PH.D. SCHOOL IN INFORMATION ENGINEERING - UNIVERSITY OF PADOVA

FINAL REPORT

Student name: **Marco Carraro**

Cycle: **XXX**

Curriculum: **ICT**

Supervisor name: **prof. Emanuele Menegatti**

Titolare di borsa di Ateneo

Thesis title: Real-time RGB-D perception of humans for robots and camera networks

PART 1 - COURSES, CONFERENCES AND MOBILITY

Courses for Ph.D. students

- *Real-Time Systems and Applications* by Prof. G. Manduchi with final mark 28/30;
- *Computational Inverse Problems* by Prof F. Marcuzzi with final mark 30/30;
- *Statistical Methods* by Prof L. Finesso with final mark A;
- *GPU Computing* by J. Pantaleoni (NVIDIA researcher) and Prof. F. Marcuzzi with final mark 28/30;
- *Principles of Cloud Computing* by Prof. T. Vardanega with final mark 28/30;
- *Bayesian Machine Learning* by Prof. G.M. Di Nunzio (only attended);

Seminars

- *Recent advances on coordination in Multi-Robot Systems* by prof. Alessandro Farinelli, associate professor at University of Verona, at UNIPD on May 15th, 2015;
- *Proof, Secrets, and Computation* by Prof. Silvio Micali, Ford Professor of Engineering CSAIL, MIT, **Turing Award 2012,** at UNIPD on May 25th, 2015;
- *Do brains compute?* by Prof. R. Sepulchre, Dept. of Engineering, University of Cambridge, UK, at UNIPD on June, 18th, 2015;
- *Human Arm Mechanics: from system identification to neural control* by prof. Davide Piovesan, Assistant Professor, Bio-Medical Engineering, Gannon University, at UNIPD on July, 7th, 2015;
- *Hybrid Approaches for Synchrony and Memory for Parallel Graph Algorithms* by prof Nancy Amato, Unocal Professor and Regents Professor in the Department of Computer Science and Engineering at Texas A&M University, at UNIPD on July, 17th, 2015;

- *Postdoctoral Research in Informatics at the Department of Information Engineering of the University of Padova*, at Centro congressi "A. Luciani" in via Forcellini 170 on July, 18th, 2015;
- *Autonomous Mobile Robot Research at Toyohashi University of Technology* by Prof. J. Miura, Colloquia at UNIPD on Oct. 5th, 2015;
- *Semantic Image Interpretation* by I. Donadello, PhD Student at FBK Fondazione Bruno Kessler in Trento, at UNIPD on Nov. 26th, 2015;
- *Building Heterogeneus Systems: A view from the Trenches* by Ryan Kastner, Professor of Computer Science, University of California at San Diego, at UNIPD on Jan 12th, 2016;
- *The Walk-Man humanoid robot: whole-body loco-manipulation planning and control* by prof. Lucia Pallottino, Univerisity of Pisa, at UNIPD on March 31st, 2016;
- *Interactive Control & Learning for Robots - What we need and why!* by K. Listmann, Laboratory Group Manager at ABB in Ladenburg, at UNIPD on May 20th, 2016;
- *Incremental Large-Scale Dense 3D Reconstruction* by Andrea Romanoni, PhD Candidate – Politecnico di Milano, at UNIPD on Nov 15th, 2016;
- *Web 3.0 and Beyond* by Vinton Cerf, Google Vice-President, **Turing Award 2004**, at UCLA on March 6th, 2017;

Participation to International Conferences and Workshops

- AIRO @ AIxIA, Corso Ercole I d'Este, 37, Università degli Studi Di Ferrara, Sept 22nd, 2015 in Ferrara (Italy) – Conference Speaker;
- IAS-14, 14th International Conference on Intelligent Autonomous Systems, July 3rd - July 7th 2016 in Shanghai (China) – Conference Speaker;
- ECMR 2017, The European Conference on Mobile Robotics, 06-08 Sept. 2017 in Paris (France) – PLANNED;
- IROS 2017, IEEE/RSJ International Conference on Intelligent Robots and Systems, 24-28 Sept. 2017 in Vancouver (Canada) – PLANNED;
- ICCV 2017, IEEE/CSV Internation Conference on Computer Vision, 22-29 Oct. 2017, In Venezia (Italy) – PLANNED

Invited talks

- Talk about my research to the research team of Fyusion Inc. on February 24th, 2017 – San Francisco, California (USA)
- "OpenPTrack, an Open Source People-Tracking Platform: Algorithms and their Implications for Human-Computer Interaction and Information." By **M. Carraro** and R. Illum at the department of Information Studies @ UCLA – Los Angeles, May 5th, 2017

Participation to European Projects:

- team member of DODICH – Creation of a machine for viticulture powered by renewable sources

Participation to International Robotic Challenges

- **team member in** RoCKIn@Work 2014 in Tolosa (France), 25-30 November 2014.
- **team member in** MBZIRC 2017, 1st Mohamed Bin Zayed International Robotics Challenge, 11-21 March 2017 in Abu Dhabi (United Arabian Emirates) – Third place in the Grand Challenge.

Mobility periods

- From Jan 8th, 2017 to July 27th, 2017: visit to the UCLA Remap Group headed by Jeff Burke at UCLA, Los Angeles, California (USA), on the topic "Multi-view, multi-person 3D body pose estimation and recognition from RGB-Depth camera networks";

Teaching activities

- Laboratory support:
  1. Co-supervisor of the master student M. Pierobon. Thesis title (in Italian): "Algoritmo di rilevazione di persone a terra da dati 3D per robot di telepresenza". Thesis title in English "An Algorithm to Detect Fallen People from 3D Data for a Telepresence Robot". Graduation date: Oct 3th, 2016;
  2. Co-supervisor of the master student F. Brea. Thesis title: "Potential field navigation for telepresence robots driven by BCI". Graduation date: Oct 3th, 2016;
  3. Co-supervisor of the bachelor student F. Vendramin. Thesis title (in Italian): "Algoritmi di navigazione per robot di servizio". Graduation date: Sept 24th, 2015;
  4. Co-supervisor of the bachelor student G. Beraldo. Thesis title (in Italian): "Sviluppo di un software per il controllo di un robot mobile di telepresenza via Brain-Computer-Interface". Graduation date: Sept 24th, 2015;
- Tutor junior:
  5. 40 hours for the laboratory of the course "Fondamenti d'Informatica", A.Y. 2015/16, by Prof. Adriano Luchetta, first semester, bachelor's degree in "Ingegneria dell'Informazione";
  6. 60 hours for the laboratory of the bachelor course "Architettura degli Elaboratori", A.Y. 2015/16, by Prof. Emanuele Menegatti, second semester, bachelor's degree in "Ingegneria dell'Informazione";
- Didattica integrativa:
  7. 12 hours for the lectures on Point Cloud Library of the master course "Elaborazione di dati tridimensionali", A.Y. 2016/2017, by prof. Emanuele Menegatti, first semester, master's degree in "Ingegneria Informatica"
- Other activities:

8. Lecture on "CUDA programming" at the master course "Elaborazione dei Dati 3D", A.Y. 2015/16, by Prof. Emanuele Menegatti, Dec 14th, 2015, master's degree in "Ingegneria Informatica";
9. Lecture on "CUDA programming" at the master course "Elaborazione dei Dati 3D", A.Y. 2016/17, by Prof. Emanuele Menegatti, Dec 7th, 2016, master's degree in "Ingegneria Informatica";
10. 12 hours as E-Tutor at the course "Architettura degli Elaboratori", A.Y. 2015/16, by Prof. Emanuele Menegatti and Antonio Rodà for the project "Integrating technology in higher education to enhance work life balance", second semester, bachelor's degree in "Ingegneria dell'Informazione".
11. Design and implementation of the lab activities and homeworks for the course "Computer Vision" by prof. Ghidoni Stefano and Zanuttigh Pietro A.Y. 2017/2018 – PLANNED

PART 2 - RESEARCH ACTIVITY

The aim of my research in these three years is to provide new efficient perception algorithms for robotics and camera networks, in particular, when humans are involved. Perception provides important feedback to prevent failures, recover from them and also making smart choices and, therefore, to create autonomous robots as well as new artificial intelligences that can interact actively with humans. In this three years, we contributed on different aspects to i) enhancing perception algorithm performance, ii) providing efficient and real-time solutions and iii) making state-of-the-art works available to the community as open-source.

Given the recent advances of new generation RGB-Depth sensors, we focalized our research on the data provided by such sensors. Exploiting these new sensors, many researchers in the Computer Vision community have contributed to improve the performance on tasks that were before impossible or difficult to solve e.g: face recognition, people re-identification or body pose estimation. In many of such works, the recognition capabilities of the algorithms developed reached recognition rates close to and, in certain cases beyond, human capabilities. Nevertheless, the majority of those works require a large computational effort and provide non-real-time outputs. While such solutions can be very useful in particular applications or for academic purposes, they are not applicable to robots or when the goal is to interact with humans in real-time.

The objective of our work is to provide efficient and ready-to-use algorithms validated with real-time applications (most of the times, mobile robotics), while keeping (or even enhancing) state-of-the-art recognition quality. One of the main focus is also on the exploitation of visual data from different point of views (i.e. using camera networks). The major contributions of our work is about Human 3D Body Pose Estimation in camera networks, but we contributed with novel technologies also in the following fields:

- Human Body Pose Recognition,
- Object Tracking in camera networks,
- People Tracking,
- People re-identification,

- Ambient Assisted Living,
- Smart Cameras

As a further contribution, in order to provide benefit and a future baseline for other researchers, most of our work has been released as open-source and our datasets made available to the community.

The **capability to segment the body parts** or more generally the skeleton of a person in an unsupervised way, is a killer application in many fields: from health-care to Ambient Assisted Living, from surveillance to action-recognition and people re-identification. Our goal is to provide a marker-less motion-capture system using low cost RGB-D cameras. When multiple views are available and the camera network extrinsic calibration is known, we can fuse the 3D information about the body of the person extracted by each single camera by taking advantage of the different points of view of the different cameras. The contribution of this work is two-fold: i) we enhanced a state-of-the-art body part detector by applying a people detection pre-filtering to the input depth image, ii) we proposed a novel way to generate a virtual depth image of each person in the scene from multiple views which takes into account the extrinsic calibration of the network and generates the frontal view of each person. The frontal view of each person is computed by means of Principal Components Analysis of the set of points obtained by projecting the different 3D points belonging to each person on the ground floor. The virtual depth image is then generated by filling the possible holes obtained by the reprojection of the points on the reference camera image plane. The final step is to compute the skeleton by using an enhanced single-view state-of-the-art body part detector. Experiments show how this method improves both the single-view version of the body part detector and a baseline multi-view skeleton estimator. Poses are generated at 9 fps using a non-optimized version of the code.
Last years have seen a general improvement in the field of body pose estimation also called **skeletal tracking systems.** We propose a framework to compute the body pose of each person in the scene using an RGB-D camera network. We chose a different approach from the one used in our previous work for the same goal by exploiting the very good results achievable by state-of-the-art single view algorithms in RGB-D camera networks. Indeed, if the single-view algorithm used for computing the current body pose is good enough, also a fusion between outcomes instead of raw data can improve the overall performance. We proposed a multi-view system to exploit current state-of-the-art Deep-Learning-based body pose estimation algorithms on the single cameras. We first bring the 2D imagers to provide 3D skeletons by combining the depth information together with the RGB one (the body pose estimation algorithms works just on the RGB). Then, we send each skeleton obtained by each camera to a central node in the network. This node, called master computer, is in charge of fusing the different skeletons obtained asynchronously by the different cameras. The fusion is performed using different Unscented Kalman Filters, one for each joint. Indeed, while the single-view performance is very good, occlusions and illumination variance can result in noisy or missing joints. The filters used are able to predict new positions for the different joints or to reject noisy ones. They use a linear velocity model, since they describe well the motion of the joints in the little time in between two different detections. The system has been successfully validated with different camera networks using 4 to 8 RGB-D cameras and a space up to 7x20 meters, with no constraints about the fixed background or particular clothing to wear. As a further contribution, the final algorithm has been released as open-source as part of an open-source project about human-computer interaction: the OpenPTrack library. Efficient **Pose Recognition** is a fundamental capability of Human-Machine-Interaction systems. For this reason, we implemented a novel system that classifies the similarity of the current

pose of each user in the scene as part of the OpenPTrack library. The input consists of the skeletons computed by the multi-view pose estimation. The library reads some *gallery* poses recorded offline and classifies each pose stroke by each user with a similarity score for each pose in such *gallery*. If one similarity score is good enough, the pose is classified as an instance of the same gallery pose. This ability is not just important for Human-Machine-Interaction, but also for Action Recognition, as it can be considered as a more informative input as the raw joint positions. One of the novelty of our approach is that each gallery pose is recorded regardless of the people to recognize. This is possible thanks to two phases: a normalization and a frontal pose generation. The normalization phase generates a unit vector skeleton starting from the current one by transforming each skeleton link in its unit vector. In this way, we can directly compare skeletons coming from different people regardless of their size. The frontal view generation is important since a person can face different directions and we want to keep the gallery pose recording as simple as possible. For each skeleton in the scene, we compute the orientation vector by using linear algebra relationship between the joints and we roto-translate the skeleton using a common reference system. Once the skeleton is normalized and in standard position it is directly compared with the gallery poses. The similarity score is computed as a sum of differences between respective joints, then a threshold is used to classify the current pose. The final algorithm runtime is negligible compared to the body pose estimator one, therefore the final pose recognition rate is roughly the same as the final pose estimation one.

Ambient Assisted Living (AAL) is the discipline that aims at assisting people at home. Last years have seen a worldwide lengthening of life expectancy and, as a consequence, an increment of advanced assistive solutions for integrated care models. In this perspective, home robots will play a crucial role. Not only they will keep the house safe by monitoring and detecting anomalies or sources of hazards, but they can also be companions able to enhance their social life, e.g. by better connecting them with their relatives and friends. During these years, we proposed a new mobile robot for AAL. The robot, called Orobot, is equipped with several sensors as a Microsoft Kinect v2, a Laser Range Finder, three bumpers, and some infrared receivers for autonomous docking to the recharging station. We implemented different state-of-the-art modules such as autonomous navigation in partially known environments, obstacle avoidance of both static and dynamic obstacles, people tracking, people recognition, speech recognition and fallen people detection. The latter software module represents a novel contribution we developed. Fallen people detection is the ability of a robot to find people lying on the floor. Many recent studies, indeed, proved that this is one of the principal death causes in people who live alone at home. Since it is unlikely that a robot can capture the act of falling while patrolling, we concentrate our work on the detection of lying people, enabling the robot to call for help.

The fallen people detection algorithm we developed works in two stages. The first one is in charge of finding clusters of points (i.e. group of contiguous points in the 3D point cloud we obtain from the 3D sensor) which could belong to a human body. The second one is in charge of splitting the 3D data in *patches* (a common technique used in the semantic mapping field) and classify them as part of a person or not. As a final step, the robot takes advantage of the knowledge of the map of the environment. It does an accumulation of the fallen people location found, localizing them in the map and periodically rejecting the false positives or validating them. As part of this contribution we recorded a new dataset of fallen people in two different scenes. This new dataset has been released to be available to the scientific community, Results on this dataset show the ability of our approach to recognize all the people in both scenes and very good results of our classifiers in terms of accuracy,

precision and recall. The overall algorithm runs in real-time (about 7 fps) on the mobile robot we used (the Orobot prototype).

An important aspect for Human-Machine-Interaction is the ability to detect and track people in a scene. Nowadays, many people tracking libraries are able to track people with unique IDs. Nevertheless, in case of occlusions or when people exit and re-enter the camera network Field-Of-View (FOV), current state-of-the-art systems assign new IDs to the new tracks. **People Re-identification** is the ability that solves this problem. In particular, in our work, we exploited Face Recognition to add such capabilities to our system. Face Recognition has seen many contributions in the Computer Vision community. Nevertheless, few works consider the problem of face recognition when multiple cameras are available. We proposed a framework to solve this problem when multiple RGB-D sensors are available. We take advantage of current state-of-the-art face detectors and recognizers, in particular OpenFace, one of the most famous solution for face recognition using Deep Neural Networks, to build an incremental signature of each track. The signature is learnt and updated online and whenever a person signature is close enough to another signature in the database kept by the network, the old track ID will be reassigned to that person. The final algorithm runs in real-time and has been released as open-source.

Since the introduction of the Microsoft Kinect v1, commercially named Microsoft Kinect 360, for gaming purposes, RGB-Depth sensors have dramatically changed the Computer-Vision research. RGB-D sensors are typically composed of a standard camera (RGB) and an infrared camera in charge of providing the depth information (D). Exploiting such information, it is possible to build information-reach data structures as Point Clouds allowing computers (and therefore, autonomous machines and robots) to see the world directly in three dimensions. In 2013, Microsoft released the second generation of the Kinect sensor, commercially called Microsoft Kinect One (here called Kinect v2). Its infrared camera implements a Time-of-Flight (ToF) technology which greatly improved the performance with respect to the first generation sensor which used Structured Light projection. For example, the data obtained are better in terms of depth estimation error (linear vs quadratic), solar radiation rejection (with direct sun-light Kinect v2 works up to 2m vs 0m using Kinect v1), both depth and image resolution (1920x1080 and 424x512 vs 1024x768 and 320x240) and general surface quality. While this sensor is able to provide depth and RGB streams at 30fps, it is computationally heavy to build the Point Cloud at suitable frame rates without the usage of GPU-enabled computers. Recently, Nvidia released GPU-enabled embedded computers which can be exploited in computationally intensive applications and at the same time have low requirements in terms of cost, power and space occupancy. One of our contribution is the **creation of a new smart camera** for mobile robotics using the Kinect v2 together with one of these embedded computers (namely, NVidia TX1). The new library exploits the GPU to build the Point Cloud efficiently. The data passage between the GPU and CPU central memory which is often the main drawback for GPU-based algorithms is kept to a minimum (about 1.5% of the total bandwidth). We tested the developed smart camera in people tracking scenarios on a mobile robot proving that the solution increased the state-of-the-art library for Kinect v2 by tripling and doubling, respectively, the frame rates that were obtained in terms of point cloud generation and people tracking.

PART 3 - PUBLICATIONS

List of publications on international journals

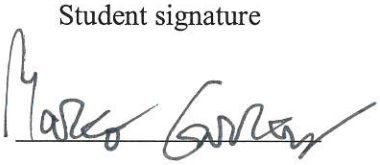
1. Human Pose Estimation from fused depth information in camera networks, **M Carraro**, M Munaro, E Menegatti, (SUBMITTED to Special Issue on the Intelligent Autonomous Systems Conference – Robotics and Autonomous Systems Journal)
2. RUR53: an Unmanned Ground Vehicle for Navigation, Recognition and Manipulation. N. Castaman, E. Tosello, M. Antonello, N. Bagarello, S. Gandin, **M. Carraro**, M. Munaro, R. Bortoletto, S. Ghidoni, E. Menegatti, E. Pagello, (SUBMITTED to Special Issue on the Mohamed Bin Zayed International Robotics Challenge 2017 - Challenges in Autonomous Field Robotics – Journal of Field Robotics)
3. Cost-efficient RGB-D smart camera for people detection and tracking, **M Carraro**, M Munaro, E Menegatti, Journal of Electronic Imaging 25 (4), 041007-041007 (JEI-2016)

List of publications on conference proceedings

1. Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks, **M. Carraro**, M. Munaro, J. Burke, E. Menegatti, (SUBMITTED, IEEE ICRA2018)
2. Fast Multiple Object Tracking in RGB-D Camera Networks, Y. Zhao, **M. Carraro**, M. Munaro, E. Menegatti (ACCEPTED, IEEE IROS2017)
3. Fast and Robust Detection of Fallen People from a Mobile Robot, M Antonello, **M Carraro**, M Pierobon, E Menegatti (ACCEPTED, IEEE IROS2017)
4. OpenPTrack: Real-time, Multi-camera Computer Vision Infrastructure for Artists, R. Illum, **M. Carraro**, E. Menegatti, J. Burke, (SUBMITTED to IEEE IROS 2017 Workshop - Towards an artist-in-the-lab Framework)
5. People Tracking and Re-Identification by Face Recognition for RGB-D Camera Networks, K. Koide, E. Menegatti, **M. Carraro**, M. Munaro and J. Miura (ACCEPTED, ECMR2017)
6. A powerful and cost-efficient human perception system for camera networks and mobile robotics, **M Carraro**, M Munaro, E Menegatti, International Conference on Intelligent Autonomous Systems, 485-497 (IAS-14)
7. Improved skeleton estimation by means of depth data fusion from multiple depth cameras, **M Carraro**, M Munaro, A Roitberg, E Menegatti, International Conference on Intelligent Autonomous Systems, 1155-1167 (IAS-14)
8. An Open Source Robotic Platform for Ambient Assisted Living. **M. Carraro**, M. Antonello, L. Tonin, E. Menegatti, AIRO@ AI* IA, 3-18

21/08/2017

Student signature



Supervisor signature