) human pose tracking and d) human action recognition.

Considering the human detection process, two novel approaches are introduced, in order to achieve human detection in stationary scenes, which cannot be currently handled by commercially available implementations such as NiTE’s BGR algorithm. Both approaches utilize the robotic platform’s localization information in order to employ a BGR algorithm and a top-down-view based human detector respectively, in order to achieve human detection even when the target user is not moving.

On the topic of person identification, the fusion of soft biometrics and colour cues provides a computationally inexpensive person re-identification implementation which is used to complement the already established faced-based person identification techniques that formulate the further part of the RAMCIP person identification capabilities needed for the project’s target subUCs.

The issue of robust human pose tracking in real environments is also tackled, as most current state-of-the-art implementations do not take into consideration the challenges presented in realistic human monitoring conditions, but are limited in controlled “ideal” laboratory-based recordings. An articulated human pose tracker is introduced, drawing upon a current state-of-the-art implementation, which is further enhanced by a series of complementary features targeting specific problems encountered in real workspaces. Moreover, the developed tracker is seamlessly integrated with NiTE’s human pose estimator, which is currently the most widely-used tracking solution available, towards efficient and accurate
human pose estimation under the monitoring circumstances involved in the RAMCIP target use cases.

Finally, a state-of-the-art method for action recognition has been further advanced by introducing a further descriptor and by incorporating into the action recognition pipeline, also information about small objects handled by the user. In this scope, we take advantage of the advances in object tracking and monitoring of the home environment, initially reported in the Deliverable D3.1 and further elaborated as reported herein. In the scope of home environment objects monitoring, an integrated change detection–based monitoring framework has been developed, capable to cover the corresponding needs of the RAMCIP target use cases.
3. Human detection

3.1 Introduction

This chapter outlines the methodologies developed and utilized within the RAMCIP project for the detection of people in the surrounding environment of the robotic platform, in order to be used in the robotic tasks related to the RAMCIP target use cases. The methodologies listed below concern the initial localization of humans within the scene, the extraction of their body silhouettes and the estimation of their body orientation in respect to the robotic platform, as these information are prerequisites for the person identification, body pose tracking and action recognition modules described in Task T3.2 of the RAMCIP project. All the possible scenarios were taken into consideration, regarding the movement and the relative positions of both the robotic platform and the human.

3.2 Background/foreground segmentation in static scene with moving foreground

The most common scenario encountered in our target use cases, includes having the robotic platform in a stationery position monitoring the human, as s/he moves within the FOV of the robot’s depth camera. In this case, background/foreground segmentation is achieved using NiTE’s built-in background removal algorithm (BGR). While the explicit details of NiTE’s BGR algorithm are not publicly available, the description below sufficiently approximates its functionality.

Changes detected in the depth values of successive frames are considered as potential foreground points and the corresponding pixels are clustered into potential foreground areas. Areas smaller than a threshold are discarded as sensor noise, while if the potential areas cover an extremely large portion of the image, the camera is considered non-stationary and the background/foreground segmentation process is aborted. The remaining potential foreground areas are then used as seeding points for a region growing algorithm which adds all the visited points to the foreground mask.

3.3 Background/foreground segmentation based on house map

NiTE’s built-in BGR algorithm implementation presents one major drawback, as it requires the human to move significantly in the first few frames in order to be detected. The initial movement prerequisite renders the NiTE BGR algorithm ineffective in a variety of scenarios where the human is initially stationary or her/his body movement is minor (i.e. seating on an armchair, eating on the dinner table etc.), as the whole scene, including any humans in the FOV, is considered as the stationary background. In order to overcome this limitation, a secondary background/foreground segmentation algorithm was developed, which could achieve efficient background/foreground segmentation even in completely static scenes.

The custom BGR algorithm developed by CERTH in the context of T3.2 takes as input the current robot pose from the robot localization module, and using the offline acquired 3D metric map of the current room, as stored within the RAMCIP hierarchical semantic mapping framework (see deliverable D3.1), it recreates the corresponding default background view. Next, the current view’s 3D point cloud is down sampled and fast ICP (~20 iterations) is utilized to align the current and the default background views, in order to compensate for any drift error produced by the localization module. Once the exact transformation matrix between the current robot view and the 3D metric map is calculated, it is used to quickly produce the default background view, as long as the robotic platform remains stationary.
Once the scene background for the current robot pose is recreated, foreground/background segmentation is achieved using Octree change detection. Any 3D points available in the current robot view, that do not appear in the default background view, are extracted as potential foreground points and clustered into potential foreground areas. Areas smaller than a threshold are discarded as sensor noise, leading to a final foreground which includes only large objects. Finally, human presence is verified by using an estimation mask which takes into consideration the overall size of the potential human blob, its height-to-width ratio and its position in regard to the floor.

![Figure 1: The CERTH BGR algorithm: Top left - current robot view, top right - reconstructed default background view, bottom left - octree change detection results, bottom right - final filtered foreground](image)

### 3.3.1 Fallen human detection

One special sub scenario of human detection is the detection of fallen humans, as described in Task T3.2 of the RAMCIP project. A fall event can be detected while the human is in the camera’s FOV and is being tracked by the human tracking module when s/he falls, or the human can be found already lying down, with the fall event having taken place outside of the camera’s FOV. In both cases the fall detection is handled by recognizing the human lying on the floor; even if the human was being tracked during the fall, the pose tracker will not be able to precisely estimate his/hers body pose due to the speed of a fall event.

Initially, the human-size blob on the floor is extracted using the NiTE or CERTH BGR algorithm, depending on whether the fall happened while the human was in the camera’s FOV or not. Next, in order to verify that the detected blob is indeed a human, the RAMCIP person identification pipeline is used (Section 4). Finally, if the identification process fails to identify the human, a poselets-based [14] human detector is employed leading to the final classification of the detected blob as either a fallen human or a fallen large object/obstacle.
3.4 Human detection in a single depth image

While BGR-based human detection methods can produce a detailed 2D/3D human silhouette within the scene, they are also susceptible to false positive detections. Often there are human-sized objects (i.e. a cupboard, or a large chair) in a scene which can be detected as humans, as they can get classified as foreground objects either due to temporary movement in consecutive frames (i.e. moving office chair) or due to permanent relocation (i.e. changing the position of a table). This drawback led to the development of a human detection algorithm that does not take into consideration the background/foreground segmentation information, but works regardless of the scene setup and is robust to different human postures and orientations with respect to the camera.

The first stage of the approach is similar to [81]. The ground plane with respect to the camera frame is assumed to be available from the robot localization module, and the human’s head should be inside the camera viewpoint. Initially, all 3D points are transformed to a coordinate frame attached to the ground with the X, Y axes parallel to it and the Z axis perpendicular. Points extremely close or far from the ground (e.g. more than the highest expected height of a human) are discarded thus effectively removing pixels that correspond to the floor or the ceiling. Then all points are projected on the X-Y plane and accumulated on a 2D grid i.e. an occupancy grid (grid cell size 5cm). The contribution of each point in the occupancy map is weighted inverse proportionally to their distance from the camera. The above process has the effect that tall vertical objects in the scene lead to local maxima in the occupancy map. To detect these maxima, we first discard grid cells with low occupancy e.g. corresponding to flat surfaces such as tables. Next, Gaussian filtering and connected component analysis are performed in-order to retain only clusters of suitable size (given prior knowledge of human figure as seen from above). Finally inside each cluster non-maxima suppression is performed. The highest 3D point inside each selected cell is used as a human head candidate.

The above procedure successfully detects humans in the scene but also finds several false positives corresponding to tall objects such as on the room walls. Therefore the second stage is to verify which of these candidates correspond to a human. Unlike the simple template approach used in [81], a more elaborated detector is employed, that can cope with arbitrary orientation of the head (both yaw and pitch rotations). A realistic 3D human body model is used and rendered from several viewpoints (about one thousand). The resulting images are cropped to form RGB and depth templates (size 96x96px) that depict the head and shoulders in various orientations. Next the LINEMOD detector [82] is trained using these templates. The detector is applied on an image window around the projection of each candidate head position on the RGB-D image plane. The window is obtained by first scaling the input image and then cropping a fixed size rectangle around the candidate point. The scaling factor is the ratio of the average depth of the rendered templates to the depth of the candidate point. The last step is important to achieve scale invariance. If the detector finds a match with a score greater than a threshold then the candidate is marked as a human. Since the detector is very fast and applied only to a limited region this step is computationally efficient.

The final step is the segmentation of the pixels corresponding to the human body from the rest of the scene. To achieve these, an approach inspired by [21] is used. Initially an undirected graph is built by connecting nearby 4-connected image points with edges weighted by their Euclidian distance. Connected component analysis performed on this graph and any small clusters are rejected. Next, nearby connected components of the graph are connected by “bridges”. In particular all pixels of one cluster that are 4-connected with a pixel from another cluster are visited and the shortest edge connecting the two clusters is computed. If the edge length is smaller than a threshold, the edge is added to the original
graph thus connecting the two clusters. This step is important to deal with self-occlusion of human body parts. Once the graph is constructed, region growing is performed starting from the “top of the head” pixel and all visited points are added to the human mask.

### 3.5 Orientation estimation

One of the roles of the human detection module is to provide input to the human pose tracking module (Section 5). Besides the human mask, a rough orientation of the human is also needed for achieving effective pose estimation, especially in the case of generative pose human pose trackers (Section 5.2), where the human template has to be initialized before proceeding with the tracking process.

While the estimation of the orientation of the human body is usually a by-product of 3D body pose estimation [19], the aim of the proposed method is to cope with challenging cases where such detectors fail i.e. when part of the body is occluded.

The orientation estimator is using as input the segmented depth image (i.e. human mask applied) and produces an estimate of the orientation around the vertical axis. This approach is similar to [19] which uses a random forest to classify each pixel in the image into one of 32 human body parts. The same features as in [19] are used, namely depth normalized differences of random pixels. The training dataset contains depth images depicting humans in various poses and orientations. The random forest is therefore trained with random pixels from randomly selected images and each pixel is annotated with the body part id and a corresponding orientation value. For efficiency, orientations are quantized to 32 bins covering the range from -120 to 120 degrees. The same classification objective as in [19] is used, but instead applied to the joint probability of parts and orientations efficiently represented by a flat histogram with 32 x 32 bins. During testing, random pixels $x$ from the input image $I$ are passed to each tree $t$ and eventually routed to a leaf node $l_t$. The resulting joint probability is obtained by averaging the joint probabilities estimated at each leaf of each tree:

$$p(c, o|x, I) = \sum_{t} p(c, o|x, l_t)$$

The orientation $o$ is then estimated by marginalizing over parts $c$, but only over parts that correspond to the torso. Including other body parts too resulted in much inferior results.

A challenging task is the creation of the training dataset since there is no such public dataset covering the wealth of human poses and orientations exhibited in common scenarios. To this end a publicly available motion capture subset from the well-known CMU dataset [83] is used. The dataset contains motion capture of several people performing various activities. About 200 of the sequences are hand-picked, excluding activities that are unlikely to happen in our scenarios such as ballet dancing or some sports. Then, an off-the-shelf software is used to rig a human skeleton for each sequence. The human body model and skeleton from the MakeHuman project [84] was selected. After this step, the corresponding 3D locations of the joints of the skeleton are obtained for each pose in each sequence. The human body model and skeleton from the MakeHuman project [84] was selected. After this step, the corresponding 3D locations of the joints of the skeleton are obtained for each pose in each sequence. Subsequently, the dataset is pruned by discarding similar poses using “furthest neighbor” clustering [85]. Distance between poses is computed as the average joint distance of associated skeletons. Finally, after this step 10000 unique poses are obtained and used to render images, also varying the orientation of the camera. Body part annotation of each image is achieved by creating a texture with a unique colour for each part which is mapped to the MakeHuman model. Thus, the rendered colour image will effectively provide the required annotation of each pixel. Three trees are trained by feeding 15 million random pixels to each tree. The maximum tree depth was 20.
3.6 Human detection during robot locomotion

The RAMCIP robot is expected to be able to detect humans within its surrounding environment. While the vision-based detection methods described above (Sections 3.2, 3.3 & 3.4) provide a detailed human mask, they are limited only to the areas of the scene that are within the FOV of the robot’s camera. In many cases, however, humans may be outside of the camera FOV (i.e. the robot has to move to another area in order to grab an object, or more than one people are within the area but they do not fit altogether inside the camera’s FOV). Moreover, their performance degrades when there are large changes between consecutive frames, especially when the robot is moving at relatively high speed. In order to overcome these limitations, a laser-scan based leg detector is utilized. While this approach does not produce detailed human detection information, it does provide a rough estimation of the humans’ position relative to the RAMCIP robot, thus allowing the execution of simple tasks that do not require high-level human detection information (i.e. following a human, route-planning etc.) or providing a reference point for the more complex vision-based human-detection methods described above.

The developed method uses the implementation from the “leg_detector” ROS package\(^1\). The input laser scans are provided by two laser scanners positioned opposite of each other near the base of the robot, creating a 360° slice around the robot, almost at floor level. The laser scans are clustered according to distance and a feature vector is extracted for each cluster, using the geometrical features listed in Table 1. Next, a random forest classifier is trained. Negative examples are obtained by moving the robot in an environment devoid of people, where all the detected clusters that meet a threshold are registered as negative train examples. Similarly, positive examples are obtained by setting up the laser in an open area with significant people traffic, with all the detected clusters that meet the threshold being registered as positive train examples. Finally, all the detected pairs of legs that produce a high probability score are considered as potential humans.

| Table 1: Geometric features used in the laser scan-based leg detector |
|-----------------|-----------------|-----------------|
| Number of points | Width | Length |
| Standard deviation | Average distance from median | Distance to adjacent cluster |
| Linearity | Circularity | Radius of best-fitting circle |
| Boundary length | Boundary regularity | Mean curvature |
| Mean angular difference | Inscribed angular variance | Distance from laser scanner |
| Occlusion (Boolean) | | |

When deployed in a realistic home environment, in contrast to a “sterile” laboratory room, the leg detector described above may produce many false positive detections, due to sensor noise, reflective surfaces or objects with structure similar to human legs (i.e. chairs, tables etc.). In order to increase the accuracy of the detector, a post processing step is added herein so as to remove any potential outliers. Specifically, any detected clusters that present one or more of the following attributes, are considered as outliers and are discarded:

- Low probability score
- No reliable detections nearby (legs are expected to come in pairs)

\(^1\) \url{http://wiki.ros.org/leg_detector}
- Existence for a very small period of time
- Overlap with objects and walls based on the global obstacle map which is a top-down projection of the 3D metric map

Finally, a Kalman filter is employed to track the position of each specific detection through subsequent frames.

### 3.7 The integrated RAMCIP human detection pipeline

The integrated RAMCIP human detection pipeline works as follows (Figure 2):

1. The laser-scan based approximate human detection is running continuously. If the human is detected in a limited space, for a long time period (e.g. \( t > 15 \text{s} \)), the robot turns its camera towards him/her in order to engage the vision based human detection implementations.
2. The NiTE human detection method is employed (Section 3.2). If there is a valid human detection, the human silhouette mask is published.
3. If NiTE fails to detect a human, the custom CERTH BGR (Section 3.3) and the single shot human detector (Section 3.4) are called in succession.
   a. If any of the two methods provides a valid detection, the silhouette mask extracted is published along with the body orientation.
   b. Otherwise the pipeline returns to step 1.

![Figure 2: The integrated RAMCIP human detection pipeline](image-url)
4. Person identification

4.1 Introduction

This chapter outlines the methodologies developed and utilized within the RAMCIP project for the identification of people in the surrounding environment of the robotic platform, in order to be used in the robotic tasks related to the RAMCIP target use cases. The methodologies listed below concern the initial identification of humans within the scene, using face-based identification techniques, as well as body and soft biometrics-based re-identification techniques, as described in Task 3.2 of the RAMCIP project.

4.2 Face-based person identification

The face classification pipeline addresses non-cooperative recognition of subjects from a closed set and identification of intruders i.e. subjects not previously seen. The non-cooperative factor means that the subjects will not be always at an ideal position in-front of the camera, but it is assumed that during the interaction with the robot an interval where the face of the subject is mostly frontal can be detected. For this reason the current approach is broken into several distinct steps which are performed continuously (i.e. in real time):

a. face detection
b. facial landmark detection
c. face alignment and cropping
d. face classification
e. temporal score fusion

An open source implementation of a state-of-the-art 2D frontal face detector [86] that returns one or more candidate face boxes, is used on the input image. Within the largest of these boxes facial landmark detection is performed, utilizing a publicly available landmark detector [87], which can withstand moderate facial yaw variation (about 20 degrees). Three landmarks (eyebrows and mouth centre) are selected, in order to perform affine rotation and centering of the image and then resize and crop it to a standard dimension (96x96px). The normalized face image is then fed to a CNN to obtain a 128-dimensional image descriptor. The public implementation of the triplet network [36] trained on the CASIA-Webface dataset [40] is used (the LFW measured accuracy of the net is about 92%) and a nearest neighbour classifier is employed, in order to find the best matching image in the enrolment dataset. Majority voting over a window of 10 image frames is used to obtain the best matching subject ID or declare an intruder in the case that the maximum score is lower than a threshold. During the enrollment phase several images of the user are recorded and processed using the above pipeline and a signature is stored for each of the user images.

Experiments were also performed using a 3D face matching algorithm but the resolution of depth images was not acceptable for the task. 3D information will be used in the final version of the system to better align faces i.e using a 3D face model.

4.3 Body-based person identification

The body-based identification implementation addresses the fast, real-time re-identification of previously identified subjects, from point of views where a close-up frontal view of the subject’s face is not available.

Based on the color and anthropometric features proposed in [44] and [61][63] respectively, a body-based identification feature set is employed, using 5 anthropometric features based on the subject’s skeleton, and 6 color features based on the subject’s clothes color in the torso and legs area (Table 2).
Table 2: List of features used for body-based person identification

<table>
<thead>
<tr>
<th>ANTHROPOMETRICS</th>
<th>COLOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Body Height</td>
<td>6. Hue in Torso</td>
</tr>
<tr>
<td>2. Shoulder Width</td>
<td>7. Saturation in Torso</td>
</tr>
<tr>
<td>3. Hip Width</td>
<td>8. Hue 2 in Torso</td>
</tr>
<tr>
<td>5. Leg / Torso Ratio</td>
<td>10. Hue in Legs</td>
</tr>
<tr>
<td></td>
<td>11. Saturation in Legs</td>
</tr>
</tbody>
</table>

Specifically the anthropometric features include:

- **Body height**: Measured as the distance between the knee joints and the head joint. The knee joints are selected instead of the ankle joints, because in many cases the subject may come close to the camera and her/his lower legs end up outside the camera’s FOV.
- **Shoulder width**: Measured as the distance between the left and right shoulder joints
- **Hip width**: Measured as the distance between the left and right hip joints
- **Upper leg length**: Measured as the average of the distances between the left/right knees and hip joints
- **Arm length**: Measured as the average of the distances between the left/right wrist, elbow and shoulder joints
- **Leg to torso ratio**: Measured as the ratio of the **Upper leg length** to the torso height (distance between the hips and neck joints)

The colour features include the two most dominant colours in the torso (to account for multicolored shirts) and the most dominant colour in the upper leg area (Figure 3), which are found by estimating each area’s colour histogram in the HSV colour space. The areas are estimated by back projecting to the image plane, the shoulder, hip and knee joints, while the HSV colour space is selected instead of the RGB colour space, as it less sensitive to illumination changes and allows more natural quantization into discrete colours. Moreover, only Hue and Saturation are used in the feature vector, as Value represents the total intensity and can be heavily affected by the ambient lightning.
Figure 3: Selected body areas for color-based feature extraction: Green - torso, Red: Legs. The arms and lower legs are not used because subjects often wear short-sleeved clothes or shorts, which may result in the skin color being most dominant under specific viewpoints.

For the classification process, an SVM classifier is employed, similarly to [61], trained using a single frame from each subject. The use of a single sample per subject along with the small feature vector size, allows for the fast retraining of the classifier when a new subject is initially identified and added to the identified persons list. This offers robust performance for relatively short time periods (i.e. for a few hours) where there are not large deviations in the appearance of the subjects (i.e. changes in clothes, lightning etc.), while retaining the real-time execution rate.

4.4 The integrated RAMCIP person identification pipeline

The integrated RAMCIP person identification pipeline works as follows (Figure 4):

1. The face DB is created offline during an enrolment phase, where the face signatures of known persons are extracted and stored in the DB.
2. The system detects a new human from the human detection module.
3. Body/colour-based identification is performed (Section 4.3)
   a. If the human is matched with any of the body/color models in the temporary human DB, update the body/colour model of the identified person, publish her/his ID and return to step 2.
4. Face-based identification is performed (Section 4.2)
   a. If the human is identified as one of the persons in the face DB, create a body/colour model for the identified person, add it to the temporary human DB, publish his ID and return to step 2.
   b. If the human is not identified, publish an alert for the presence of an unknown person and return to step 2.
Figure 4: The integrated RAMCIP person identification pipeline
5. Human pose tracking

5.1 Introduction
This chapter outlines the methodologies developed and utilized within the RAMCIP project for the efficient body pose tracking of humans inside the robotic platform’s camera FOV, in order to be used in the robotic tasks related to the RAMCIP target use cases. The methodologies listed below concern the initial estimation of the human body pose through a single-shot body pose estimator, as well as the tracking of the body pose in subsequent frames using an articulated human pose tracker, as described in Task T3.2 of the RAMCIP project. All the possible scenarios were taken into consideration, regarding the visibility of the human body parts, the variability in human body shape and size and the presence of noisy data in real home environments.

5.2 Single-shot pose estimation
Discriminative approaches, also known as single-shot pose estimators, use large training datasets and machine learning techniques in order to map the extracted features from the input data to body parts and poses. They offer body part recognition even from partial views, without the need of model fitting and explicit point correspondences. Their main advantage is that they do not require any prior knowledge about the scene, and thus can achieve human pose estimation using only a single frame.

While the main body pose tracking process in the RAMCIP project is carried out by an articulated tracker (Section 5.3), a single-shot pose estimator is still necessary in order to initialize and reset, in case of failure, the articulated tracker. Towards this end, two single-shot pose estimators were developed and tested in conditions similar to those described in the RAMCIP use cases: a) NiTE’s built-in human tracking implementation, which is based on [19], uses randomized decision trees and forests for body part detection and treats the body part segmentation as a per-pixel classification task. Each pixel in the depth image is evaluated separately and a weighted mean shift-based approach is used to estimate the body pose based on the inferred body part probability on each pixel; b) An optimized implementation based on metric regression forests [20] which also utilizes randomized decision forests, but uses an alternative training approach by employing the Metric Space Information Gain training objective. The metric regression forests-based approach was eventually discarded, as it failed to achieve an adequate processing speed (~2Hz). However, it may be used in the future, as it can be specifically trained and employed for tracker initialization in a small set of challenging views (i.e. side views, upper-body-only views etc.), in which the NiTE implementation fails to perform successfully, since it is trained mainly for full frontal views.

5.2.1 NiTE human pose estimation
The NiTE’s built-in human tracking module is based on a high-performance implementation of Shotton et al. [19]. It uses a single depth image as input, and extracts the 3D positions of 15 skeletal joints (Figure 5) in real time (30fps).
Figure 5: The skeleton joints estimated by the NiTE human tracking implementation

The algorithm uses a random forest to classify each pixel in the image into one of 32 human body parts (Figure 6), employing a simple depth comparison feature. At a given pixel $x$ the feature computes:

$$f_\theta(l, x) = d_i(x + \frac{u}{d_i(x)}) - d_i(x + \frac{v}{d_i(x)})$$

where $d_i(x)$ is the depth at pixel $x$ in image $l$, and parameters $\theta = (u, v)$ describe offset $u$ and $v$. The normalization of the offsets by $\frac{1}{d_i(x)}$ ensures the features are depth invariant: at a given point on the body, a fixed world space offset will result whether the pixel is close or far from the camera.

Next, the per-pixel body part information is pooled across pixels to generate reliable proposals for the positions of 3D skeletal joints, employing a local mode-finding approach based on mean shift with a weighted Gaussian kernel.

The final parameters of the random forest classifier are set as: 3 trees, 20 deep, 300,000 training images per tree and 2,000 training example pixels per image.

Figure 6: Body part segmentation of the human body, during pixel classification
5.3 Articulated full body model –based pose tracking

Generative approaches, also known as articulated body pose trackers, try to match the input data to articulated body templates by minimizing an objective function, utilizing various optimization techniques. These methods offer high detection rates, while also providing adequate accuracy even in cases of occlusions (which are encountered very often in real life conditions, in contrast to “ideal” laboratory setups) due to the use of a life-like human template. The main disadvantage of generative approaches is the initialization phase. In order for the initial pose estimation to be successful, the starting pose of the body template must be similar to the actual pose of the human, so that the optimization algorithm will be able to converge. Once the model converges to the user's starting pose, these methods proceed with the tracking of the consequent poses of the monitored subject, based on the prior knowledge and the new observations. As described in Section 5.2, this problem can be overcome by using a single-shot pose estimator. Moreover, if there are large body movements between consecutive observations, the optimization process can get trapped in local minima and fail to recover properly, thus making it necessary to achieve a high framerate in order ensure smooth body motion between frames.

Initially, an implementation based on [24] was tested, as the authors reported high accuracy rates and real-time (>20fps) processing speed using non-optimized CPU code. The authors in [24] use a Generalized Sum-of-Gaussians model for human shape modelling. The observed data are represented by isotropic Gaussians through octree partitioning, and multivariate Gaussian kernel correlation is employed as a similarity measure between the body template and the observation. However, the implemented articulated tracker managed to achieve the processing rate reported by the authors only by subsampling the input data and decreasing the number of iterations, which led to a significant drop in accuracy, and thus was rejected.

Next, a second articulated pose tracker was developed, based on DART [22]. Initial testing confirmed the accuracy levels reported by the authors, while through various code optimizations (implementation on CUDA) the implementation achieved real-time performance (~20fps). The DART-based CERTH articulated pose tracker, along with all the customizations and enhancements on the original tracker, is described in detail in the next section.

5.3.1 DART-based human pose tracking

The custom human pose tracker developed in T3.2 is based on an optimized implementation of Schmidt et al. [22], extending the DART approach towards a more effective and robust human tracking solution for the RAMCIP application scenarios. Our approach uses as input the depth image of a human silhouette, and it employs an articulated skinned human template to track the human pose, extracting the 3D positions of the skeletal joints. In order to ensure compatibility with the NITE pose estimator, since it is used to initialize and reset the articulated tracker, the joints estimated are the same as the joints presented in Figure 5.

5.3.1.1 Articulated human template

The articulated human template used in our approach is created using the MakeHuman [84] open source tool, and is customized in order to be suitable for use in the human tracker. The final template includes a skeleton composed of 10 rigid body parts connected to each other in a kinematic tree (Figure 7), along with 13000 vertices which represent the shape of the human body. The vertices are rigged on the skeleton, each controlled by up to 4 body parts, in order to realistically deform the skin mesh based on the skeleton movement.

The template pose is described through a pose vector \( \theta \) which includes the rotation of each body part relative to its parent, along with the global rotation and
translation of the template with reference to the camera frame. For a given pose \( \theta_0 \), the transformation matrix \( T_{i,0} \) of a body part \( i \) in relation to the camera frame can be defined recursively using:

\[
T_{i,0}(\theta_0) = T_{\text{parent}(i),0}(\theta_0)T_i(\theta_0)
\]

by composing the transforms in a chain, indexed using the \( \text{parent}(\cdot) \) function. For efficiency reasons and to avoid the problem of gimbal lock [88], the relative rotations of each body part are represented in the pose vector using unit quaternions instead of rotation matrices.

![Figure 7: The kinematic tree of the articulated human template, including the joints and the rigid body parts](image)

While, within the context of the RAMCIP project, a custom template is designed for each user, in order to make it fit exactly on her/his silhouette, two more factors are introduced to allow the dynamic manipulation of the size and shape of the template, in order to easily customize it when trying to track a non-registered human. Specifically, a global scale factor \( g \), which uniformly resizes the skeleton, to account for variations in height and limb length, and a width factor \( s \), which increases or decreases the distance between the skin vertices and the skeleton bones in order to fit on silhouettes of humans with variable body fat (Figure 8).
5.3.1.2 DART-based tracker

The main goal of the DART human pose tracker, as well as any articulated pose tracker in general, is to minimize the sum of the distances between the 3D points of an observed point cloud and the 3D points of the mesh of the human template. In traditional ICP algorithms, this requires the explicit nearest point computation for each of the input cloud points, which can become computationally expensive when using a detailed human template. However, based on the work of Fitzgibbon [89], it is possible to remove the need for this taxing process by using implicit surface representations in the form of precomputed signed distance functions, which replaces nearest point computation with a much faster lookup in the SDF. The DART tracker extends this approach by introducing an articulated model signed distance function $\text{SDF}_{\text{mod}}(x, \theta): R^3 \rightarrow R$, which defines for all of $R^3$ the distance to the surface of the model when articulated according to the pose vector $\theta$. Thus, the pose estimation process is achieved by estimating the pose vector $\hat{\theta}$ that minimizes the DART energy function:

$$\hat{\theta} = \arg\min_{\theta} \text{SDF}_{\text{model}}(\theta) = \arg\min_{\theta} \sum_{x \in \Omega} \text{SDF}_{\text{mod}}(x, \theta)$$

where $\Omega$ the 3D point cloud generated by back projecting the input depth frame.

Model representation

For each rigid body part $i$ of the human template, the 3D mesh points controlled by it form a geometry that is defined implicitly by a signed distance function $\text{SDF}^i(x, \theta): R^3 \rightarrow R$ [90], which takes on negative values inside the geometry, positive values outside, and has a zero value at the surface interface. This allows to approximate the global signed distance function $\text{SDF}_{\text{mod}}(x, \theta)$, as a composition of the precomputed local signed distance functions $\text{SDF}^i(x, \theta)$, which are not affected by the pose alterations, since the geometry they are defined from

---

Figure 8: Human template customization. Left: $[g=1, s=0]$, Right $[g=1.2, s=0.5]$
remains unaltered. As a result, this alleviates the need to recompute the full global SDF every time the pose is updated.

**Data association**

In order to calculate the value of the DART energy function $SDF_{model}(\theta)$ for a given pose $\theta$, the following steps are followed:

1. Transformation of the observed point cloud to body part space
   All the points of the input point cloud are transformed to the local coordinates system of each rigid body part.

2. SDF lookup
   For each point of the input point cloud, the distance of the closest point on the surface of each body part is estimated by trilinear interpolation of the precomputed SDF of each body part.

3. Body part assignment
   Based on the distances estimated in the previous step, each point of the input cloud is assigned to a body part. The sum of the distances of each point to its corresponding body part is the value of the energy function.

One of the main advantages of the described approach is that it is highly parallelizable, since every 3D point of the input point cloud can be processed independently, thus making it suitable for implementation on modern GPUs.

**Optimization**

For the estimation of the optimal pose, a Quasi-Newton method is employed, using the BFGS algorithm [91]. In Quasi-Newton methods, the Hessian matrix of second derivatives does not need to be evaluated directly. Instead, the Hessian matrix is approximated using rank-one updates specified by gradient evaluations.

The gradient values are calculated by estimating the data association error for each point, as described above, and analytically computing its first derivatives according to the model structure and current pose estimate. The resulting pose step $\Delta \theta$ is then composed to the pose estimate, and the algorithm iterates as needed until it converges to a final pose vector or reaches a predefined iteration limit.

**5.3.1.3 Complementary tracking features**

Besides the main tracking algorithm described above, a series of complementary tracking features are also introduced herein, in order to increase the tracker’s overall robustness and accuracy, as well as tackle specific problems that are encountered in the real home environment scenarios described in RAMCIP’s target use cases.

**2D reciprocal error**

The authors in [22] propose an “observation SDF term”, in order to avoid gross free space violation. In a similar fashion, our tracker also employs a reciprocal SDF term. However, instead of computing the 3D distance transform, as in the original DART tracker, the SDF of the 2D projection of the body template on the camera frame is used.

Specifically, the SDF of the current depth image is computed. Next, the deformed template is projected on the image, with each vertex of the template contributing towards a reciprocal SDF error, based on the value of the corresponding pixel on the SDF image. Finally, the energy minimization function takes the form:
\[ \theta = \arg\min_{\theta} (SDF_{\text{model}} + \lambda SDF_{\text{reciprocal}}) \]
where \( \lambda \) is a weighing factor \((0 < \lambda \leq 1)\).

**Outliers removal**

In ideal recording conditions (i.e. dataset capture in empty laboratory rooms) all the captured points correspond to the human body, and are consequently taken into consideration during the optimization process. However, in realistic home environments, similar to those that the RAMCIP robot is expected to work in, there are often noisy data (i.e. sensor noise, reflective surfaces, small objects near the human etc.) which degrade the tracker's performance when included in the optimization process. To counter this effect, a threshold \( D_{\text{thresh}} \) is introduced. Any observation points, whose distance from the corresponding template body part they are assigned to is larger than \( D_{\text{thresh}} \), are considered to be outliers and are discarded.

**Skeleton joints visibility**

Another problem encountered in realistic home environments is the occlusion of body parts due to the tracked human standing behind an obstacle or at the edge of the camera’s FOV. In this case only the visible body parts should be used in the optimization process, as the non-visible body parts may converge to the position of the closest visible body part.

To this end a pre-processing step is employed, calculating the visibility of each body part before moving to the optimization step. The last tracked skeleton pose is projected to the camera’s image plane. Any body parts that are outside the camera’s FOV or do not have any non-zero depth observations in their surrounding pixels, are considered non-visible. For the body parts without any children (lower arm, shin, head), visibility is determined by the position of the body part’s endpoint, while for the rest of the body parts the midpoint is used.
Non-visible body parts are not taken into consideration during the optimization process, by setting to zero the contribution of their corresponding 3D points to the energy function, thus keeping them in the last tracked pose, until a new valid depth observation appears.

Figure 10: Joint visibility correction in occluded view. Left: without correction, both legs converge to the visible left leg, Right: With correction, the right occluded leg remains at the default position and is not rendered. Moreover, the natural leg pose anchors the lower body on the floor plane, thus improving the fit of the upper body as well.

Leg crossing correction

During preliminary testing of the articulated tracker, it was noticed that there were instances where the tracker would misplace the legs of the human, mixing the left and right hips or shins, creating an unnatural pose, while the same behaviour was noticed in the NiTE pose estimator as well.

This mix-up is mainly caused by two factors: (a) noisy observations between the two legs (due to long distance from the camera, or baggy clothes) making it hard to separate the legs and (b) quick turn-arounds of the human which may create a local minimum which “traps” the optimization algorithm.

To counteract for this issue, a leg collision error is introduced in our pose tracking approach based on [92]. Each of the four leg body parts (hip R/L and shin R/L) are approximated by 7 spheres, $s$, equally spaced along the body parts, with centres $c_{s}$ on the body parts’ bone and radii $r_{s}$ small enough so that the spheres lie within the corresponding body geometry of each body part.

For a given pose $\theta$, intersection between spheres $s$ and $t$ occurs when

$$\|c_{s}(\theta) - c_{t}(\theta)\| < r_{s} + r_{t}$$

Thus the collision error is defined as

$$E_{\text{int}} = \sum_{(s,t) \in P} \frac{1}{1 + e^{-(r_{s} + r_{t} - \|c_{s}(\theta) - c_{t}(\theta)\|)/\gamma}}$$

where $P$ is the set of pairs of spheres.
Initially, the collision penalty was incorporated into the energy function. However, based on the fact that there were exactly two subcases to take into consideration (mix up between right/left hip or right/left shin; collision between the shin and the hip is not considered unnatural as it is regularly observed when someone is sitting), the leg crossing correction is now performed as a post processing step, after the optimization process.

Specifically, if the leg collision error for the current pose is found larger than a threshold, then it is recalculated for two more poses: a) right/left knees interchanged and b) right/left ankles interchanged. The pose that produces the minimum leg collision error is retained as the optimal pose.

![Figure 11: Leg collision fix](image)

### 5.3.1.4 Tracker initialization and failure detection

As noted above, in order for an articulate pose tracker to successfully track the human pose, the human template should be initialized from a pose close enough to the actual pose of the human. To this end, a two-step initialization process is employed in the RAMCIP human pose tracker.

The first step concerns the calculation of the global rotation and translation of the human template related to the camera frame. This is achieved by estimating the rigid transformation between the chest area of the template (formed by the chest and shoulders joints of the human template) and the corresponding area of the observed human (estimated either from the NiTE pose estimator (Section 5.2.1) or the body orientation estimator described in Section 3.5).

The second step concerns the initialization of each of the human template's body parts. If the human actual pose is close to the default template pose (standing, with the hands falling naturally by his side), then this step can be omitted. However, in more complex initial poses, it is essential to correctly initialize each body part, so as to allow the articulated tracker to converge to the correct pose, and not get trapped to a local minimum. Similarly to the optimization process of the DART tracker, the initial pose is estimated by minimizing, using the BFGS solver, an energy function which in this case is defined as the sum of the
distances between the corresponding joints of the template and the target skeleton provided by NiTE. The energy minimization function for the initialization process takes the form:

$$\theta = \arg\min_\theta \sum_i |j_t_i(\theta) - j_t_{i,\text{init}}|$$

where $j_t_i(\theta)$ and $j_t_{i,\text{init}}$ the 3D positions of the joint $i$ of the human template and the initialization skeleton respectively.

Figure 12: The initialization pipeline of the RAMCIP human pose tracker

5.3.1.5 Tracker failure detection

A human tracking failure metric is introduced in order to ensure that the tracker is properly reset if the optimization process converges to a wrong human pose. The tracking confidence is estimated by projecting on the camera image plane the deformed human template, and calculating its overlay ratio in respect to the observed human silhouette mask. If the calculated score falls below a confidence threshold (currently set at 0.75), the estimated pose is considered incorrect and the articulated tracker is re-initialized following the procedure described in Section 5.3.1.4. An example of a successful and failed pose estimation is presented in Figure 13.

Figure 13: Examples of successful (left) and failed (right) human pose tracking
5.4 The integrated RAMCIP human pose tracking pipeline

The integrated RAMCIP human pose tracking pipeline works as follows (Figure 14):

1. The system receives the new RGB and Depth frames from the imaging module.
2. Human detection (Section 3) is performed. If no humans are detected in the scene, return to step 1.
3. Person identification (Section 4) is performed. If the identified person is not the designated user, issue a corresponding signal so as for the rest of RAMCIP ROS nodes to become aware of this (e.g. ADM to be accordingly informed) and return to step 1.
4. The human skeleton from Nite is extracted (Section 5.2) and used to initialize the CERTH/DART tracker.
   a. If the Nite skeleton tracker fails to extract a human skeleton, use the CERTH body orientation estimator (Section 3.5) to initialize the CERTH/DART tracker. If this fails as well return to step 1.
5. Extract the human skeleton using the CERTH/DART tracker (Section 5.3).
   a. If the tracking succeeds, publish the extracted skeleton, go to steps 1 (data acquisition) & 2 (human detection) and then jump to step 5 to continue tracking the human skeleton.
   b. If the tracking fails, go to steps 1 (data acquisition) & 2 (human detection) and then jump to step 4 to re-initialize the tracker.

![Figure 14: The integrated RAMCIP human pose estimation pipeline](image-url)
6. Action Recognition

6.1 Introduction

This chapter outlines the methodologies developed and utilized within RAMCIP for the detection of human actions related to home objects, which will be employed for human activity recognition and monitoring as described in Task T3.2 of the RAMCIP project. The methodologies described below concern the detection of specific human postures and gestures which compose an action, as well as an extension of the action recognition framework in order to incorporate information of home object usage.

6.2 EigenJoints-based action recognition

The eigenJoints [73] algorithm focuses on extracting information about the relative positions of the joints between frames in a video sequence. It employs the calculation of three metrics; namely $F_{ci}$, $F_{cc}$ and $F_{cp}$, which are essentially pairwise differences of the joints in the current frame ($f_{cc}$), between the initial frame ($f_{ci}$) and with the previous frames ($f_{cp}$). The next step is to normalize the data in order to avoid domination of the attributes in higher numerical ranges. Next, in order to reduce redundancy and noise, Principal Components Analysis (PCA) is employed to obtain a compact representation of the concatenated $f_c = [f_{cc} f_{ci} f_{cp}]$ vector, by keeping the first 128 eigenvectors. The data processing pipeline mentioned above is depicted in Figure 15.

![Figure 15: The eigenJoints algorithm data processing pipeline [73]](image)

After the eigenJoints features are extracted, a Naive-Bayes-Nearest-Neighbor (NBNN) classifier is employed. The developed implementation exploits FLANN (Fast Library for Approximate Nearest Neighbors) and calculates Video-to-Class distance in order to output the most probable action given an input video sequence. In order to increase the discriminative potential of the algorithm, the original eigenJoints implementation can also be further customized by adding joint-selection for each action classifier, so as to use only a subset of the available joints (i.e. only hands and torso for “arm-waving”) when trying to detect a specific target action.

In our implementation, we further extended the original eigenJoints approach described in [73], by taking into account also the next frame of the video sequence instead of restricting the analysis only on the previous ones. In essence, we added a further feature, $f_{cn}$, which was analogous to the $f_{cp}$ feature, yet extracted from the next frame to the current one. This led to an improvement...
in the performance of the action recognition method, when tested on the CERTH action dataset, as described in Section 8.6 of the deliverable.

It can be noted at this point that an alternative action recognition framework was also implemented, based on the HOJ3D algorithm [75]. The HOJ3D method focuses on a histogram-based representation of the 3D human posture, where the 3D space is partitioned in $N$ bins using a modified spherical coordinate system, and a series of HMMs are trained to handle the recognition task. However, this approach failed to outperform the eigenJoints algorithm (Section 8.6) and as a result, the eigenJoints-based approach described above was implemented for the RAMCIP action recognition module.

### 6.2.1 Action recognition in realistic input streams

The aforementioned action recognition approach works on segmented sequences, i.e. short sequences containing only action-specific frames. On the other hand, the input video of the RAMCIP robot during its actual operation in the target application scenarios will be a continuous stream of frames, containing sequences of actions. In order to be able to recognize the action occurring at each time frame, a method to segment the input stream is required. Since the incoming video is provided as a stream, finding an end time for an action is not trivial to be robustly obtained, so the segmentation method focuses on finding a probable start time for an action, and using an average length for each action, based on the train set.

Given that user activities in need of monitoring in the RAMCIP target use cases include a strong relation to objects of the home environment (e.g. “drink water” involves interaction with a cup, “open the fridge” involves interaction with the fridge, etc.), the most effective way to detect the start of an action can be considered the detection of the event where the user starts interacting with a specific related object (e.g. grab a cup for a drinking activity). This way, by utilizing information from the object detection and tracking module (see Section 7.2), the start time for the majority of actions of interest is determined. As indicative examples with key relation to the RAMCIP target use cases, user interaction with the following objects denotes a probable action:

- Cup $\rightarrow$ drinking
- Pill box $\rightarrow$ take pill
- Cupboard/fridge $\rightarrow$ open/close cupboard/fridge

When input from the object detection and tracking module about an object being used is received, the action recognition module uses this time index as the possible start time of an action, and creates a segment from the incoming stream, with its length based on the average sequence length for the specific action from the training set. Alongside, the corresponding scene semantics derived from the object recognition module provide the necessary information so as to allow us infer the specific action that is being performed by the user within our hierarchical approach, which concerns the recognition of the core actions involved in the RAMCIP subUCs, by jointly considering detected body pose and gestures, in relation to the home objects.

In addition, in order to further increase the recognition robustness of key actions involved in the RAMCIP target use cases (e.g. water drinking, pill intake), geometric criteria can as well be applied on the joints of the user's tracked skeleton, so as for instance, to detect when the user's hand has approached her/his face after the detection of an object-related candidate activity start event, giving extra clues for the segmentation of e.g. a possible drinking or pill intake action.

At this point, it should be noted that the specific target action of pill intake, involved in the RAMCIP subUC3.1 (Taking medication/food supplement – reminders, bringing and monitoring), is highly similar in terms of user joint
movements to further user eating activities, and is thus detected in a similar way to those. However, within the RAMCIP action recognition approach, this activity is being differentiated through the object that is involved in it, i.e. the pills box. Given that the resolution of the current RGBD sensors, along with occlusions, do not allow a pill to be tracked during the monitoring of a user’s pill intake activity, we have to rely for the detection of pill intake actions on the further scene semantics. More specifically, during user monitoring, it is assumed that an eating activity, occurring in direct relation to the detected user’s manipulation of the pills box, corresponds to a pill intake activity. Notably, the fact that the user may manipulate a pills box and thereafter, perform a hand-to-mouth action, without however taking the pill, is something that can be misinterpreted also by a human caregiver as an actual pill intake activity; RAMCIP is no exception to this. In order however to counteract for such cases to the extent possible, and also for further false positives or false negatives in the pill intake action recognition, RAMCIP employs specific rules for the subUC3.1 to be established. While the specific approach followed for the overall establishment of the subUC3.1 and the robotic actions related to user monitoring will be described in detail in the RAMCIP deliverables D3.4 and D3.6, a brief overview of some key corresponding factors is provided below, for the shake of completeness.

The pills involved in the subUC3.1 during the RAMCIP robot’s operation, will be organized within specific pill boxes, where each pill box (a) will have a different colour from the other and (b) will contain all pills that need to be taken from the user at a specific time during the day. This way, upon the triggering of the specific subUC, only the exact pills that need to be taken by the user at that time point will be involved. Alongside, in order to counteract for false positive or false negative activity detections, the scenario of subUC3.1 will involve also a robot’s request to the user to further check whether the specific pills box is empty. Through the above, it will be guaranteed that the user will not be stimulated to take again pills that have already been taken during an undetected corresponding activity, while also, the possibility for pills left untaken will be further reduced.

Focusing again on the overall RAMCIP user activity monitoring framework, for further actions, which do not involve interaction with an object, motion rules applied on the user’s tracked skeleton joints are utilized, in order to detect a probable start time; for instance, no motion indicates the possible appearance of “idle” actions, change in skeleton height indicates stand2sit or sit2stand actions etc., again creating segments to be further processed based on the average sequence length from the training set.
7. Home Environment Monitoring

The main purpose of this section is to exhibit the advances of the algorithms described in D3.1 and to illustrate their functionality considering the human factor. Specifically, in this section the additional features that have been developed to make the home environment monitoring and modelling algorithms operable when the human user interacts with the objects are analytically described herein and assessed on dynamic environments.

7.1 Dynamic Update of the Hierarchical Semantic Map

The RAMCIP robot should be capable of performing fetching tasks of small objects (e.g. proactive bringing of bottle of water in subUC5.2) or to assist the user when a small object has fallen on the floor (e.g. subUC 4.3; assistance for fallen objects) in order to successfully fulfil the target Use Cases. To achieve this, continuous situation awareness –where possible- about the small objects that are involved in the user’s daily activities is mandatory.

Towards this direction, the hierarchical semantic mapping framework presented in the deliverable D3.1 has been enhanced in order to support dynamic updates about the new position of the currently detected small objects. More precisely when a small object e.g. cup is detected to be at a new place, e.g. Kitchen Table, the hierarchical map updates simultaneously the old “parent” LargeObject of the cup i.e. Cupboard to the new “parent” large object i.e. Kitchen Table. This functionality provides continuous situation awareness of the robot and enables it to fulfil successfully its fetching task by recalling from the hierarchical semantic map the last known positions of the small object of interest.

Alongside, by keeping track of the history of past detected positions of small objects, the robot is capable to look for an object at a series of possible locations derived from the past observations. More specifically, the hierarchical semantic map maintains a registry of past detected positions of each core object of interest (in respect to a large object, e.g. bottle of water found on the kitchen table, at the kitchen bench etc.), along with the frequency of those appearances. In case that an object is not detected at the position where it was last known to be, the robot, based on the above frequencies of past object detections, can perform further object recognition tries, at those most common object positions.

7.1.1 Change detection during home environment monitoring

An additional functionality of the hierarchical semantic map is to support the detection of changes on the positions of the misplaced objects towards the realization of the target subUC2.3. A change detection algorithm has been developed to detect small and large changes that correspond to objects that may block the user’s path, in accordance to the subUC2.3 needs, as defined in the deliverable D2.1.

This method extends the short term human motion intention modelling already described in D5.1 and is part of the human aware robot navigation as it takes into consideration the frequently visited human standing positions denoted in the hierarchical semantic map and the current human pose in the scene and predicts the human paths within the environment.

The developed change detection algorithm examines fragments on the most probable human path and determines notable changes that may belong to small or large objects. After subtracting the supporting surface i.e. the floor, the remaining clusters are assessed in a vertical range of [0, 0.4] utilizing the procedure described in Section 7.2, to detect already known small objects of large objects, i.e. large clusters on the examined region, which may belong to a large misplaced object and block the user’s path.
7.2 Monitoring of small objects during human activities

The small objects monitoring module aims to capture interactions between the human and the household objects, which are involved in the user’s daily activities. This module is deployed in collaboration with the human action recognition module and through the fusion of the extracted knowledge, the RAMCIP robot gains awareness of the performed human action, implying that the human actions are associated with the manipulated objects.

For the monitoring of the human manipulated small objects, it is essential to define a workspace which the robot observes constantly during the execution of the user’s activities. Small objects are expected to be found inside those workspaces. The processing of the 3D scene captured by the RAMCIP RGBD sensors is significantly boosted through information related to large scale objects that are extracted by the RAMCIP hierarchical semantic mapping approach (i.e. the hierarchical semantic map of the specific house environment).

Following the structure of the hierarchical semantic map, the large objects are tightly related to the small objects, providing on demand specific attributes to the method such as the supporting surface. More precisely, descending the spatial semantic hierarchy, for each room, there is a detailed description list of the large scale objects it contains, alongside with some general semantic information including the pose of the large object and the supporting surface. This semantic information is used to apply a coarse filtering of the input scene as a first step of our process.

Once the initial scene is filtered, RANSAC plane segmentation is applied to detect a more detailed supporting surface of the involved large scale objects, which will further constrain the monitored environment and expedite the small objects detection and tracking procedure. The Convex Hull of the points that belong to the supporting surface is the geometrical structure that defines the workspace. All points above the plane are projected onto it and the ones inside the convex hull are considered to belong inside the workspace. The following figure outlines the reference image as acquired by the RAMCIP robot and the detected workspace.

Figure 16: Calculation of the monitoring workspace, a) upper left: the RGB reference image, b) upper right: the retrieved supporting surface from the hierarchical semantic map, c) lower left: the detected plane and d) lower right: the workspace defined Convex Hull, with the remaining clusters of interest
The next processing step involves the segmentation of the workspace into separated clusters; each cluster corresponds to a small object. The RAMCIP object detection algorithm is executed for each of the extracted clusters and the best hypothesis, in terms of scene interpretation, is the object id assigned to the corresponding cluster. In case no known-object is detected in the area of a cluster then we consider the cluster as an “unknown” object. The following figure illustrates the sequential object recognition on each cluster.

![SCENE CLUSTER vs DETECTION RESULT](image)

**Figure 17:** Sequential parsing of calculated clusters and inference of the object detection and recognition algorithm.

The above pipeline is executed only when the monitoring procedure commences for the first time for a certain scenario. The RAMCIP robot keeps track of the changes in the object clusters thereafter and, therefore, is able to infer which object is currently being manipulated by the user. In order to be able to detect changes concerning only the objects in the scene, we have to discard any other changes captured by RACMIP sensors, namely the human movement. To this end, we draw upon the NiTE’s built-in background removal algorithm described in Section 3.2. The following figure illustrates the masked human from the scene.
The clusters initially present in the workspace define the first key frame. All new frames are compared to the current key frame. If correspondences between all the clusters of the current frame and the key frame are detected, then it is considered that no change is taking place in the workspace. Correspondences are checked in terms of the positions of the clusters and their dimensions. If a cluster of the key frame is not matched with any cluster for a certain period of time, then the object corresponding to that cluster is considered missing from the scene and therefore being manipulated by the user. The key frame is updated accordingly. The following figure illustrates the case of a person picking up a cup from the workspace.

Figure 18: Human excluded from the scene utilizing Nite’s background removal algorithm, i.e. scene analyser; removing user from background

On the other hand, if a new cluster appears in the workspace and does not match any cluster in the key frame, it is considered either that the user finished manipulating the object picked up previously or that a new object has entered in the workspace. The tracking of the last object left from the workspace is maintained. In case that the new object entered in the scene is not similar to the already tracked one in terms of shape and size, the object detection algorithm is executed for the new cluster. The following figure illustrates the case of a person placing back the cup to the monitored workspace.

Figure 19: User picking a cup from the monitored workspace
7.3 Detection of fallen objects during human activities monitoring

In the context of home environment monitoring during the target subUCs, the RAMCIP robot should be in position to detect a fallen or misplaced object, specifically towards the fulfilment of subUCs 4.3 and 2.3. In case an object of adequate size (given the corresponding restrictions of the RGBD sensors used) slips through the user’s hands and falls to the ground, the robot should be able to detect the fall event, determine the position of the fallen object and finally recognize it and detect its 6DOF pose in order to grasp it if possible, and bring it back to the user. Therefore, in order to increase situation awareness, the RAMCIP robot is at a corresponding monitoring state when the user performs specific activities such as cooking.

Similarly to the method described in Section 7.3, the first step comprises the detection of the dominant plane, where we rely on the assumption that the floor will be the planar surface perpendicular to the RGBD frame with the largest distance from it. As the workspace of this monitoring task, an area of 30cm height above the dominant detected plane, i.e. floor, is considered.

The first frame captured as soon as an activity commences is considered the reference frame. An approach following Octree-based Change Detection [93] is then applied to detect any changes in the workspace with respect to the workspace in the reference frame. In order to avoid taking into consideration changes in the 3D scene due to the user’s movement, the background removal methods described in Sections 3.2 and 3.3 are also utilized herein. Any new cluster above the floor that remains unchanged for a certain period of time is considered a candidate fallen object and initializes the object detection algorithm. This is based on the assumption that, once an object falls on the ground, it will eventually stabilize. In case that the fallen object is not a known to RAMCIP robot object, it is marked as “unknown” and the robot notifies the user for its existence. The figure below exhibits the sequence of monitoring for fallen object.
Figure 21: Sequence of snapshots from robot platform during monitoring for fallen object. The first row contains reference frames exhibiting a user with a
fallen cup. The second row of images corresponds to the detection of the user and his/her exclusion of the scene using the background/foreground segmentation human detection module. The third row of images describes the detected workspace along with the defined clusters. The last row presents the output of the detection algorithm operating over all the defined clusters.

At this point, it should be mentioned that the triggering policy that initiates the monitoring algorithms described in Section 7.2 and 7.3 is embedded within the higher level cognitive functions of the robot i.e. tightly connected with the Task 3.6. Therefore, this will be thoroughly described in D3.5, where details about the functionalities of the Assistance Decision Maker along with the robot task planner will be provided.

7.4 Object reconstruction methods and improvements

During the integration of the implemented modules for the home environment monitoring and modelling on the ROS environment, specific refinements have been conducted on the 3D reconstruction toolkit methods, which have already extensively described in D3.1. The main purpose of the 3D reconstruction toolkit is the rapid modelling of new objects when the robot is deployed in a human inhabited environment and some of the existing objects are not included in the hierarchical semantic map of the robot. Within the scope of the small objects 3D modelling, an enhancement of the method described initially in D3.1 (Section 3.3.3.3) has been conducted following the D3.1 delivery. Specifically, during the estimation of the object poses, a refinement step is performed by applying a global optimization step using the g2o methodology. Comparing to similar state-of-the-art methods [94] the qualitative results closely resemble each other. However, in terms of computational efficiency the method developed within the RAMCIP scope reveals better operational performance since it utilizes g2o global optimization step instead of the commonly used Levemberg-Marquard, where the convergence is slower and, thus, subsampling maybe required leading to a less accurate reconstructed model.

In this scope, it should be noted that following our corresponding further elaboration and testing efforts (experimental results can be found at Section 8.7.1 of the deliverable), the further advanced turntable-based object reconstruction method, originally developed within T3.1 and reported in D3.1, is now considered as the main method of the RAMCIP robot for the reconstruction of small household objects, which will be used for modelling such objects that will be involved in the RAMCIP trials, both at the LUM apartment and in the real end user homes of the ACE pilot site. As demonstrated through the experimental evaluation results of Section 8.7.1, the developed method is capable to lead into the rapid development of object models that can be subsequently used for effective object recognition and 6-DoF pose estimation through the corresponding RAMCIP methods described in D3.1.
8. Experimental Evaluation

8.1 Introduction
This chapter presents the experimental evaluation of the methods described in the previous section of the deliverable. As a starting point, Section 8.2 provides a detailed description of the public datasets that were used for the evaluation process, as well as of the datasets that were created within the RAMCIP project, while Sections 8.3 - 8.7 present qualitative and qualitative experimental results for each method.

8.2 Datasets

8.2.1 Public Datasets
This section lists the public datasets that were used for training and testing the algorithms described in the deliverable. The dataset selection was based on availability, data variability and frequency of use in the corresponding research fields.

8.2.1.1 CASIA WebFace Database
The CASIA WebFace Database [40][95] is a publicly available face dataset created by crawling and semi-automatically annotating face images from the IMDB website [96]. It contains in total 494,414 face images from 10,575 subjects with each subject having at least 15 images, and is the second largest face database, behind Facebook’s SFC dataset [97], which however is not currently publicly available.

8.2.1.2 EVAL Dataset
The EVAL Dataset [21][98] is a publicly available human pose tracking dataset, recorded using a Microsoft Kinect camera. It includes 3 subjects, with each one performing 8 sequences, and contains in total 10,207 depth frames.
Ground truth data for the human pose in each frame is recorded using the Vicon motion capture system, which captures the position of 14 joints using 33 markers located on the subject (Table 3).

Figure 22: Sample frames from the EVAL dataset
Table 3: Joints and markers tracked by the Vicon motion capture system, in the EVAL dataset

<table>
<thead>
<tr>
<th>JOINTS</th>
<th>MARKERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Chest</td>
<td>1</td>
</tr>
<tr>
<td>2. Head</td>
<td>2</td>
</tr>
<tr>
<td>3. Shoulder Left</td>
<td>2</td>
</tr>
<tr>
<td>4. Shoulder Right</td>
<td>2</td>
</tr>
<tr>
<td>5. Elbow Left</td>
<td>2</td>
</tr>
<tr>
<td>6. Elbow Right</td>
<td>2</td>
</tr>
<tr>
<td>7. Hand Left</td>
<td>3</td>
</tr>
<tr>
<td>8. Hand Right</td>
<td>3</td>
</tr>
<tr>
<td>9. Hip Left</td>
<td>1</td>
</tr>
<tr>
<td>10. Hip Right</td>
<td>1</td>
</tr>
<tr>
<td>11. Knee Left</td>
<td>3</td>
</tr>
<tr>
<td>12. Knee Right</td>
<td>3</td>
</tr>
<tr>
<td>13. Foot Left</td>
<td>4</td>
</tr>
<tr>
<td>14. Foot Right</td>
<td>4</td>
</tr>
</tbody>
</table>

8.2.1.3 MSR Action3D Dataset

The MSR Action3D Dataset [99][100] is a publicly available human action dataset, that contains sequences of 20 actions, listed in Table 4. Each action is performed three times by 7 subjects, with each subject facing the camera during the performance. The depth maps are captured at about 15 frames per second by a depth camera that acquires the depth through structure infra-red light. The size of the depth map is $640 \times 480$. Altogether, the dataset contains 23,797 frames of depth maps for 4020 action samples.

Table 4: Actions recorded in the MSR Action3D Dataset

| 1. High arm wave         | 11. Two hand wave |
| 2. Horizontal arm wave   | 12. Side boxing   |
| 3. Hammer                | 13. Bend          |
| 4. Hand catch            | 14. Forward kick  |
| 5. Forward punch         | 15. Side kick     |
| 7. Draw X                | 17. Tennis swing  |
| 8. Draw tick             | 18. Tennis serve  |
Figure 23: Sample sequences of the actions in the MSR Action3D dataset. Top - Draw tick, bottom - Tennis serve

8.2.2 CERTH Datasets

Besides the public datasets described above, two additional datasets were created within the premises of CERTH, in order to further test the developed methods in environmental conditions similar to those described in the target use cases of the RAMCIP project.

8.2.2.1 CERTH face dataset

The CERTH face dataset includes frontal upper body RGB images recorded with a Microsoft Kinect sensor, from a point of view similar to that of the robotic platform of the RAMCIP project, under various camera positions, head positions and lightning conditions.

The environmental/camera conditions taken into consideration are:

a. subject posture (standing / sitting)
b. camera height (low - 1.2m / high - 1.4m)
c. ambient lightning (bright / dark)
d. camera tilt (straight - 0° / tilted - 11°)

Based on the above settings, 14 recording scenarios are formed (Table 5), with each subject recording a sequence for each scenario where s/he moves her/his head into various poses. In total, 104,937 frames were recorded, from 13 subjects, with two of the subjects recording sequences both with and without glasses.
Table 5: Recording scenarios in the CERTH face dataset

<table>
<thead>
<tr>
<th>SCENARIO NUMBER</th>
<th>CAMERA HEIGHT</th>
<th>SUBJECT POSTURE</th>
<th>AMBIENT LIGHTNING</th>
<th>CAMERA TILT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Low</td>
<td>Sitting</td>
<td>Dark</td>
<td>Straight</td>
</tr>
<tr>
<td>2.</td>
<td>Low</td>
<td>Sitting</td>
<td>Dark</td>
<td>Tilted</td>
</tr>
<tr>
<td>3.</td>
<td>Low</td>
<td>Sitting</td>
<td>Bright</td>
<td>Straight</td>
</tr>
<tr>
<td>4.</td>
<td>Low</td>
<td>Sitting</td>
<td>Bright</td>
<td>Tilted</td>
</tr>
<tr>
<td>5.</td>
<td>Low</td>
<td>Standing</td>
<td>Dark</td>
<td>Straight</td>
</tr>
<tr>
<td>6.</td>
<td>Low</td>
<td>Standing</td>
<td>Dark</td>
<td>Tilted</td>
</tr>
<tr>
<td>7.</td>
<td>Low</td>
<td>Standing</td>
<td>Bright</td>
<td>Straight</td>
</tr>
<tr>
<td>8.</td>
<td>Low</td>
<td>Standing</td>
<td>Bright</td>
<td>Tilted</td>
</tr>
<tr>
<td>9.</td>
<td>High</td>
<td>Sitting</td>
<td>Dark</td>
<td>Straight</td>
</tr>
<tr>
<td>10.</td>
<td>High</td>
<td>Sitting</td>
<td>Bright</td>
<td>Straight</td>
</tr>
<tr>
<td>11.</td>
<td>High</td>
<td>Standing</td>
<td>Dark</td>
<td>Straight</td>
</tr>
<tr>
<td>12.</td>
<td>High</td>
<td>Standing</td>
<td>Dark</td>
<td>Tilted</td>
</tr>
<tr>
<td>13.</td>
<td>High</td>
<td>Standing</td>
<td>Bright</td>
<td>Straight</td>
</tr>
<tr>
<td>14.</td>
<td>High</td>
<td>Standing</td>
<td>Bright</td>
<td>Tilted</td>
</tr>
</tbody>
</table>

Figure 24: Sample frames from the CERTH face dataset
8.2.2.2 CERTH action dataset

The CERTH action dataset includes recordings of human actions encountered in the target use cases of the RAMCIP project, under conditions similar to those in real home environments.

The action sequences have been recorded using a Microsoft Kinect camera, which captures RGB and Depth frames, as well as skeleton tracking information for 15 joints from NiTE, at 30 fps. The actions include interactions of the subject with 3 large household objects (chair, cupboard, table), 3 small household objects (pillbox, cup, plate), random walking around the room and user idling. The action sequences are recorded from two different robot monitoring positions, as well as with and without occlusion, thus offering a camera POV similar to the POV that the robotic platform would have in a real home environment. Finally, 25 action scenarios are defined, along with a long recording where the subject is asked to perform most of the discrete actions in one continuous sequence (Table 6).

In total, 15 users were recorded, resulting in 350,000 captured RGB, Depth and Skeleton frames.

Table 6: List of action sequences recorded in the CERTH action dataset

<table>
<thead>
<tr>
<th>ACTION DESCRIPTION</th>
<th>CAMERA POSITION</th>
<th>REPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Standing to sitting</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2. Standing to sitting occluded</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3. Sitting to standing</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4. Sitting to standing occluded</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5. Random Walking</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6. Sitting idle</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7. Sitting idle occluded</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8. Standing idle</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>9. Drinking standing</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10. Drinking sitting</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>11. Drinking sitting occluded</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12. Handle pillbox sitting</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>13. Handle pillbox sitting occluded</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>14. Eating sitting</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15. Open cupboard door from mid handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>16. Open cupboard door from top handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>17. Open cupboard door from bottom handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>18. Reach cupboard door mid handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>19. Reach cupboard door top handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20. Reach cupboard door bottom handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>21. Retract hand from cupboard door mid handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>22. Retract hand from cupboard door top handle</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>23. Retract hand from cupboard door bottom handle</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
24. Reach cup on table sitting occluded  | 1  | 1  
25. Take pill sitting occluded       | 1  | 1  
26. Long multiple-action sequence    | 1  | 1  

| Figure 25: Sample sequences from the CERTH action dataset. a) 1 - Standing to sitting, b) 12 - Handle pillbox sitting, c) 14 - Eating sitting, d) 15 - Open cupboard door from mid handle |

### 8.2.3 LUM activities dataset

The LUM activities dataset was recorded in the premises of LUM, specifically in a 16.2m² single-room apartment, which included a sink, a table, an electric stove, cupboards and a chair (Figure 26). The subjects that took part in the dataset creation were elderly patients at LUM and were selected by the LUM personnel based on their mental and physical state and their ability to perform the required activities.

It should be noted that this dataset has been collected in the scope of the RAMCIP Task T3.4 (Modelling and monitoring user behaviour), and specifically, with the aim to obtain data that would facilitate the development of the T3.4 algorithms and methods towards complex activity and behaviour monitoring. As such, the specific dataset, as well as the results obtained through its use, will be reported in detail in the RAMCIP deliverable D3.4. Nevertheless, given the fact that the sequences recorded therein included participants of our target user group, in a realistic environment, performing activities closely related to the RAMCIP target...
use cases, they were also used so as to provide a further evaluation testbed for several of the T3.2 methods.

Figure 26: Rendering of the top-down view of the LUM room

During the recording process of the LUM activities dataset, each subject was asked to perform a series of everyday activity scenarios, for at least three repetitions, while being monitored by a Microsoft Kinect camera. The camera was positioned in predetermined monitoring spots, similar to the “parking positions” of the RAMCIP robot, and was recording RGB and Depth frames, as well as skeleton tracking information for 15 joints from NiTE, at 30 fps. As a result, the captured data provide a realistic depiction of the conditions that the RAMCIP robotic platform will encounter when it will be deployed in the LUM apartment, as well as in real home environments, and thus can be used to evaluate the algorithms developed for the RAMCIP project, under realistic conditions.

In total, 19 subjects took part in the dataset creation, amassing 886 sequences of 14 activity scenarios and resulting in 2,700,000 captured RGB, Depth and Skeleton frames. Table 7 provides a detailed description of the recorded activity scenarios, while some sample RGB and Depth frames are presented in Figure 27.

Table 7: Activity scenarios recorded in the LUM activities dataset

<table>
<thead>
<tr>
<th>1. DRINKING WATER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 1.1</strong></td>
</tr>
<tr>
<td>a. Subject is in the kitchen area.</td>
</tr>
<tr>
<td>b. Bottle is on the kitchen bench and the cup in the cupboard.</td>
</tr>
<tr>
<td>c. Subject opens the cupboard, gets the cup and closes the cupboard.</td>
</tr>
<tr>
<td>d. Subject gets the bottle and pours into the cup.</td>
</tr>
<tr>
<td>e. Subject closes the bottle.</td>
</tr>
<tr>
<td>f. Subject drinks water from the cup.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. HAVING A MEAL: PREPARATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenario 2.1</strong></td>
</tr>
<tr>
<td>a. Subject is in the kitchen area.</td>
</tr>
<tr>
<td>b. Knife and spoon are in the drawer, cup and dish in the cupboard, the pot is on the oven.</td>
</tr>
<tr>
<td>c. Subject opens the drawer, gets knife and spoon and closes the drawer.</td>
</tr>
<tr>
<td>d. Subject opens the cupboard, gets cup and dish and closes the cupboard.</td>
</tr>
</tbody>
</table>
Subject gets the pot and pours food in the dish.
Subject leaves the pot back on the bench.
Subject moves the rest of the objects to the table.

3. HAVING A MEAL: ACTUAL EATING PROCEDURE

Scenario 3.1
a. Subject moves towards the chair.
b. Subject sits down.
c. Subject starts eating and drinking.
d. Subject gets up from the chair.

4. HAVING A MEAL: COLLECTION OF OBJECTS

Scenario 4.1
a. Subject moves near the table.
b. Subject gets spoon, knife, dish and cup.
c. Subject moves towards the sink with utensils in hand.
d. Subject places them in the sink.

5. COOKING: PREPARATION

Scenario 5.1
a. Subject is in the kitchen area.
b. Subject opens the fridge.
c. Subject gets an object from the fridge.
d. Subject closes the fridge.
e. Subject moves the object towards the kitchen bench.

Scenario 5.2
a. Subject is in the kitchen area.
b. Subject opens the drawer.
c. Subject gets an object from the drawer.
d. Subject closes the drawer.
e. Subject moves the object towards the kitchen bench.

6. COOKING: CHOPPING

Scenario 6.1
a. Subject is in the kitchen area and the object to be chopped on the bench.
b. Subject moves towards the drawer.
c. Subject opens it and gets the knife, then closes the drawer.
d. Subject moves towards the ingredient of interest and starts chopping it.
e. Subject places the knife down, on the bench.

7. COOKING: STIRRING

Scenario 7.1
a. Subject is in the kitchen area and the object to be stirred on the bench (and in the pot).
b. Subject moves towards the drawer.
c. Subject opens it and gets the scoop, then closes the drawer.
d. Subject moves to the pot and starts stirring.
e. Then, the subject places the knife down, on the bench.
8. COOKING: CORE BAKING/BOILING PROCEDURE

**Scenario 8.1**

a. Subject is in the kitchen area.
b. The ingredient of interest is in the drawer (e.g. spaghetti box), the pot is in the cupboard.
c. Subject opens the drawer and gets the ingredient of interest.
d. Subject places the ingredient on the kitchen bench.
e. Subject closes the drawer.
f. Subject opens the cupboard and gets the pot.
g. Subject places the pot on the bench.
h. Subject closes the cupboard.
i. Subject gets the pot and places it under the sink.
j. Subject opens the tap of the faucet and then closes it.
k. Subject places the pot on the burner grate.
l. Subject turns on the burner grate.
m. Subject waits.
n. Subject turns off the burner grate.

9. PILL INTAKE: POUR WATER INTO CUP

**Scenario 9.1**

a. Subject is in the kitchen area.
b. The pillbox is in the drawer and the cup is in the cupboard and nearby a bottle of water.
c. Subject opens the cupboard, gets a cup and leaves it on the kitchen bench.
d. Subject closes the cupboard.
e. Subject opens the bottle and pours water into the cup.
f. Subject closes the bottle.
g. Subject opens the drawer and gets the pillbox, leaves the pillbox on the bench.
h. Subject closes the drawer.

10. PILL INTAKE: INTAKE THE PILL

**Scenario 10.1**

a. Subject sits on the chair near the table.
b. The pillbox and a cup of water are nearby, on the table.
c. Subject takes a pill and drinks water.

**Scenario 10.2**

a. Subject is standing in front of a drawer.
b. A cup of water is nearby.
c. Subject opens the drawer and takes the pillbox.
d. Subject takes the pill.
e. Subject drinks water.
f. Subject puts the pillbox back to its storage place (drawer).

11. PILL INTAKE: PLACE THE PILLBOX BACK TO STORAGE PLACE

**Scenario 11.1**
a. Subject sits on the chair near the table.
b. The pillbox is on the table.
c. Subject takes the pillbox and moves towards the kitchen area.
d. Subject opens the drawer and places the pillbox in it.
e. Subject closes the drawer.

**Scenario 11.2**

a. Subject is standing in front of a drawer.
b. The pillbox is on the bench.
c. Subject opens the drawer.
d. Subject gets the pillbox and places it in the drawer.
e. Subject closes the drawer.

---

**8.3 Human Detection**

The single shot human detection algorithm (Section 3.4) was evaluated using a randomly selected subset from the LUM activities dataset. Each sample frame was used without taking into consideration any temporal or other information, in order...
to evaluate the human detection capacity of the implemented algorithm. The ground truth human silhouettes were extracted offline using a semi-automatic BGR process and manual post-processing.

The detection confidence threshold was set to a high level, in order to completely eliminate any false human detections. While this may lead to a lower final detection rate, such behavior is preferable in real conditions, as the robot has to find only one valid detection among multiple successive frames; after an initial high confidence detection, the mask of the human silhouette in subsequent frames can be extracted using seeding points from the previous frame. The final detection rate achieved was 65.51%. The detector failed to successfully identify the human in the scene, mainly in three types of scenes, where the top down view of the upper-body deviated significantly from the template view (Figure 28):

1. The human was close to the edges of the FOV
2. The human was very close to a large object of increased height (i.e. large cupboard)
3. The human was in a low bending pose

However, the detector showed great detection capacity in scenes where the human was seated, partially occluded and relatively stationary (i.e. eating on a table) (Figure 29), which are scenes where the BGR algorithms fail to perform adequately, due to lack of movement and small foreground area. This demonstrates that the implemented single shot human detector can be efficiently used as a complementary human detection module, along with the BGR methods described in Section 3.

The human mask silhouette accuracy was also evaluated. The accuracy $A_C$ was defined as:

$$A_C = \frac{\text{pixels} \left( \text{overlap}(M_{gd}, M_{rec}) \right)}{\text{max} \left( \text{pixels}(M_{gd}), \text{pixels}(M_{rec}) \right)}$$

where $M_{gd}, M_{rec}$ are the ground truth and recreated human masks respectively, and $\text{pixels}()$ the number of pixels of the mask. The average mask accuracy was 83.69% as in most cases the whole human silhouette was successfully extracted.

Table 8: Experimental evaluation of the RAMCIP human detection and silhouette extraction implementation, on a random subset of the LUM activities dataset

<table>
<thead>
<tr>
<th>TOTAL SAMPLES</th>
<th>DETECTIONS</th>
<th>DETECTION RATE</th>
<th>AVERAGE MASK ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>4616</td>
<td>3042</td>
<td>65.61%</td>
<td>83.69%</td>
</tr>
</tbody>
</table>

Figure 28: Sample scenes where the single shot human detection algorithm failed
Figure 29: Sample scenes where the single shot human detection algorithm successfully detected the human and extracted the silhouette mask
8.3.1 Orientation Estimation

The RAMCIP body orientation estimation implementation (Section 3.5) was evaluated using a randomly selected subset of the CERTH action dataset (Section 8.2.2.2). Each frame was processed without taking into consideration any temporal information available. The ground truth body orientation was provided by using the right and left shoulder joints from the NiTE skeleton tracker, with each frame being manually checked in order to ensure the validity of the corresponding skeleton joints. The average deviation between the ground truth and the estimated body orientation was 7.47°, which falls within the limits of a single bin (the estimated values range from -120° to 120° degrees and are quantized in 32 bins each with a range of 7.5°). Moreover, 89% of the orientation estimations laid within the range of 2 bins (15°) (Figure 30), while in 2% of the samples the algorithm failed to provide a valid estimation.

Table 9: Experimental evaluation of the RAMCIP body orientation estimation implementation, on a randomly selected subset of the CERTH action dataset

<table>
<thead>
<tr>
<th>TOTAL SAMPLES</th>
<th>VALID ORIENTATION DETECTION</th>
<th>DETECTION RATE</th>
<th>AVERAGE ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>4451</td>
<td>4369</td>
<td>98.16%</td>
<td>7.47°</td>
</tr>
</tbody>
</table>

Figure 30: Percentage of samples for which the deviation from ground truth lies within the range of n bins

8.4 Person Identification

8.4.1 Face-based Identification

The face-based RAMCIP human identification implementation was evaluated using the CERTH face dataset (Section 8.2.2.1). The evaluation process tested both the face detection and human identification capacity of the developed algorithm.

Initially, a training set was created by selecting random frames for each of the 15 subjects. Table 10 presents the face detection results for all the frames selected for the training set. It should be noted that face detection performed very poorly...
in the case of subjects 14 and 15, as they were both very tall and had to be moved further away from the camera, in order to fit in the camera’s FOV for the defined scenarios. As a result, very few face detections were achieved for these specific subjects, mainly from frontal views with the subjects seated and they were removed from the testing process. Most of the non-detected images were replaced with images closer to a frontal view and the updated training set was used to train the classifier.

Table 10: Face detection rates for the initial randomly selected training set

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>TOTAL SAMPLES</th>
<th>FACE DETECTIONS</th>
<th>DETECTION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>365</td>
<td>286</td>
<td>78.36%</td>
</tr>
<tr>
<td>ID2</td>
<td>314</td>
<td>232</td>
<td>73.89%</td>
</tr>
<tr>
<td>ID3 (ID2 + glasses)</td>
<td>463</td>
<td>349</td>
<td>75.38%</td>
</tr>
<tr>
<td>ID4</td>
<td>523</td>
<td>300</td>
<td>57.36%</td>
</tr>
<tr>
<td>ID5</td>
<td>519</td>
<td>307</td>
<td>59.15%</td>
</tr>
<tr>
<td>ID6</td>
<td>390</td>
<td>175</td>
<td>44.87%</td>
</tr>
<tr>
<td>ID7 (ID6 + glasses)</td>
<td>468</td>
<td>238</td>
<td>50.85%</td>
</tr>
<tr>
<td>ID8</td>
<td>506</td>
<td>231</td>
<td>45.65%</td>
</tr>
<tr>
<td>ID9</td>
<td>537</td>
<td>194</td>
<td>36.13%</td>
</tr>
<tr>
<td>ID10</td>
<td>498</td>
<td>239</td>
<td>47.99%</td>
</tr>
<tr>
<td>ID11</td>
<td>464</td>
<td>332</td>
<td>71.55%</td>
</tr>
<tr>
<td>ID12</td>
<td>337</td>
<td>175</td>
<td>51.93%</td>
</tr>
<tr>
<td>ID13</td>
<td>428</td>
<td>270</td>
<td>63.08%</td>
</tr>
<tr>
<td>ID14</td>
<td>618</td>
<td>163</td>
<td>26.38%</td>
</tr>
<tr>
<td>ID15</td>
<td>562</td>
<td>45</td>
<td>8.01%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>6992</td>
<td>3536</td>
<td>50.57%</td>
</tr>
</tbody>
</table>

Next, a test set was generated by randomly selecting frames from each subject. In the case of more than one matches, the first successful match was selected. The detailed face detection and recognition rates per subject are presented in Table 11.

Table 11: Face detection and identification rates for the randomly selected test set

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>TOTAL SAMPLES</th>
<th>DETECTION SAMPLES</th>
<th>DETECTION RATE</th>
<th>IDENTIFICATION SAMPLES</th>
<th>RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>280</td>
<td>277</td>
<td>98.93%</td>
<td>145</td>
<td>52.35%</td>
</tr>
<tr>
<td>ID2</td>
<td>243</td>
<td>226</td>
<td>93.39%</td>
<td>140</td>
<td>61.95%</td>
</tr>
<tr>
<td>ID3</td>
<td>280</td>
<td>251</td>
<td>89.64%</td>
<td>143</td>
<td>56.97%</td>
</tr>
</tbody>
</table>
During testing, it was noticed that in a significant portion of the misclassified samples (~26% of the total identified samples), the classifier had matched the sample to the correct subject along with one or more wrong subjects. However, since the classifier selects the first successful match, it would misclassify the sample showing bias in favour of the subjects that were tested first.

To this end, a memory module was introduced, which would give a small boost to a match depending on the number of matches it received in the last few samples. This behaviour also corresponds to real life scenarios, where usually a small subset of the enrolled persons is present in the robot’s operational area and is identified during a relatively short time period. The use of the memory module increased the identification rate by over 20% (Table 12).

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>TOTAL SAMPLES</th>
<th>DETECTION</th>
<th>IDENTIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TOTAL</td>
<td>DETECTION</td>
<td>IDENTIFICATION</td>
</tr>
<tr>
<td></td>
<td>SAMPLES</td>
<td>RATE</td>
<td>SAMPLES</td>
</tr>
<tr>
<td>ID1</td>
<td>280</td>
<td>256</td>
<td>91.43%</td>
</tr>
<tr>
<td>ID2</td>
<td>243</td>
<td>207</td>
<td>85.19%</td>
</tr>
<tr>
<td>ID3</td>
<td>280</td>
<td>233</td>
<td>83.21%</td>
</tr>
<tr>
<td>ID4</td>
<td>290</td>
<td>242</td>
<td>83.45%</td>
</tr>
<tr>
<td>ID5</td>
<td>280</td>
<td>222</td>
<td>79.29%</td>
</tr>
<tr>
<td>ID6</td>
<td>280</td>
<td>177</td>
<td>63.21%</td>
</tr>
<tr>
<td>ID7</td>
<td>280</td>
<td>211</td>
<td>75.36%</td>
</tr>
<tr>
<td>ID8</td>
<td>280</td>
<td>184</td>
<td>65.71%</td>
</tr>
<tr>
<td>ID9</td>
<td>280</td>
<td>224</td>
<td>80.00%</td>
</tr>
<tr>
<td>ID10</td>
<td>280</td>
<td>230</td>
<td>82.14%</td>
</tr>
<tr>
<td>ID11</td>
<td>275</td>
<td>227</td>
<td>82.55%</td>
</tr>
<tr>
<td>ID12</td>
<td>280</td>
<td>218</td>
<td>77.86%</td>
</tr>
<tr>
<td>ID13</td>
<td>280</td>
<td>217</td>
<td>77.50%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3608</td>
<td>2848</td>
<td>78.94%</td>
</tr>
</tbody>
</table>
Finally, all the images in which face detection failed, were analysed in order to find a correlation between the recording conditions, as defined in the scenarios in Table 5, and the face detection rate. As it was expected, the majority of the no-face detection samples came from the scenarios 5 and 7 (human standing, camera low not tilted) where the human face appears on the upper limit of the camera’s FOV and its clarity can be affected by the large distance from the camera, the large deviation from the frontal view and the camera’s radial distortion which is more significant at the edges. The per recording scenario allocation of the no-face detection samples is presented in Table 13.

Table 13: Allocation of samples where face detection failed, per recording scenario

<table>
<thead>
<tr>
<th>RECORDING SCENARIO</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Low / Sitting / Dark / Straight</td>
<td>4.31%</td>
</tr>
<tr>
<td>2. Low / Sitting / Dark / Tilted</td>
<td>3.50%</td>
</tr>
<tr>
<td>3. Low / Sitting / Bright / Straight</td>
<td>2.02%</td>
</tr>
<tr>
<td>4. Low / Sitting / Bright / Tilted</td>
<td>3.50%</td>
</tr>
<tr>
<td>5. Low / Standing / Dark / Straight</td>
<td>30.32%</td>
</tr>
<tr>
<td>6. Low / Standing / Dark / Tilted</td>
<td>3.23%</td>
</tr>
<tr>
<td>7. Low / Standing / Bright / Straight</td>
<td>31.40%</td>
</tr>
<tr>
<td>8. Low / Standing / Bright / Tilted</td>
<td>5.80%</td>
</tr>
<tr>
<td>9. High / Sitting / Dark / Straight</td>
<td>4.85%</td>
</tr>
<tr>
<td>10. High / Sitting / Bright / Straight</td>
<td>0.00%</td>
</tr>
<tr>
<td>11. High / Standing / Dark / Straight</td>
<td>4.31%</td>
</tr>
<tr>
<td>12. High / Standing / Dark / Tilted</td>
<td>0.67%</td>
</tr>
<tr>
<td>13. High / Standing / Bright / Straight</td>
<td>2.96%</td>
</tr>
<tr>
<td>14. High / Standing / Bright / Tilted</td>
<td>3.10%</td>
</tr>
</tbody>
</table>

8.4.2 Body-based identification

The body-based identification implementation was evaluated using a randomly selected subset of 6 subjects from the CERTH action dataset. The number of subjects was intentionally selected to be small, since in realistic environment conditions, the robot will usually have a small number of people in its active workspace. Moreover, the small number of subjects allows for quick retraining of the classifier when a newly identified person appears.

The NiTE pose estimator was used to provide the human silhouette mask and the skeleton tracking information, with all the selected frames being manually checked to ensure the validity of the data.

Initially, the anthropometric and color-based features were evaluated separately. Both the anthropometric and color features achieved an average classification accuracy when used separately as can be seen in Table 14. However, when combined, the two feature sets reached an accuracy rate of 92.66%. Table 15 presents the detection rate for each subject.
Table 14: Identification accuracy for each feature set, on 6-subject subset of the CERTH action dataset

<table>
<thead>
<tr>
<th>ONLY ANTHROPOMETRICS</th>
<th>ONLY COLOR</th>
<th>BOTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.77%</td>
<td>68.67%</td>
<td>92.66%</td>
</tr>
</tbody>
</table>

Table 15: Identification accuracy for each subject, on 6-subject subset of the CERTH action dataset

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>IDENTIFICATION ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID_1</td>
<td>82.03%</td>
</tr>
<tr>
<td>ID_2</td>
<td>96.99%</td>
</tr>
<tr>
<td>ID_3</td>
<td>82.96%</td>
</tr>
<tr>
<td>ID_4</td>
<td>98.39%</td>
</tr>
<tr>
<td>ID_5</td>
<td>99.11%</td>
</tr>
<tr>
<td>ID_6</td>
<td>95.65%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>92.66%</td>
</tr>
</tbody>
</table>
Figure 31: Sample frames from the human body-based identification evaluation on the CERTH action-dataset: Left-training template, right-identified frames

8.5 Human pose tracking

8.5.1 DART-based human pose tracking

The DART-based RAMCIP human pose tracking method was first evaluated using all the sequences of the EVAL dataset (Section 8.2.1.2). By following the evaluation processes typically used in the relevant literature [18]-[28], the metrics used herein were:

a. Average distance error: Average distance between the ground truth and the estimated joint positions
b. Joint accuracy at 0.1m: Percentage of body joints where the distance between the ground truth and the estimated joint position is less than a max allowed distance

A generic human body template was used, manually customized to match the body shape of the actors in the sequences. The evaluation process was run both with and without the RCP error correction, resulting in an average distance error of 0.045m and average accuracy of 91.38% at 0.1m, which are in line with the results reported by the authors in [22]; notably, average accuracy results surpassing the 90% level have been obtained through our proposed approach. The detailed results for each sequence of the dataset are presented in Table 16 below.

<table>
<thead>
<tr>
<th>SEQUENCE</th>
<th>NO RCP ERROR CORRECTION</th>
<th>WITH RCP ERROR CORRECTION, (\lambda=0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVERAGE DISTANCE ERROR (m)</td>
<td>JOINT ACCURACY, 0.1m</td>
</tr>
<tr>
<td>a_0</td>
<td>0.038</td>
<td>97.45%</td>
</tr>
<tr>
<td>a_1</td>
<td>0.042</td>
<td>94.53%</td>
</tr>
<tr>
<td>a_2</td>
<td>0.044</td>
<td>93.54%</td>
</tr>
<tr>
<td>a_3</td>
<td>0.047</td>
<td>92.37%</td>
</tr>
<tr>
<td>a_4</td>
<td>0.050</td>
<td>90.33%</td>
</tr>
<tr>
<td>a_5</td>
<td>0.043</td>
<td>92.27%</td>
</tr>
<tr>
<td>a_6</td>
<td>0.047</td>
<td>91.53%</td>
</tr>
<tr>
<td>a_7</td>
<td>0.048</td>
<td>90.03%</td>
</tr>
<tr>
<td>b_0</td>
<td>0.038</td>
<td>96.03%</td>
</tr>
<tr>
<td>b_1</td>
<td>0.049</td>
<td>86.30%</td>
</tr>
<tr>
<td>b_2</td>
<td>0.041</td>
<td>94.14%</td>
</tr>
<tr>
<td>b_3</td>
<td>0.049</td>
<td>90.21%</td>
</tr>
<tr>
<td>b_4</td>
<td>0.051</td>
<td>89.52%</td>
</tr>
<tr>
<td>b_5</td>
<td>0.044</td>
<td>91.13%</td>
</tr>
<tr>
<td>b_6</td>
<td>0.046</td>
<td>91.99%</td>
</tr>
<tr>
<td>b_7</td>
<td>0.053</td>
<td>86.42%</td>
</tr>
<tr>
<td>c_0</td>
<td>0.039</td>
<td>95.38%</td>
</tr>
<tr>
<td>c_1</td>
<td>0.037</td>
<td>95.67%</td>
</tr>
<tr>
<td>c_2</td>
<td>0.046</td>
<td>91.47%</td>
</tr>
<tr>
<td>c_3</td>
<td>0.050</td>
<td>88.67%</td>
</tr>
<tr>
<td>c_4</td>
<td>0.042</td>
<td>93.07%</td>
</tr>
<tr>
<td>c_5</td>
<td>0.053</td>
<td>82.16%</td>
</tr>
<tr>
<td>c_6</td>
<td>0.045</td>
<td>91.60%</td>
</tr>
<tr>
<td>c_7</td>
<td>0.065</td>
<td>82.07%</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.046</td>
<td>91.16%</td>
</tr>
</tbody>
</table>

**Figure 32: Sample frames from the human pose tracking evaluation on the EVAL dataset**

In order to test the performance of the DART-based RAMCIP human pose tracking implementation in more realistic conditions, closer to the target RAMCIP use cases, as well as to compare its performance to the NiTE pose estimator, both trackers were evaluated on the CERTH action dataset and the LUM activities dataset. Since there are no ground truth data (e.g. VICON marker measurements) available for these datasets, only qualitative results could be obtained (Figure 33). Preliminary analysis of the results showed that the DART/CERTH implementation offers better performance in sitting scenarios and near the limits of the camera’s FOV, as well as more robust leg tracking when the tracked human is moving. As further shown in the indicative examples of Figure 34, the RAMCIP tracker was found to significantly outperform the NiTE estimator in a series of cases strongly related to the human activity monitoring needs of the RAMCIP target use cases.

However, both implementations failed to successfully handle scenes where the tracked human interacted with large objects (i.e. cupboard doors), as in this case the large object is added to the foreground area, and both algorithms try to incorporate it in the pose estimation process. In the specific application scenarios of RAMCIP, and given the hierarchical semantic map built for the robot’s operational environment, along with the large objects state tracking module implemented in T3.1, this problem can be overcome by including information about the location of large articulated objects in the monitored area (Section 7), and thus removing them from the foreground area before proceeding to the pose estimation step. An example of such a scene is presented in Figure 35.
Figure 33: Sample frames of successful human pose tracking in the CERTH action dataset. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NiTE estimator.
Figure 34: Sample frames of tracking in realistic conditions where the CERTH tracker outperforms the NiTE estimator. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NiTE estimator. a) Lower body occlusion while standing, b) Turn around, c) Leg
crossing, d) Sitting, e) Lower body occlusion while seated, f) heavy body occlusion near the edges of the FOV

![Sample frames of failed human pose tracking in the CERTH action dataset, due to the human's interaction with a large object. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NITE estimator](image)

**Figure 35:** Sample frames of failed human pose tracking in the CERTH action dataset, due to the human's interaction with a large object. Left - input scene, red - skeleton produced by the CERTH/DART tracker, green - skeleton produced by the NITE estimator

### 8.6 Action Recognition

The eigenJoints-based RAMCIP action recognition implementation was evaluated using the MSR Action3D dataset (Section 8.2.1.3) and the CERTH action dataset (Section 8.2.2.2), while preliminary experimentation over sequences of the LUM activities dataset was also performed.

In the evaluation on the MSR Action3D dataset, first, all 20 classes were included in our analysis, in contrast to the original eigenJoints study of [73], where the test set was dissected into 3 action sets. Hence the numerical difference in the corresponding results described below; the original work achieved an average recognition rate of 96% on non-cross subject testing and 82% on cross-subject testing, in contrast to our implementation of the original eigenJoints algorithm, which achieved 94% in non-cross subject and 70% in cross subject testing. Table 17 and Table 18 present the detailed results of the evaluation process for non-cross subject and cross subject testing respectively.

<table>
<thead>
<tr>
<th>EXPERIMENT DESCRIPTION</th>
<th>TOTAL TEST SAMPLES</th>
<th>CORRECT DETECTIONS</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leave-out all e_01 sessions for all subjects</td>
<td>187</td>
<td>170</td>
<td>90.90%</td>
</tr>
<tr>
<td>Leave-out all e_02 sessions for all subjects</td>
<td>191</td>
<td>182</td>
<td>95.28%</td>
</tr>
<tr>
<td>Leave-out all e_03 sessions for all subjects</td>
<td>185</td>
<td>178</td>
<td>96.21%</td>
</tr>
<tr>
<td>EXPERIMENT DESCRIPTION</td>
<td>TOTAL TEST SAMPLES</td>
<td>CORRECT DETECTIONS</td>
<td>PCTG</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>--------------------</td>
<td>--------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Leave-out all S10 subject related actions</td>
<td>60</td>
<td>46</td>
<td>76.66%</td>
</tr>
<tr>
<td>Leave-out all S09 subject related actions</td>
<td>60</td>
<td>54</td>
<td>90.00%</td>
</tr>
<tr>
<td>Leave-out all S08 subject related actions</td>
<td>60</td>
<td>48</td>
<td>80.00%</td>
</tr>
<tr>
<td>Leave-out all S07 subject related actions</td>
<td>59</td>
<td>49</td>
<td>83.05%</td>
</tr>
<tr>
<td>Leave-out all S06 subject related actions</td>
<td>58</td>
<td>47</td>
<td>81.03%</td>
</tr>
<tr>
<td>Leave-out all S05 subject related actions</td>
<td>57</td>
<td>37</td>
<td>64.91%</td>
</tr>
<tr>
<td>Leave-out all S04 subject related actions</td>
<td>38</td>
<td>31</td>
<td>81.57%</td>
</tr>
<tr>
<td>Leave-out all S03 subject related actions</td>
<td>56</td>
<td>22</td>
<td>39.28%</td>
</tr>
<tr>
<td>Leave-out all S02 subject related actions</td>
<td>59</td>
<td>31</td>
<td>52.54%</td>
</tr>
<tr>
<td>Leave-out all S01 subject related actions</td>
<td>60</td>
<td>45</td>
<td>75.00%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>567</strong></td>
<td><strong>410</strong></td>
<td><strong>72.31%</strong></td>
</tr>
<tr>
<td>Leave-out all S08-09-10 subject related actions</td>
<td>180</td>
<td>138</td>
<td>76.66%</td>
</tr>
</tbody>
</table>
Further to the above, and in order to obtain a more direct comparison so as to better understand the improvements introduced from our modified eigenJoints-based low-level action recognition approach described in Section 6.2, we performed the exact same experiments as reported in [73]. Specifically, we focused on cross-subject testing, over the Action Sets AS1, AS2 and AS3. The obtained results through the original method and our proposed extended implementation are summarized in Table 19 below.

Table 19: Comparison between the original eigenJoints approach and the RAMCIP extension

<table>
<thead>
<tr>
<th></th>
<th>Original eigenJoints method implementation [73]</th>
<th>RAMCIP eigenJoints based action recognition approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1</td>
<td>74/106 (69.81%)</td>
<td>86/106 (81.13%)</td>
</tr>
<tr>
<td>AS2</td>
<td>71/113 (62.83%)</td>
<td>90/113 (79.65%)</td>
</tr>
<tr>
<td>AS3</td>
<td>101/112 (90.18%)</td>
<td>108/112 (96.43%)</td>
</tr>
</tbody>
</table>

In addition to the above, the evaluation of our eigenJoints-based low-level action recognition method on the CERTH action dataset yielded the following results:

Table 20: Experimental evaluation of the RAMCIP action recognition pipeline on the CERTH action dataset

<table>
<thead>
<tr>
<th>TOTAL SEQUENCES</th>
<th>TRAIN SET</th>
<th>TEST SET</th>
<th>CORRECT DETECTIONS</th>
<th>DETECTION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>789</td>
<td>591</td>
<td>198</td>
<td>151</td>
<td>76.26%</td>
</tr>
</tbody>
</table>

As shown in the confusion matrix presented in Figure 36 below, a large number of low-level action classes were incorporated in this analysis, in order to better understand the limits of the developed eigenJoints-based method for realistic applications, when trying to monitor user activities at a highly detailed level through a service robot. For a series of actions, perfect recognition rate was obtained, even without incorporating higher-level, object information, while several low-level actions were mixed with ones that involved rather similar skeleton joints movements. By applying our overall hierarchical activity monitoring approach, incorporating further information on the tracked manipulated objects and geometric criteria of the user's joints (as described in Section 6.2.1), and by restricting the analysis on key actions of interest involved in the RAMCIP target subUCs (e.g. eating, drinking, pill intake, open the fridge or cupboard) for complex user activities monitoring and inferencing on robot assistive decisions, the overall detection rate is further improved, towards robust user activity monitoring even in the RAMCIP realistic settings.
We performed preliminary evaluation experiments of our overall action recognition approach on the LUM activities dataset. In this scope, our aim was to obtain a preliminary assessment on key actions involved in the RAMCIP target use cases, over a dataset collected from a population of our target end user group, in the same realistic environment where the RAMCIP robot will operate during both the preliminary trials with end users and the final pilot trials.

Specifically, we manually annotated the existence of target actions “open the fridge”, “close the fridge”, “eating” and “drinking” within the data recorded from a random subjects pool of 6 persons, as these formulate key actions towards robot decisions in the scope of a series of the target subUCs (e.g. subUC3.1, subUC4.2, subUC5.2). Following the annotation process, we ended up with a test set of 24 action instances in total. For this experiment, the train set used in our action recognition method implementation, comprised corresponding action instances obtained from the LUM activities dataset, after excluding the LUM dataset test sequences mentioned above. Testing was thus performed on a cross-subject basis.

As concerns the eating and drinking activities, a perfect detection rate (100%) was obtained for both of them on the test set, without false positives. Despite the realistic nature of the dataset, including occlusions and sensor noise, the RAMCIP human pose tracker (Section 5.4) was capable to monitor all users’ upper body

Figure 36: Confusion matrix of the RAMCIP action recognition evaluation on the CERTH action dataset
pose robustly enough, while they were sitting on the table of the LUM apartment and performed eating and drinking activities. In addition, the objects monitoring framework managed to correctly detect all cases of objects of interest involved in these activities and their manipulation. As a result, our overall action recognition approach detected correctly all eating and drinking action instances.

As concerns the fridge open/close actions, not all test instances of our preliminary testbed were correctly detected. Specifically, although a perfect (100%) recognition rate was obtained for the “open fridge” action, 50% accuracy was achieved for the detection of “close fridge” actions. The difference in these results was mainly due to self-occlusions that appeared during the close fridge action, given the specific camera pose in the dataset collection space. Moreover, errors in human tracking during the user’s interaction with large environment objects, such as the fridge (as described in the previous section), can play a further negative role.

Given the importance of the specific actions in the RAMCIP target subUCs (specifically subUC4.2, where the robot detects the fridge left open), the approach to further increase the robustness of the corresponding actions recognition from the integrated robot in its realistic applications, concerns the incorporation of further object-related information, derived from the known large object positions in the house hierarchical map and the RAMCIP large objects state tracking module. At a first level, this information can help towards better human pose tracking results, by discarding large object segments initially included in the user’s extracted foreground silhouette. Moreover, upon the detection of a candidate event for a fridge open action (related for e.g. to the presence of user’s skeleton joints in the door handle’s vicinity), our large objects state tracking module can be run (once the target object is detected not to be masked from the user’s silhouette), so as for the robot to be further facilitated while inferencing whether the fridge is indeed open or closed, towards the triggering of the subUC4.2-related robot assistive actions.

More thorough assessment of the performance of the overall RAMCIP activity recognition framework on the LUM activities dataset, jointly focusing both on low-level and high-level complex activities monitoring in the scope of the needs of the RAMCIP target subUCs, will be provided in the RAMCIP deliverable D3.4.

8.7 Home environment monitoring

8.7.1 Evaluation of 3D object reconstruction method

The elaborated object 3D reconstruction method for the modelling of small objects has been evaluated on the real objects that are utilized within the RAMCIP use cases. Five different objects has been reconstructed using the turntable i.e. pillbox, cup, bottle of water, spaghetti box and pot. The queried objects have been reconstructed accurately using the developed method while the quality of the produced models has been assessed by utilizing in the model based object detection and recognition component. The following figure graphically illustrates the outcome of the reconstruction method and appends as ground-truth the manually created models using the Blender software. By observing the visual outcome of the reconstruction it is summarized that the method is adequate to create consistent 3D models that closely resemble the manually created models using the Blender SW. However, the impact of the minor differences among the rendered and the reconstructed models should be quantitative measured.
In order to further examine the accuracy of the 3D reconstructed models, in terms of capacity to lead into accurate object recognition and 6-DoF pose estimation, an additional experiment has been conducted. Specifically, the above-illustrated objects were placed on various, yet specific distances from the RGBD camera in accordance with the SubUC where each object is involved. In each different spot, the poses of the objects have been manually measured and retained as ground truth. Then, the object detection method has been firstly trained with the manually created models of the objects using the Blender SW and was then trained with the models of the objects that have been created using the RAMCIP turntable-based 3D reconstruction method. The aim of this experiment is to prove that the reconstructed models are adequate to be utilized for the object detection and recognition module. Therefore, during the evaluation phase, the object detection algorithm has been assessed over the manually created models and the 3D reconstructed ones and evaluated against the ground-truth measurements. The position and orientation deviations from the ground-truth for the different distances are summarized in the following table for the two setups. Specifically, the error is measured for the deviation in the translation in meters and for the rotation deviation along each axis in degrees. Note that for symmetrical objects such as the bottle of water and the pillbox, only the Z axis is the one of interest while the rest are omitted. It is exhibited that although there are some negligible differentiations among the estimations for the two different setups, the estimated poses for each object remain within the acceptable for the grasping limits as set in D5.2.

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Groundtruth Model</th>
<th>Reconstructed object –View1</th>
<th>Reconstructed object –View2</th>
<th>Reconstructed object –View3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cup</td>
<td>![Cup Image]</td>
<td>![Cup Image]</td>
<td>![Cup Image]</td>
<td>![Cup Image]</td>
</tr>
<tr>
<td>Spaghetti Box</td>
<td>![Spaghetti Box Image]</td>
<td>![Spaghetti Box Image]</td>
<td>![Spaghetti Box Image]</td>
<td>![Spaghetti Box Image]</td>
</tr>
</tbody>
</table>

**Figure 37: Evaluation of the object 3D reconstruction algorithm for small objects.**
Table 21: Summary of the experimental evaluation of the 3D reconstruction method. The deviation in the Translation is measured in mm, and the deviations from the three rotational axes dRx, dRy, dRz are in degrees. For the symmetric objects only the orientation of the one principal axis is measured. The detection method is assessed for various Distances (0.9, 1.4 and 1.8)m between the RGBD sensor and the object.

<table>
<thead>
<tr>
<th>Object</th>
<th>Bottle</th>
<th>Pot</th>
<th>Spaghetti Box</th>
<th>Pillbox</th>
<th>Cup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>Translation</td>
<td>dRx</td>
<td>dRy</td>
<td>dRz</td>
<td>Translation</td>
</tr>
<tr>
<td>0.9m</td>
<td>6.8mm</td>
<td>-</td>
<td>-</td>
<td>3.7°</td>
<td>6.5 mm</td>
</tr>
<tr>
<td>1.4m</td>
<td>6.3mm</td>
<td>-</td>
<td>-</td>
<td>8.1°</td>
<td>5.5 mm</td>
</tr>
<tr>
<td>1.8m</td>
<td>7.4mm</td>
<td>-</td>
<td>-</td>
<td>4.9°</td>
<td>5.0 mm</td>
</tr>
</tbody>
</table>

Deviation of the ground-truth with manually created models (Using Blender SW) for various distances from the camera

<table>
<thead>
<tr>
<th>Object</th>
<th>Bottle</th>
<th>Pot</th>
<th>Spaghetti Box</th>
<th>Pillbox</th>
<th>Cup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>Translation</td>
<td>dRx</td>
<td>dRy</td>
<td>dRz</td>
<td>Translation</td>
</tr>
<tr>
<td>0.9m</td>
<td>5.8 mm</td>
<td>-</td>
<td>-</td>
<td>1.4°</td>
<td>6.9 mm</td>
</tr>
<tr>
<td>1.4m</td>
<td>5.5 mm</td>
<td>-</td>
<td>-</td>
<td>3.1°</td>
<td>6.6 mm</td>
</tr>
<tr>
<td>1.8m</td>
<td>3.3 mm</td>
<td>-</td>
<td>-</td>
<td>7.9°</td>
<td>1.7 mm</td>
</tr>
</tbody>
</table>
8.7.2 **Evaluation of large object state tracking**

As part of the home environment monitoring component of this deliverable, we also describe herein results from the large object state tracking algorithms initially presented in D3.1, applied in a realistic setting of the target RAMCIP application scenarios. More precisely, the effectiveness of the method for the large objects state tracking has been examined on a realistic setup, at the premises of the LUM apartment that will host project pilot trials.

During the data acquisition at the LUM apartment, various activities were performed (meal preparation, cooking etc.), while the actors omitted to close some electric appliances towards simulation of the subUC 4.2. Using the acquired data, tests have been executed for two different large scale articulated objects that were involved during the executed activities, i.e. fridge, cupboard. The method exhibited remarkable performance in the scenarios where the query large object was within the point cloud obtained from the RGBD mounted onboard the robot, while the occlusions from the user were negligible. Even if the user interacted with the object during the execution, the state of the object was correctly recognized and the positions and states of its articulated parts at the given time were identified correctly. In the Figures appended below, a series of testing results, of two different large articulated objects during interaction with the human are illustrated.

![Figure 38](image1.png)

**Figure 38:** The cupboard model (red), with all its articulated models in different colours. All cupboard doors were labelled as "closed".

![Figure 39](image2.png)

**Figure 39:** The cupboard model (red) is registered to the scene, and the left door of the cupboard (pink) is found to be open. From the other two doors of the cupboard, only the rightmost one was detected to be closed. The middle door was occluded by the user and its state couldn't be detected.
Figure 40: A series of instances where the user interacted with the fridge. The parts of the fridge (base, upper door, lower door) are illustrated with different colours. The fridge model was registered to the scene, and the states of its doors were recognized in each case. In the first row, the lower fridge door was
recognized as “Open” and the upper door as “Closed”. In the last three rows, the upper door was recognized again as “Closed” and the state of the lower door was correctly recognized as “Open” during the human interaction with partial occlusions.
9. Conclusions

This deliverable presented the developed methodologies and their respective results that have been achieved in the scope of the RAMCIP Task 3.2 “Human activity monitoring”, as well as the progress that have been made from Task 3.1 “Modelling and monitoring the home environment” since the delivery of D3.1. Specifically, it summarized the efforts and outcomes towards the establishment of the RAMCIP methods for monitoring human activity, including human detection and person identification, user body pose tracking and low-level action recognition in relation to objects of the home environment. Alongside, the final efforts and outcomes of T3.1 have been described, which have led to the implementation of the RAMCIP integrated framework for home environment monitoring, on the basis of the dynamic hierarchical semantic mapping approach initially described in D3.1.

In this scope, the development of accurate and robust solutions suitable to operate in realistic environments comprised the main concern of the T3.1 and T3.2 efforts. In order to ensure that the newly developed, adopted and extended methodologies will be operable for the variety of applications that necessitate from the identified use cases (D2.1), state-of-art algorithms have been implemented, tested and extended where necessary, so as to establish algorithms and methods adequate to cover the RAMCIP robot needs and its application scenarios, which involve the robot’s operation in realistic environments.

The first part of the reported research focused on the problem of human detection during the robot’s operation. An integrated human detection pipeline has been developed, through the fusion of vision-based and laser-based approaches. Specifically, when the user is within the FoV of the robot’s camera, further to the well-established approach of foreground/background segmentation in a static scene with a moving subject, an approach based on the house map, as encoded in the RAMCIP hierarchical semantic mapping framework has been implemented, along with a single-shot human detection solution that is capable to further help in cases of significant changes between the current static part of the scene and the registered map. The last two approaches enable human detection when both the scene and the human are static, while the human detection approach is complemented by a user orientation estimation, which is necessary for the subsequent human pose tracking. Alongside, our integrated human detection pipeline employs a solution for also detecting human presence outside the robot camera’s FoV; this is achieved by utilizing the robot’s laser sensor readings.

The second part of the reported research has focused on person identification. On one hand, the RAMCIP robot should be capable to recognize its primary user, even among the presence of multiple humans in the user’s home. On the other hand, the robot should be able to recognize the user’s known relatives and friends, being able to detect cases where for e.g. a stranger tries to enter the user’s home and therefore, the subUC2.4 (Detection of unknown persons/strangers) should be triggered. To this end, the core part of person identification is conducted through facial expression recognition. Alongside, a further person re-identification approach has been implemented through the fusion of SoA soft biometric characteristics with color information, offering the robot the capability to detect in real-time its primary user as soon as s/he enters again the robot’s FoV, as well as further co-existing persons.

As concerns the third highly challenging part of the reported research, focusing on robust human pose tracking, an integrated pipeline has been developed, comprising the fusion of a well-established discriminative pose estimator with a generative pose tracking approach based on articulated full body model fitting. Through the T3.2 efforts, a state-of-art articulated pose tracker has been extended through the addition of a series of features focusing on the advanced
processing of the user silhouette 3D point clouds, towards better occlusion handling and compensating for the noisy data that appear in realistic environments. As a result, the developed integrated human pose tracking pipeline leads to improved robustness in the realistic settings foreseen in the RAMCIP target application scenarios.

Having established through the above, the capability for the robot to detect humans, identify its target user and track her/his pose, the fourth research part of the T3.2 research concerned the recognition of low-level user actions in relation to the home environment objects. In this scope, a state-of-art action recognition method has been adopted and extended, enabling improved performance in segmented action sequences. Alongside, given the continuous nature of the RAMCIP data acquisition process during human monitoring in realistic applications, a segmentation solution based on user interactions with household objects has been developed. Notably, the latter enables the recognition of user actions in relation to the home environment, which are thereafter essential to be provided as inputs to the user high-level activity and behaviour monitoring module of T3.4.

Finally, as concerns the home environment monitoring part, the T3.1 advances reported in the present deliverable focus first of all on the realization of an integrated, change detection -based household objects monitoring approach, which enables the RAMCIP robot not only to keep track of the home environment state, but to also augment its knowledge on the user’s detected actions with information on the objects (small and large ones) with which s/he interacts. Moreover, further elaboration and testing has been performed on the main method of the RAMCIP toolkit for small household objects reconstruction, originally presented in D3.1, leading to a further validated turntable-based method that can be applied for developing object models at the project’s pilot trials environments.

The above algorithms and methods, described in the present deliverable, have been experimentally evaluated on situations closely resembling the RAMCIP target application scenarios, on public datasets and on datasets collected by the RAMCIP consortium. Overall, the experimental results have demonstrated the capability of the T3.2 and T3.1 outcomes to result into adequately robust solutions once integrated on the RAMCIP robot, capable to be applied in the project’s trial scenarios. At present, the research outcomes presented herein have led to the development of corresponding functional software modules, whose integration on the ROS environment of the RAMCIP robot is currently in progress. The majority of the developed modules will be part of the integrated V1 RAMCIP robot that will be used in the project’s preliminary trials with end users, planned to start at the premises of the LUM RAMCIP partner by the end of November 2016.
References


