Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Curriculum in Ingegneria Informatica, Gestionale e dell'Automazione

---

Human Behaviour Understanding using Top-View RGB-D Data

Ph.D. Dissertation of:
Daniele Liciotti

Advisor:
Prof. Emanuele Frontoni

Curriculum Supervisor:
Prof. Francesco Piazza

XVI edition - new series

Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Curriculum in Ingegneria Informatica, Gestionale e dell'Automazione

Human Behaviour Understanding using Top-View RGB-D Data

Ph.D. Dissertation of:
Daniele Liciotti

Advisor:
Prof. Emanuele Frontoni

Curriculum Supervisor:
Prof. Francesco Piazza

XVI edition - new series

Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Facoltà di Ingegneria
Via Brecce Bianche – 60131 Ancona (AN), Italy

To my parents

Acknowledgments

Firstly, I would like to thank my supervisors Emanuele Frontoni and Primo Zingaretti who always offered valuable advise and insight, helped me along my journey as a PhD candidate. Especially Emanuele Frontoni always offered support, contributed very many good ideas. Thank you both for giving me this great opportunity and helping me along the way.

Secondly, I would like to express my gratitude to Prof. Tom Duckett and Prof. Nicola Bellotto for the ideas and suggestions received during the visiting period at University of Lincoln (UK).

I would like to thank everybody who made my time at university unforgettable. First and foremost, I want to thank Annalisa Cenci and Marina Paolanti for being the best laboratory mates. I consider you as sisters. We have done a great job together!

I am grateful to all the members of DII, in particular the actual and former members of VRAI (Vision, Robotics and Artificial Intelligence) team.

Finally, I would like to thank my parents, my brother with his family, and Laura. Everything would have been more difficult without your constant encouragement.

Ancona, October 2017

Daniele Liciotti

Abstract

The capability of automatically detecting people and understanding their behaviours is an important functionality of intelligent video systems. The interest in behaviour understanding has effectively increased in recent years, motivated by a societal needs.

This thesis is focused on the development of algorithms and solutions for different environments exploiting top-view RGB-D data. In particular, the addressed topics refer to Human Behaviour Understanding (HBU) in different research areas.

The first goal is to implement people detection algorithms in order to monitor the people activities. To this aim, a thorough study of the state of the art has been conducted to identify the advantages and weakness. An initial approach, proposed in this thesis, is based on Computer Vision (CV) techniques, it regards the extraction the head of each person using depth data. Another approach is based on deep learning and is proposed to simplify the heads detection implementation in chaotic environments and in the presence of people with different heights. These solutions are validated with a specific dataset.

The second goal is to extract several feature from subject and to identify possible interactions that they have with the surrounding environment.

Finally, in order to demonstrate the actual contribution of algorithms for understanding the human behaviour in different environments, several use cases have been realized and tested.

Declaration

I, Daniele Liciotti, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that this work was done wholly while in candidature for a research degree at the Università Politecnica delle Marche and that this thesis has not previously been submitted for a degree or any other qualification at this University or any other institution. I also confirm that where I have consulted the published work of others, this is always clearly attributed and that where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work and I have acknowledged all main sources of help. I confirm that where the thesis is based on work done by myself jointly with others. Parts of this work have been published as:

- D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, and V. Placidi. Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network. In *International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage*, pages 146–157. Springer, Cham, 2014.

- D. Liciotti, E. Frontoni, A. Mancini, and P. Zingasretti. Pervasive system for consumer behaviour analysis in retail environments. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, volume 2. 2017.

- D. Liciotti, E. Frontoni, P. Zingaretti, N. Bellotto, and T. Duckett. Hmm-based activity recognition with a ceiling rgb-d camera. In *ICPRAM (International Conference on Pattern Recognition Applications and Methods)*, 2017.

- D. Liciotti, G. Massi, E. Frontoni, A. Mancini, and P. Zingaretti. Human activity analysis for in-home fall risk assessment. In *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pages 284–289. IEEE, 2015.

- D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti. Person re-identification dataset with rgb-d camera in a top-view configuration. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 1–11. Springer, 2016.

- M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini, M. Contigiani, and P. Zingaretti. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognition Letters*, 2016.

Contents

List of Acronyms

AAL	Ambient Assisted Living
ADLs	Activities of Daily Living
CMC	Cumulative Matching Characteristic
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CT	Computer Tomography
CV	Computer Vision
DoS	Degrees of Semantics
GMM	Gaussian Mixture Models
HBU	Human Behaviour Understanding
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
IoU	Intersection over Union
IRE	Intelligent Retail Environments
LRN	Local Response Normalization
PSA	Pixel State Analysis
RADiAL	Recognition of Activity DAily Living
Re-id	Re-Identification
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
TOF	Time of Flight
TVHeads	Top-View Heads
TVPR	Top-View Person Re-Identification

Chapter 1.

Introduction

This Thesis addresses the subject of HBU using top-view RGB-D data. The objective of the Thesis is described in this Chapter, together with the definition of the research problem, the main contributions, and the Thesis organization.

1.1. Research problem

In recent years, a lot of researchers have focused the attention on automatic analysis of human behaviour because of its important potential applications and its intrinsic scientific challenges. In several technological fields the awareness is emerging that a system can provide better and more suitable services to people only if it can understand much more about users' preferences, personality, social relationships etc., as well as about what people are doing, the activities they have been concerned in the past, their life-styles and routines, etc.

CV and deep learning techniques are currently the most interesting solutions to analyse the human behaviour. In particular, if these are used in combination with RGB-D data that provide high availability, reliability and affordability.

Detection and tracking algorithms allow to generate motion descriptions of subjects which are used to identify actions or interactions. Consequently, it is possible to associate to a certain sequence of actions a particular behaviour. In this view, investigating technological solutions aimed at improving the environments and adapting them to the specific user requirements, can be very useful.

The problem remains largely open due to several serious challenges, such as occlusions, change of appearance, complex and dynamic background. To counter these challenges, several studies adopt the top-view configuration because it eases the task and makes simple to extract different trajectory features. This setup also introduces robustness, due to the lack of occlusions among individuals.

Different domains are analysed in this Thesis, such as those of video surveillance, Intelligent Retail Environments (IRE) and Activities of Daily Living (ADLs).

1.2. Objectives and contributions

The objective of this Thesis is to understand the human behaviour in different real scenarios using CV techniques applied on RGB-D data in top-view configuration. To this aim, a thorough study of the literature will be presented, identifying advantages, challenges and issues related to the use of this particular configuration. Furthermore, in order to support this research, several use cases will be presented. In particular, one of these was conducted during the five months of Ph.D. visiting period at Lincoln Centre for Autonomous Systems (LCAS) in the School of Computer Science at the University of Lincoln (UK).

1.3. Structure of the Thesis

The Thesis is organized in five Chapters, which describe and detail the different approaches and applications for human behaviour understanding using top-view RGB-D data. The Thesis has the following structure.

Chapter 2 reviews the state-of-the-art about the two main topics addressed: human behaviour analysis and RGB-D data from top-view.

In Chapter 3 are proposed some algorithms for people detection using RGB-D data in top-view configuration.

Chapter 4 describes different use cases, in particular are analysed applications on: video surveillance and analytics, intelligent retail environment, and ADLs.

Finally, conclusions and discussion are drawn in Chapter 5 where, after clarifying the contribution of this work, some future research directions are identified. Furthermore, this Chapter besides arguing over the possibilities that the proposed applications opens up in different topics, summaries also the challenges, the open issues and the limitations that require further investigations.

Chapter 2.

State of art

The aim of this Chapter is to review and discuss the most relevant works on HBU using RGB-D data from top-view. An overview of techniques and solutions is provided, then, the discussion is focused on main scenarios and challenges.

2.1. Human behaviour understanding

Understanding human behaviours is a challenging problem in CV that has recently seen important advances. HBU combines image and signal processing, feature extraction, machine learning and 3D geometry. Application scenarios range from surveillance to indexing and retrieval, from patient care to industrial safety and sports analysis.

The capabilities of automatically detecting and tracking people, and of understanding their behaviours are the crucial key functionalities of intelligent video systems. The interest in HBU has quickly increased in recent years, motivated by a societal needs [13] that include security, natural interfaces, gaming, affective computing, and assisted living.

An initial approach can be the detecting and tracking of the subjects of interest, which in this case are the people. This way it is possible to generate motion descriptors which are used to identify actions or interactions.

Recognising particular behaviours requires the definition of a set of templates that represent different classes of behaviours. Nevertheless, in many scenarios not all behaviours can be characterised by a predefined number of classes nor can be known a priori. Alternatively, it is used the concept of *anomaly*, namely a deviation from the learned behaviours.

In the literature the use of terminology on HBU is ambiguous. In following paragraph a consistent definition of the terms used in HBU is proposed.

2.1.1. Taxonomy

The works of Moeslund *et al.* [82] and Borges *et al.* [17] have been used to create a taxonomy on HBU. Human activities can be categorized into four main groups, namely gesture, action, activity, and behaviour (figure 2.1).

- *Gestures* are movements of body parts that can be used to control and to manipulate, or to communicate. These are the atomic components describing the motion of a person.
- *Actions* can be seen as temporal concatenations of gestures. Actions represent voluntary body movements of an arbitrary complexity. An action implies a detailed sequence of elementary movements.
- *Activities* are a set of multiple actions that can be classified in order to understand human behaviours.
- *Behaviours* are the responses of subjects to internal, external, conscious, or unconscious stimuli. A series of activities may be related to a particular behaviour.

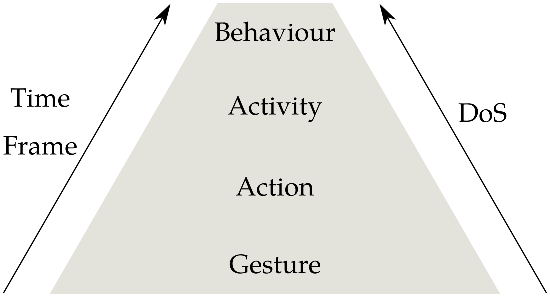


Figure 2.1.: HBU tasks - Classification.

Table 2.1 summarises the different Degrees of Semantics (DoS) considered by the taxonomy, along with some examples. Not only time frame and semantic degree grow at higher levels of this hierarchy, but also complexity and computational cost lead to heavy and slow recognition systems, as each level requires most of the previous level tasks to be done too.

2.2. RGB-D data from top-view

Detecting and tracking people is an important and fundamental component for many interactive and intelligent systems. The problem remains largely open due to several serious challenges, such as occlusion, change of appearance, complex and dynamic background [70].

Table 2.1.: Classification of tasks according to the DoS involved

DoS	**Time lapse**
Gesture	frames, seconds
Action	seconds, minutes
Activity	minutes, hours
Behaviour	hours, days

Popular sensors for this task are RGB-D cameras because of their availability, reliability and affordability. Studies have demonstrated the great value (both in accuracy and efficiency) of depth camera in coping with severe occlusions among humans and complex background. The appearance of devices, such as Microsoft's Kinect[1] and Asus's Xtion Pro Live[2] Sensors motivates a revolution in CV and vision related research. The combination of high-resolution depth and visual information opens up new challenges and opportunities for activity recognition and people tracking for many application fields. Reliable depth maps can provide valuable additional information to significantly improve tracking and detection results.

The task of detecting and tracking people in such image and sequences has proven very challenging although sustained research over many years has created a range of smart methods. Techniques involve extracting spatially global features and using statistical learning with local features and boosting, such as EOH [28], Histogram of Oriented Gradients (HOG) [22] and edgelet [117]. Other challenges such as high variation in human poses, self-occlusions and cross-occlusions make the problem even more complicated.

To counter these challenges, several research papers adopt the top-view configuration because it eases the task and makes simple to extract different trajectory features. This setup also introduces robustness, due to the lack of occlusions among individuals. Figure 2.2 depicts a people counting system from top-view configuration with an RGB-D camera.

The objective of this section is to provide a comprehensive overview of recent developments of people detection and tracking with RGB-D technologies from the top-view perspective, mainly published in the CV and machine intelligence communities. The criteria for topic selection arises from our previous experience with approaches with RGB-D cameras installed in a top-view configuration.

More specifically, this section includes person tracking and recognition, human activity analysis, hand gesture recognition, and fall detection in different fields. The broad diversity of topics clearly shows the potential impact of top-

[1] https://developer.microsoft.com/en-us/windows/kinect/develop

[2] https://www.asus.com/us/3D-Sensor/Xtion_PRO_LIVE/

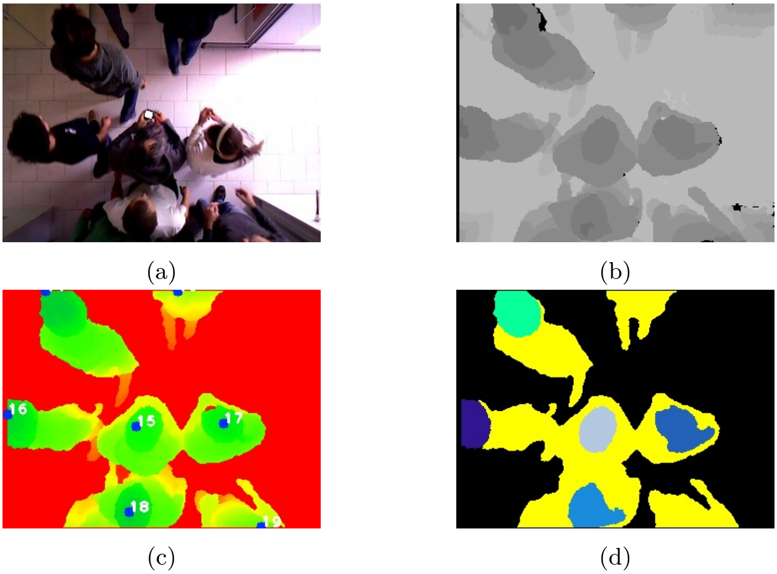

(a) (b) (c) (d)

Figure 2.2.: People counting system from top-view configuration with RGB-D camera.

view configuration in CV. I also summarize main paths that most approaches follow and point out their contributions. I categorize and compare the reviewed approaches from multiple perspectives, including information modality, representation coding, structure and transition, and feature engineering methodology, and analyse the pros and cons of each category.

2.3. Algorithms and approaches

Many vision techniques and algorithms for person detection and tracking have been proposed during the last years and these greatly restrict the generality of the approach in real-world settings. In this section, a survey of current methods, covering both early and recent literature related to algorithms and techniques applied for tracking and detecting humans from top view RGB-D data is presented. In particular, the approaches related to segmentation using background subtraction and statistical algorithms are reviewed.

Kouno *et al.* in [54] describe an image-based person identification task focusing on an image from an overhead camera. The process is based on the background subtraction approach. They apply four features to the identification method, i.e. estimated body height, estimated body dimensions, estimated body size and depth histogram.

In [119], the authors propose a system for passengers counting in buses based on stereovision. The processing chain corresponding to this counting system involves different steps dedicated to the detection, segmentation, tracking and counting. In fact, they have segmented the height maps for highlighting the passengers' heads at different levels (i.e. adults, teenagers, children). The result is binary images that contain information related to the heads, called "kernels". The extraction part attributes a number of parameters to the kernel such as, size of the kernel, shape, average grey level, average height level. Then, with the kernel information, a tracking procedure is applied to analyse the trajectories of the kernels.

The top-view camera setting is also adopted in [68]. In this paper, each depth image in a sequence is segmented into K layers as the Computer Tomography (CT) slides where the depth spacing between two adjacent layers is set to be a fixed value, distance and the number K is an a priori chosen parameter. After that, the region of each slide can be found based on the classic contour finding algorithm. Dynamic time warping algorithm is also applied to address the different sequence length problem. Finally, a Support Vector Machines (SVM) classifier is trained to classify the activities.

In another work the authors with methods of low-level segmentation and tracking develop a system that maps the customers in the store, detects the interactions with products on the shelves and the movement of groups of people within the store [67].

Another segmentation approach is the one proposed in [103]. In this paper, a pipeline verifies that only a single, authorized subject, can enter inside a secured area. Verification scenarios are carried out by using a set of RGB-D images. They used an adaptive Gaussian mixture-based background/foreground segmentation method to exclude parts from the sample image that have the same texture as the background image.

Microsoft Kinect depth sensor is employed in [34] in an "on-ceiling" configuration based on the analysis of depth frames. Elements acquired in the depth scene are recognized by a segmentation algorithm, which analyses the raw depth data directly provided by the sensor. The system extracts the elements, and implements a solution to classify all the blobs in the scene. Anthropometric relationships and features are used to recognize human subjects among the blobs. Once a person is detected, he is followed by a tracking algorithm between different frames.

Dittrich *et al.* [27] present an approach for low-level body part segmentation based on RGB-D data. The RGB-D sensor is installed at the ceiling and observed a shared workspace for human-robot collaboration in the industrial domain. The object classes are the distinct human body parts: Head, Upper Body, Upper and Lower Arm, Hand, Legs and the background rejection. For the generation of data for the classifier training, they use a synthetic representation of the human body in a virtual environment, where synthetic sensors generate depth data. The features used for the description of the object class samples are based on the depth information only, and have been extracted by a centred pixel patch with constant size. As an innovation an optimized training strategy allows a reduced number of training samples while preserving the classification performance.

Further segmentation approach is [46]. Hernandez *et al.* have described a system that operates in troublesome scenarios where illumination conditions can suffer sudden changes. They have been focused on the people counting problem with Re-Identification (Re-id) and trajectory analysis.

A variant of classical segmentation is the one proposed by Tseng in [110]. In this paper, they present a real-time indoor surveillance system which installs multiple depth cameras from vertical top-view to track humans. The system with a framework tries to solve the traditional challenge of surveillance through tracking of multiple people, such as severe occlusion, similar appearance, illumination changes, and outline deformation. To cover the entire space of indoor surveillance scene, the image stitching based on the cameras' spatial relation is also used. The background subtraction of the stitched top-view image has been performed to extract the foreground objects in the cluttered environment. The detection scheme involves different phases such as the graph-based segmentation, the head hemiellipsoid model, and the geodesic distance map. Furthermore, the shape feature based on diffusion distance has been designed to verify the human tracking hypotheses within particle filter.

In [76], the processing of the combined depth image in multiple steps to identify the location, orientation, and people formations is executed. The algorithm adopted in this case finds the highest point of all detected people identified in the previous step, and shifted the depth values of all remaining outlines of people by the difference in height.

An improvement of the classical segmentation techniques is the algorithm proposed by Kepski *et al.* [51]. The first step of the algorithm is nearest neighbor interpolation to fill the holes in the depth map and to get the map with meaningful values for all pixels. Then, the median filter with a 5 × 5 window on the depth array is executed to make the data smooth. The algorithm also extracts the floor and removes their corresponding pixels from the depth map. Given the extracted person in the last depth frame, the region growing is performed to delineate the person in the current frame. To confirm the presence of the tracked subject as well as to give head location a SVM based person finder is used. On the basis of the person's centroid the pan-tilt head rotates the camera to keep the head in the central part of the depth map. Finally, a cascade classifier consisting of lying pose detector and dynamic transition detector is carried out.

An additional paper that describes a method for people counting in public transportation with a segmentation approach is [75]. Kinect sensor mounted vertically has been employed to acquire an images database of $1 - 5$ persons, with and without body poses of holding a handrail. However, in this case the image is processed in blocks in order to find potential local maxima, which are subsequently verified to find head candidates. Finally, non-head objects have been filtered out, based on the ratio of pixels with similar and near-zero value, in the neighbourhood of the maxima.

In [116], the depth images acquired from Kinect camera have been analysed for detecting moving objects using the background subtraction technique. The heads of person are identified by object segmentation in the U-disparity representation.

The approach in [9] investigates a real time people tracking system able to work even under severe low-lighting conditions. The system relies on a novel active sensor that provides brightness and depth images based on a Time of Flight (TOF) technology. This is performed by means of a simple background subtraction procedure based on a pixelwise parametric statistical model. The tracking algorithm is efficient, being based on geometrical constraints and invariants. Experiments are performed under changing lighting conditions and involving multiple people closely interacting with each other.

The same technique is the one applied in [120]. In this paper, the method is composed by two behaviour estimators. The first one is based on height of hand with depth information the second instead on SVM with depth and Pixel State Analysis (PSA) based features and these estimators are used by cascading them.

A method to detect human body parts in depth images based on an active learning strategy is proposed in [12]. The approach is evaluated on two different scenarios: the detection of human heads of people lying in a bed and the detection of human heads from a ceiling camera. The proposal is to reduce both the training processing time and the image labelling efforts, combining an online decision tree learning procedure that is able to train the model incrementally and a data sampling strategy that selects relevant samples for labelling The data are grouped into clusters using as features the depth pixel values, with an algorithm such as k-means.

Tian *et al.*, in [109] have adopted the median filtering to noise removal, because it could well filter the depth image noise obtained by Kinect, and at the same time could protect edge information well. A human detection method using HOG features, that are local descriptors, of head and shoulder based on depth map and detecting moving objects in particular scene is used. SVM classifier has isolated regions of interest (features of head and shoulder) to achieve real-time detection of objects (pedestrian).

A method for human detection and tracking in depth images captured by a top-view camera system is presented in [97]. They have introduced feature descriptor to train a head-shoulder detector using a discriminative class scheme. A separate processing step has ensured that only a minimal but sufficient number of head-shoulder candidates is evaluated. A final tracking step reliably propagated detections in time and provides stable tracking results. The quality of the method has allowed to recognise many challenging situations with humans tailgating and piggybacking.

An interesting binary segmentation approach is the one proposed by Wu *et al.* [117] that have used a Gaussian Mixture Models (GMM) algorithm and reduced depth-sensing noise from the camera and background subtraction. Moreover, the authors have smoothed the foreground depth map using a 5 by 5 median filter. The real-time segmentation of a tracked person and their body parts has been the first phase of the EagleSense tracking pipeline.

In [45] authors described and evaluated a vision-based technique for tracking many people with a network of stereo camera sensors. They have modelled the stereo depth estimation error as Gaussian and track the features using a Kalman filter. The feature tracking component starts by identifying good features to track using the Harris corner detector. It has tracked the left and right image features independently in the time domain using Lucas-Kanade-Tomasi feature tracking. The approach has been evaluated using the MOTA-MOTP multi-target tracking performance metrics on real data sets with up to 6 people and on challenging simulations of crowds of up to 25 people with uniform appearance. This technique uses a separate particle filter to track each person and thus a data association step is required to assign 3D feature measurements to individual trackers.

Migniot in papers [80] and [79] has addressed the problem of the tracking of 3D human body pose from depth image sequences given by a Xtion Pro-Live camera. Human body poses have been estimated through model fitting using dense correspondences between depth data and an articulated human model. Two trackers using particle filter have been presented.

A CV algorithm adopted by many researchers in case of RGB-D cameras placed in top-view configuration is Water filling.

Zhang *et al.* [122] have built a system with vertical Kinect sensor for people counting, where the depth information is used to remove the effect of the appearance variation. Since the head is closer to the Kinect sensor than other parts of the body, people counting task found the suitable local minimum regions. The unsupervised water filling method finds these regions with the property of robustness, locality and scale-invariance.

Even in [1] and in [21], the authors have presented a water filling people counting algorithm using depth images acquired from a Kinect camera that is installed vertically, i.e., pointing toward the floor. The algorithm in [1] is referred to as Field seeding algorithm. The people head blobs are detected from the binary images generated with regard to the threshold values derived from the local minimum values. In [21] the approach called as people tracking increases the performance of the people counting system.

2.4. Challenges and opportunities in the research fields

In this section, the main motivating factors for the installation of RGB-D cameras in top-view configuration are presented. I will discuss the reliable and occlusion free counting of people that is crucial to many applications.

Most of previous works can only count moving people from a single camera, which can fail in crowded environment where occlusions are very frequent. This Thesis focuses the attention on the works with RGB-D data in top-view configuration in three particular fields of research:

- *Video surveillance*;
- *Intelligent retail environment*;
- *Activities of daily living*;

Applications here described cover several fields such as the ones listed below. Datasets collected with RGB-D camera in top-view configuration are reported in 2.4.1.

Video surveillance

Applications developed in this field are related to safety and security in crowded environments, people flow analysis and access control as well as counting. Actual tracking accuracy of top-view cameras over-performs all other point of view in crowded environments with accuracies up to 99%. When there are special security applications or the system is working in usually crowded scenarios the proposed architecture and point of view are the only suitable.

In [15], authors focus their approach on the development of an embedded smart camera network dedicated to track and count people in public spaces. In the network, each node is capable of sensing, tracking and counting people while communicating with the adjacent nodes of the network. Each node uses a 3D-sensing camera positioned in a downward-view. This system has performed background modelling during the calibration process, using a fast and lightweight segmentation algorithm.

A vision based method for counting the number of people which cross a virtual line is presented in [25]. The method analyses the video stream acquired by a camera mounted in a zenithal position with respect to the counting line, allowing to determine the number of people that cross the virtual line, and providing the crossing direction for each person. This approach was designed to achieve high accuracy and computational efficiency. An extensive evaluation of the method has been carried out taking into account the factors that may impact on the counting performance and, in particular, the acquisition technology (traditional RGB camera and depth sensor), the installation scenario (indoor and outdoor), the density of the people flow (isolated people and groups of persons), the acquisition frame rate, and the image resolution. They also analysed the combination of the outputs obtained from the RGB and depth sensors as a way to improve the counting performance.

Another work for people counting discussed in [31]. An algorithm by multimodal joint information processing for crowd counting is developed. In this method, the authors have used colour and depth information along with an ordinary depth camera (e.g. Microsoft Kinect). First, they have detected each head of the passing or still person in the surveillance region using an adaptive modulation approach when the depth scenes vary. Then, they have tracked and counted each head detected in the colour data.

An image-based person identification task is performed by Kouno *et al.* [54]. They have adopted an overhead camera, because of the restriction reduction of the installation location of a camera and the problem solution of occluded images.

The approach in [9] is a real time people tracking system able to work even under severe low-lighting conditions. The system is based on a novel active sensor that provides brightness and depth images based on a TOF technology. Human detection and tracking is also the main goal in [97].

In order to guarantee security in critical infrastructure a pipeline is presented in [103]. It verifies that only a single, authorized subject can enter a secured area. Verification scenarios are carried out by using a set of RGB-D images. Features, invariant to rotation and pose, are used and classified by different metrics to be applied in real-time.

Even, in security field, Tian *et al.* [109] have proposed a human detection method using HOG features of head and shoulder based on depth map and detecting moving objects.

Another people counting application is the technique that uses the mixture of colour and depth images from top-view camera [122]. The U-disparity as depth image projection is introduced in order to increase the accuracy of counting number [116].

The combination of the people counting problem with Re-id and trajectory analysis is faced in [46]. They have extracted useful information using depth cameras. The Re-id task is studied by [71]. Authors have introduced a study on the use of different features exclusively obtained from depth images captured with top-view RGB-D cameras. Top-View Person Re-Identification (TVPR) is the dataset for person Re-id with an RGB-D camera in a top-view configuration. The registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling [66].

Intelligent retail environment

Another important research field is represented by the detection of the interaction between people and environment. More precisely, in the following I will discuss the IRE and intelligent shelf, such as Shopper Analytics systems. [61]. The author of this work presented a low cost integrated system consisted of a RGB-D camera and a software able to monitor shoppers. The camera installed above the shelf detects the presence of people and uniquely identifies them. Through the depth frames, the system detects the interactions of the shoppers with the products on the shelf, and determines if a product is picked up or if the product is taken and then put back, and, finally, if there is not contact with the products.

The same authors, in [63] have described the monitoring of consumer behaviours. The autonomous and low cost system employed is based on a software infrastructure connected to a video sensor network, with a set of CV algorithms, embedded in the distributed RGB-D cameras.

GroupTogether is another system that explores cross-device interactions using two sociological constructs [76]. It supports fluid, minimally disruptive techniques for co-located collaboration by leveraging the proximity of people as well as the proximity of devices.

Migniot *et al.* have explored the problem of people tracking with a robust and reliable markerless camera tracking system for outdoor augmented reality using only a mobile handheld camera. The method was particularly efficient for partially known 3D scenes where only an incomplete 3D model of the outdoor environment was available [79].

Activities of daily living

ADLs recognition is another research field that may widely benefit from RGB-D data top view configuration. In this field the application range goes from high reliability fall detection to occlusion free HBU at home for elders in Ambient Assisted Living (AAL) environments. All these applications have relevant outcomes form the current research with the ability to identify users while performing tracking, interaction analysis, or HBU. Furthermore, all these application scenarios, can gather data using low cost sensors and processing units.

An example is the system for real-time human tracking and predefined human gestures detection that uses depth data acquired from Kinect sensor installed right above the detection region described in [8]. The tracking part is based on fitting an articulated human body model to obtained data using particle filter framework and specifically defined constraints which originate in physiological properties of the human body. The gesture recognition part has used the timed automaton conforming to the human body poses and regarding tolerances of the joints positions and time constraints.

In [12], a method to detect human body parts in depth images that is based on an active learning strategy is presented. The goal is to build an accurate classifier using a reduced number of labelled samples in order to minimize the training computational cost as well as the image labelling cost. The authors have validated the approach on two different scenarios: the detection of human heads of people lying in a bed and the detection of human heads from a ceiling camera.

The work proposed in [68] describes a feature for activity recognition from vertical top-view depth image sequences. The approach performance were verified on Top-View 3D Daily Activity Dataset.

For advanced analysis of human behaviours, the authors of [62] have developed a highly-integrated system. The video framework exploits vertical RGB-D sensors for people tracking, interaction analysis, and users activities detection in domestic scenarios. The depth information has been used to remove the effect of the appearance variation and to evaluate users activities inside the home and in front of the fixtures. In addition, group interactions have been monitored and analysed. The audio framework has recognised voice commands by continuously monitoring the acoustic home environment.

As previously stated, another important issue to monitor and evaluate during the people tracking is the fall detection, as reported for example in [65, 51, 52, 34]. The solutions implemented in these papers with RGB-D camera in a top-view configuration are suitable and affordable for this aim.

An automated RGB-D video analysis system that recognises human ADLs activities, related to classical daily actions is described in [64]. The main goal is to predict the probability of an analysed subject action. Thus, abnormal behaviours can be detected. The activity detection and recognition is performed using an affordable RGB-D camera. Action sequence recognition is then handled using a discriminative Hidden Markov Model (HMM).

2.4.1. Datasets

The most relevant available datasets with RGB-D data installed in a top-view configuration are listed below.

TST Fall detection dataset v1[3][34]

It stores depth frames collected using Microsoft Kinect v1 in top-view configuration. Four volunteers, aged between $26-27$ years and height in $1.62-1.78m$, have been recruited for a total number of 20 tests. The dataset is separated in two main groups: Group A (test 1-10): two or more people walk in the monitored area; Group B (test 11-20): a person performs some falls in the covered area (figure 2.3).

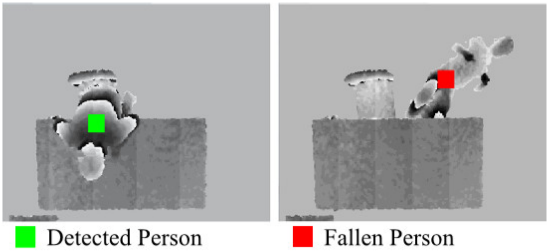


Figure 2.3.: ST Fall detection dataset v1. Image taken from [34]

TST Intake Monitoring dataset v1[4][33]

It is composed of food intake movements, recorded with Kinect V1, simulated by 35 volunteers for a total of 48 tests. The device is located on the ceiling at a $3m$ distance from the floor. The people involved in the tests are aged between 22 and 39 years, with different height and build (figure 2.4).

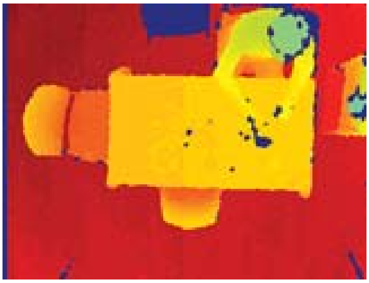


Figure 2.4.: TST Intake Monitoring dataset v1. Image taken from [33]

[3] http://www.tlc.dii.univpm.it/blog/databases4kinect#IDFall1
[4] http://www.tlc.dii.univpm.it/blog/databases4kinect#IDFood

UR Fall Detection Dataset[5][52]

It contains 70 (30 falls + 40 activities of daily living) sequences. Fall events are recorded with 2 Microsoft Kinect cameras (parallel to the floor and ceiling mounted) and corresponding accelerometric data (figure 2.5).

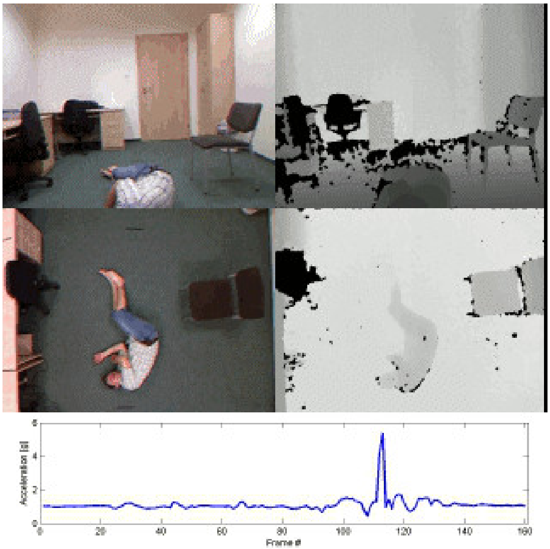


Figure 2.5.: UR Fall Detection Dataset. Image taken from [52]

Depthvisdoor[6][46]

It is a database of images captured by a single and stationary Kinect camera covering, from a top view, the entrance of a classroom at the University of Las Palmas de Gran Canaria. Two sessions have been recorded per day in 3 different days, with a one week gap every two recording sessions. For each of the 6 recordings, the Kinect sensor is located roughly at a similar location, approximately $2.7m$ height, looking at the scenario floor (figure 2.6). The illumination conditions change from one day to another and even within the same day due to the two hours of difference between the start and the end of the class, and the sensor makes use of its auto adjustment. The database has been used to develop, test, and evaluate people counting, description and Re-id algorithms.

[5] `http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html`

[6] `http://berlioz.dis.ulpgc.es/roc-siani/descargas-en/depthvisdoor-database-1`

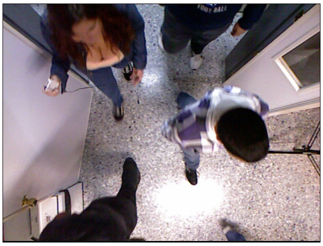
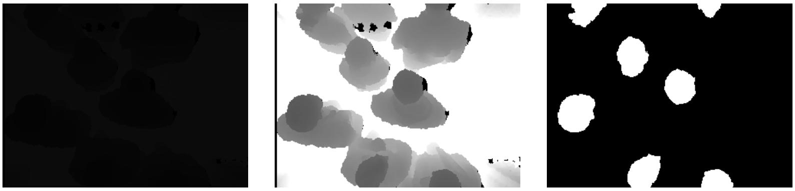


Figure 2.6.: Depthvisdoor Dataset.

TVHeads Dataset[7]

The Top-View Heads (TVHeads) dataset contains depth images of people from top-view configuration. In particular, the purpose of this dataset is to localize the heads of people who are present below the camera. It contains a total of 1815 depth images (16 bit) with a dimension of 320 × 240 pixels. Furthermore, after an image preprocessing phase, the depth images are also converted, with an appropriate scaling, in order to obtain an images (8 bit) where the heads silhouette is highlighted by improving image contrast and brightness. The ground truth was manually labelled by 6 human annotators. Figure 2.7 shows an example of a dataset instance that includes the three images described above.



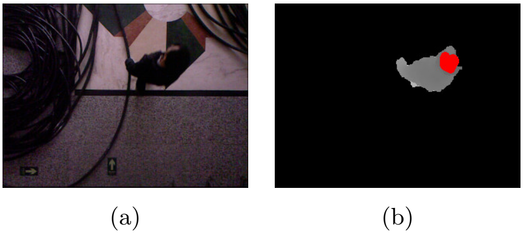
(a) 16 bit Depth image. (b) 8 bit Depth image. (c) Ground truth.

Figure 2.7.: TVHeads Dataset.

CBSR[8]

The dataset includes a total of 3884 images with 6094 heads. It contains depth images after background subtraction and in the groundtruth the heads are manually painted as red colour.

[7] `http://vrai.dii.univpm.it/tvheads-dataset`
[8] `http://www.cbsr.ia.ac.cn/users/xczhang/HeadData-CBSR.zip`



(a) (b)

Figure 2.8.: CBSR Dataset.

TVPR Dataset[9][66]

The 100 people of TVPR were recorded in 23 registration session. The recording time for the session and the number of persons of that session are reported in the following table. Each of the 23 folders contains the video of one registration session. Acquisitions have been performed in 8 days and the total recording time is about 2000 seconds. Registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling. More details will be provided in the following Chapter.

[9] `http://vrai.dii.univpm.it/re-id-dataset`

Chapter 3.

RGB-D data for top-view HBU: algorithms

This Chapter describes several solutions and main contributions for people detection using RGB-D data from top-view configuration. The algorithms are based on CV techniques and these will be used in the Chapter 4 for every use case. The problem of people detection has been simplified in order to find only the head of each subject. From top-view configuration, the head is the part of body that hardly has contact with objects and people. This facilitates the tracking procedure.

In section 3.2 are illustrated two important image processing algorithms, while in section 3.3 different semantic segmentation approaches based on deep learning techniques are presented. The most important metrics used for segmentation problems are reported in section 3.1.

3.1. Adopted metrics

Typically, to measure the segmentation accuracy and performance, different types of metrics are used. In particular, in this section, a two stages process is adopted: the first metrics measure how much the system is able to separate the heads from the background, whereas the second metric measures the ability of the system to correctly classify the heads.

One of the first metrics is the Jaccard index, also known as Intersection over Union (IoU), measures similarity between finite sample sets, and defined as the size of the intersection divided by the size of the union of the sample sets:

$$IoU = \frac{true_{pos}}{true_{pos} + false_{pos} + false_{neg}} \tag{3.1}$$

Another metric is the Sørensen–Dice index [26], also called the overlap index, is the most used metric in semantic segmentation, and is computed as:

$$Dice = \frac{2 \cdot true_{pos}}{2 \cdot true_{pos} + false_{pos} + false_{neg}} \tag{3.2}$$

where the positive class is the heads and the negative is all the rest.

The second set of metrics is composed by the average accuracy, precision, recall, and $f1$ score averaged across all the test images. These metrics are evaluated just on the heads pixels. The metrics are thus computed as:

$$accuracy = \frac{true_{pos} + true_{neg}}{true_{pos} + true_{neg} + false_{pos} + false_{neg}} \tag{3.3}$$

$$precision = \frac{true_{pos}}{true_{pos} + false_{pos}} \tag{3.4}$$

$$recall = \frac{true_{pos}}{true_{pos} + false_{neg}} \tag{3.5}$$

$$f1\,score = \frac{2 \times precision \times recall}{precision + recall} \tag{3.6}$$

3.2. Image processing approaches

In this section are reported two innovative algorithms based on image processing techniques using RGB-D data in top-view configuration.

3.2.1. Multi level segmentation

Multi level segmentation algorithm is explained in detail in the pseudo-code Algorithm 1. The MultiLevelSegm function has in input the foreground image ($f(x,y)$). First of all, FindPointMax function calculates the highest point of whole image (max) and its coordinates ($point_{max}$). In line 3, the *level* counter assumes the *threshold* value, that is a fixed value corresponding to average height of a human head (we adopted the value $10cm$). So, the number of segmentations is strictly related to the height of the tallest person.

The output condition of the while loop is verified when the segmentation level becomes negative (above the floor). In line 5 there is a segmentation function that yields in output a binary image with blobs representative of moving objects that are above the segmentation level ($max - level$). This binary image is the input of FindContours, an OpenCV function that returns a vector of points for each blob. Then, the FilterContours function deletes noise (blobs with a little dimension and/or a bad shape).

The for loop from line 8 to line 14 inserts in the vector *points* the highest point/depth value (FindPointMax function) of each blob identified by means of the FilterMask function. Finally, MultiLevelSegm function returns a vector with all maximum local points. The length of this vector indicates the number of people that are in the image.

Algorithm 1 Multi level segmentation algorithm

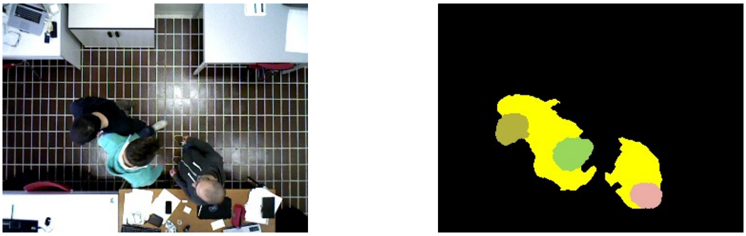
```
function multiLevelSegm($f(x, y)$)
    $(max, point_{max}) = $ findPointMax($f(x, y)$)
    $level = threshold$
    while $(max - level) > 0$ do
        $f_{level}(x, y) = f(x, y) > (max - level)$
        $contours = $ findContours($f_{level}(x, y)$)
        filterContours($contours$)
        for each contour $i \in contours$ do
            $f_{mask}(x, y) = $ filterMask($f_{level}(x, y), i$)
            $v_{max}, p_{max} = $ findPointMax($f_{mask}(x, y)$)
            if $p_{max} \notin points$ then
                $points$.pushBack($p_{max}$)
            end if
        end for
        $level = level + threshold$
    end while
    return $points$
end function
```

The multi level Segmentation algorithm overcomes the limitations of the binary segmentation method proposed in [61] in case of collisions among people. In fact, using a single-level segmentation, in case of a collision, two people become a single blob (person), without distinguishing between head and shoulders of the person. By using this approach, when a collision occurs, even if two people are identified with a single blob, the head of each person is anyway detected, becoming the discriminant element. Figure 3.1 highlights the head of each person obtained by the multi level segmentation algorithm: different colours highlight the head of a person detected by the camera. In case of collisions (Figures 3.1a and 3.1b), the yellow blob contains two people and both heads are detected.

3.2.2. Water-Filling

A further algorithm for people detection, using an RGB-D sensor in vertical position, is proposed by Zhang *et al.* in [122] and it is called “Water Filling”. In this paragraph an improvement of this algorithm is suggested.

It finds, in a depth image, the local minimum regions simulating the rain and the flooding of ground. According to an uniform distribution, it simulates the rain with some raindrop. The algorithm moves the raindrops towards the local



(a) RGB image.

(b) Segmented image.

Figure 3.1.: Multi level segmentation algorithm. Head recognition (3.1b): different colours of the blob highlight the head of the people detected in the scene (3.1a).

minimum points, but if a point is wet, it wets the point of the higher level. Then, puddles are formed because the water flows to some local minimum regions. It computes the contour lines considering the distance from the local minimum as a function of the total raindrops.

It is possible to consider the depth image as a function $f(x,y)$ that can be non-derivable or even discontinuous, due to the noise of depth sensor. Finding people in depth image equals to finding local minimum regions in f. Mathematically, the problem can be defined as finding the region A and N that satisfy the following constraint:

$$E_A(f(x,y)) + \eta \leq E_{N\backslash A}(f(x,y)) \tag{3.7}$$

where $A \in N$, A is the local region and N is its neighbourhood, both can be of arbitrary shape, $E(\cdot)$ is an operation to pool the depth information in the region to a real value that reflects the total depth information in the region. η is a pre-defined threshold to ensure that depth in A should lower than $N\backslash A$ with a margin.

In order to solve effectively the problem and be robust to noise, can be mathematically defined an additional measure function $g(x,y)$ as:

Definition 1. *$g(x,y)$ is a measure function of $f(x,y)$ if and only if $\exists\, \epsilon > 0, \forall\,(x_1,y_1),(x_2,y_2)$, s.t. $\left\|(x_1-x_2)^2+(y_1-y_2)^2\right\| < \epsilon$, if $f(x_1,y_1) \leq f(x_2,y_2)$*

$$f(x_1,y_1) + g(x_1,y_1) \leq f(x_2,y_2) + g(x_2,y_2) \tag{3.8}$$

$$g(x_1,y_1) \geq g(x_2,y_2)$$

$$g(x_1,y_1) \geq 0,\ g(x_2,y_2) \geq 0$$

The form of $g(x,y)$ can be trivial, for example a zero function. The use of $g(x,y)$ it allows to infer the $f(x,y)$.

In this context, it is not important to get a general solution of $g(x, y)$, but proper non-trivial form it is acceptable.

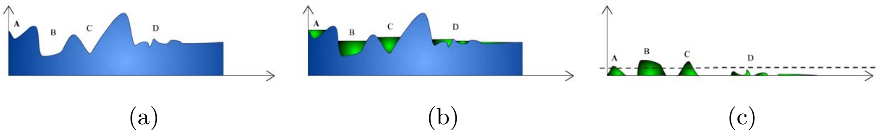
The form of function $f(x, y)$ can be seen as a land with humps and hollows. The raindrop in the hump will flow directly to the neighbourhood hollow under force of gravity. Little by little, the hollow region will gather a lot of raindrops. The function $g(x, y)$ reflects the quantity of raindrop at point (x, y). After the rain stops, the regions with a lot of rain drop can be classified as a hollow.

The algorithm 2 is an improved version of original proposed in [122]. In particular, the drops are chosen according to the segmentation of foreground image (line 4). This procedure improves the execution time of the algorithm.

Algorithm 2 Water Filling algorithm

```
1:  function WaterFilling($f(x,y)$, $T$, $K$)
2:      $g(x,y) = 0$
3:      $M, N = size(f(x,y))$
4:      $fg(x,y) = (bg(x,y) - f(x,y)) > T$
5:      for $k = 1 : K$ do
6:          $x = rand(1,M)$, $y = rand(1,N)$ with $(x,y) \in fg(x,y)$
7:          while $True$ do
8:              $d(x_n,y_n) = f(x_n,y_n) + g(x_n,y_n) - (f(x,y) + g(x,y))$ where $(x_n,y_n)$ is the neighbourhood of $(x,y)$.
9:              $(x^*,y^*) = arg \min d(x_n,y_n)$
10:             if $d(x^*,y^*) < 0$ then
11:                 $x = x^*$, $y = y^*$
12:             else
13:                 $g(x,y) = g(x,y) + 1$
14:                 $break$
15:             end if
16:         end while
17:     end for
18:     return $g(x,y) > T$
19: end function
```

The total number of raindrops is $K = tMN$, where t is usually set to be 100. At every loop (line 5), (x, y) is randomly generated through a discrete uniform distribution (line 6). If there is a point (x^*, y^*) in the neighbourhood of (x, y) that satisfies Equation 3.8 then the raindrop in (x, y) flows towards to (x^*, y^*) and restart the loop until a local minimum is reached. When this is reached, the measure function $g(\cdot)$ is increased (line 13). After all the K raindrops find their stable place, measure function $g(x, y)$ is calculated and, by applying a threshold T, it is possible to extract the heads of people that pass under the RGB-D sensor (line 18).



(a) (b) (c)

Figure 3.2.: The main characteristics water filling algorithm. A, B, C correspond to three people respectively and D is a noise region (figure 3.2a). Region A has smaller scale compared with B and C, and the absolute height of A is larger than noise region D. After the water filling process (figure 3.2b), the measure function $g(x, y)$ which reflects the property of $f(x, y)$ is obtained (figure 3.2c). Finally the people are detected by a threshold operation on measure function $g(x, y)$. Images taken from [122]

3.2.3. Results and performance

In order to test the performance of multi level segmentation and water filling algorithms, a restricted part of CBSR Dataset is used. Table 3.1 shows the results of algorithms in term of precision, recall and, f1-score.

The algorithms reach high values of performances, but often fail when the heads are along the edge of image. Multi level segmentation algorithm looks more accurate than water filling algorithm.

Table 3.1.: Image processing algorithms performances.

Algorithm	**Precision**	**Recall**	**F1-Score**
Multi level segmentation	0.9390	0.9872	0.9625
Water filling	0.9365	0.7564	0.8369

3.3. Semantic segmentation with deep learning approaches

Nowadays, one of the key problems in the field of CV is the semantic segmentation that is applied to 2D images, video, and even 3D data. Looking at the big picture, semantic segmentation is one of the high-level task that paves the way towards complete scene understanding.

Scene understanding started with the goal of building machines that can see like humans to infer general principles and current situations from imagery, but it has become much broader than that. Applications such as image search engines, autonomous driving, computational photography, vision for graphics, human machine interaction, were unanticipated and other applications keep arising as scene understanding technology develops [32]. As a core problem of high level CV, while it has enjoyed some great success in the past 50 years, a lot more is required to reach a complete understanding of visual scenes.

In the past, such a problem has been addressed using different traditional CV and machine learning techniques. Despite the popularity of those kind of methods, the deep learning marked a significant change so that many CV problems are being tackled using deep architectures, usually convolutional neural networks, which are surpassing other approaches by a large margin in terms of accuracy and sometimes even efficiency.

This section presents a particular case study describing five approaches from literature based on Convolutional Neural Network (CNN) architectures and implementation methods for semantic segmentation. Furthermore, a novel approach with better performances is presented.

3.3.1. U-Net

U-Net architecture proposed in [99] is shown in figure 3.3. It is composed of two main parts:

- *contracting* path (left side);
- *expansive* path (right side).

The first path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3 × 3 convolutions (unpadded convolutions), each followed by a Rectified Linear Unit (ReLU) and a 2 × 2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. Every step in the expansive path consists of an upsampling of the feature map followed by a 2 × 2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3 × 3 convolutions, each followed by a ReLU. At the final layer a 1 × 1 convolution is used to map each 32-component feature vector to the desired number of classes.

Similarly, the authors of [98] revisited the classic U-Net by removing two levels of max polling and changing the ReLU activation function with a LeakyReLU (figure 3.4).

Another U-Net architecture is proposed in this Thesis. The structure remains largely the same, but some changes are made at the end of each layer. In particular a batch normalisation is added after the first ReLU activation function and after each max polling and upsampling functions (figure 3.5).

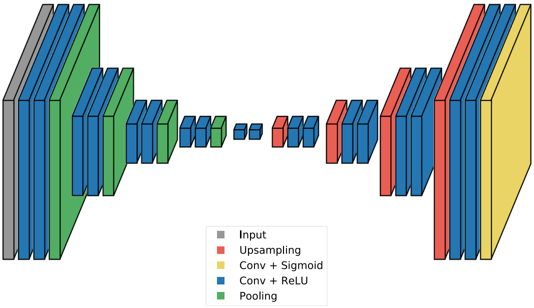


Figure 3.3.: U-Net architecture.

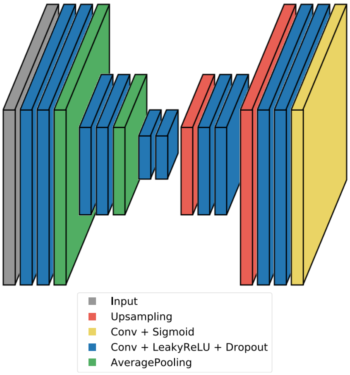


Figure 3.4.: U-Net2 architecture.

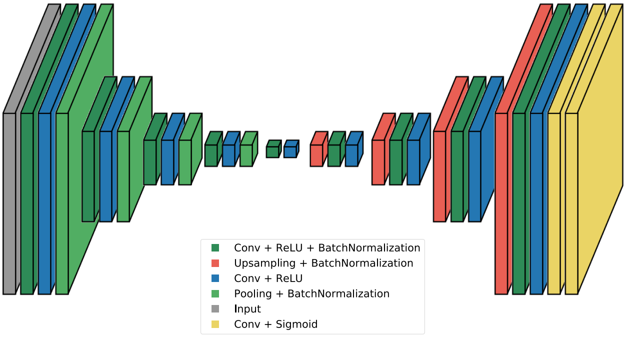

Conv + ReLU + BatchNormalization
Upsampling + BatchNormalization
Conv + ReLU
Pooling + BatchNormalization
Input
Conv + Sigmoid

Figure 3.5.: U-Net3 architecture.

3.3.2. SegNet

SegNet, presented by Vijay *et al.* in [3], is shown in Figure 3.6. The architecture consists of a sequence of non-linear processing layers (encoders) and a corresponding set of decoders followed by a pixelwise classifier. Typically, each encoder consists of one or more convolutional layers with batch normalisation and a ReLU non-linearity, followed by non-overlapping max pooling and subsampling. The sparse encoding, due to the pooling process, is upsampled in the decoder using the max pooling indices in the encoding sequence. One key ingredient of the SegNet is the use of max pooling indices in the decoders to perform upsampling of low resolution feature maps. This has the important advantages of retaining high frequency details in the segmented images and also reducing the total number of trainable parameters in the decoders. The entire architecture can be trained end-to-end using stochastic gradient descent. The raw SegNet predictions tend to be smooth even without a Conditional Random Fields (CRF) based post-processing.

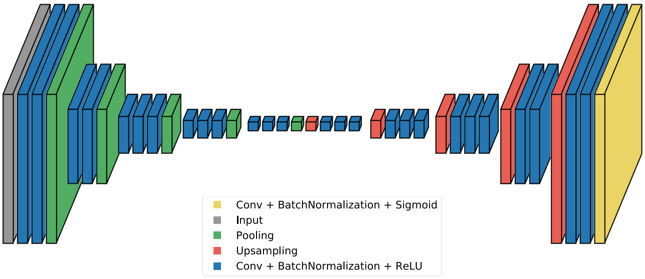

🟨 Conv + BatchNormalization + Sigmoid
⬜ Input
🟩 Pooling
🟥 Upsampling
🟦 Conv + BatchNormalization + ReLU

Figure 3.6.: SegNet.

3.3.3. ResNet

He *et al.* in [44] observed that deepening traditional feedforward networks often results in an increased training loss. In theory, however, the training loss of a shallow network should be an upper bound on the training loss of a corresponding deep network. This is due to the fact that increasing the depth by adding layers strictly increases the expressive power of the model. A deep network can express all functions that the original shallow network can express by using identity mappings for the added layers. Hence a deep network should perform at least as well as the shallower model on the training data.

The violation of this principle implied that current training algorithms have difficulties in optimizing very deep traditional feedforward networks. He *et al.* proposed residual networks that exhibit significantly improved training characteristics, allowing network depths that were previously unattainable. A

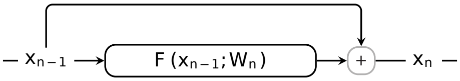


Figure 3.7.: Residual Unit.

ResNet is composed of a sequence of residual units shown in Figure 3.7. The output x_n of the $n-th$ RU in a ResNet is computed as:

$$x_n = x_{n-1} + F(x_{n-1}; W_n) \tag{3.9}$$

where $F(x_{n-1}; W_n)$ is the residual, which is parametrized by W_n. In this way, instead of computing the output x_n directly, F only computes a residual that is added to the input x_{n-1}. This design can be referred as skip connection, since there is a connection from the input x_{n-1} to the output x_n that skips the actual computation F. It has been empirically observed that ResNets have superior training properties over traditional feedforward networks. This can be explained by an improved gradient flow within the network.

3.3.4. FractalNet

Fractal network is introduced by Larsson *et al.* in [59]. Let C denotes the index of a truncated fractal $f_C(\cdot)$ (i.e., a few stacked layers) and the base case of a truncated fractal is a single convolution:

$$f_1(z) = conv(z)$$

According to the expansion rule:

$$z' = conv(z)$$

$$f_{C+1}(z) = conv(conv(z') \oplus f_C(z'))$$

can be defined recursively the successive fractals, where $\oplus$ is a join operation and $conv(\cdot)$ is a convolution operator. Two blobs are merged by the join operation $\oplus$. As these two blobs contain features from different visual levels, joining them can enhance the discrimination capability of our network. Generally, this operation can be summation, maximization and concatenation.

In order to enlarge the receptive field and enclose more contextual information, downsampling and upsampling operations are added in the above expansion rule. In particular, a max pooling with a stride of 2 and a deconvolution also with a stride of 2 are added. After the downsampling operation, the receptive field of a fractal becomes broader. When combining different receptive fields through the join operation, the network can harness multi-scale visual cues and promote itself in discriminating.

The Fractal Net used in this section is depicted in figure 3.8.

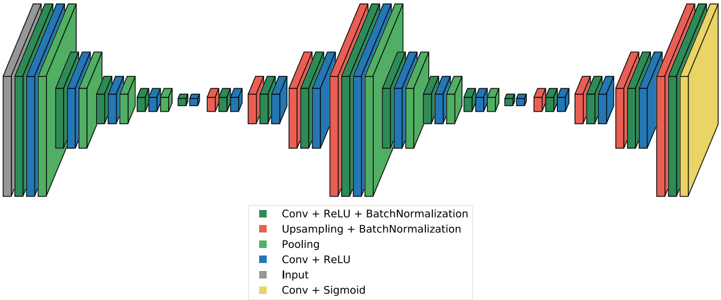

🟩 Conv + ReLU + BatchNormalization
🟥 Upsampling + BatchNormalization
🟩 Pooling
🟦 Conv + ReLU
◼ Input
🟨 Conv + Sigmoid

Figure 3.8.: FractalNet.

3.3.5. Results and performance

This paragraph presents the results from several CNN architectures. Graphs and tables show the cost and performance during training and validation for heads detection task on *TVHeads dataset*. Each CNN implementation is trained with two types of depth images:

- *16-bit*: original images acquired by depth sensor;
- *8-bit*: scaled images in order to highlight the heads silhouette, improving the images contrast and brightness.

The general procedure for training neural network based models is to take the dataset and split it into three parts: training, validation, and test. In this case, train, test and validation are chosen respectively to learn model parameters. Once this is complete, the best model is also evaluated over the never before seen test set.

In following experiments, 70%, 10% and 20% of dataset are chosen respectively for train, test, and validation. Furthermore, different combinations of hyperparameters are tested, but a learning rate equal to 0.001 and an Adam optimization algorithm have been used.

Quantitative Evaluation

Semantic segmentation performances are divided into two different tables. Tables 3.2 shows Jaccard and Dice indices for training and for validation respectively. While, Table 3.3 reported the results in term of accuracy, precision, recall and f1-score. Both tables refer to a learning process conducted during 200 epochs.

Table 3.2.: Jaccard and Dice indices of different CNN architectures.

Net	Bit	Jaccard *Train*	Jaccard *Validation*	Dice *Train*	Dice *Validation*
Fractal [59]	8	0.960464	0.948000	0.979833	0.973306
	16	0.961636	0.947762	0.980443	0.973180
U-Net [99]	8	0.896804	0.869399	0.945595	0.930138
	16	0.894410	0.869487	0.944262	0.930188
U-Net2 [98]	8	0.923823	0.939086	0.960403	0.968586
	16	0.923537	0.938208	0.960249	0.968119
U-Net3	8	0.962520	0.931355	0.980902	0.964458
	16	0.961540	0.929924	0.980393	0.963690
SegNet [3]	8	0.884182	0.823731	0.938531	0.903347
	16	0.884162	0.827745	0.938520	0.905756
ResNet [44]	8	0.932160	0.856337	0.964889	0.922609
	16	0.933436	0.848240	0.965572	0.917889

In Table 3.2, the best CNN architecture is the U-Net3 8-bit version. Indeed, Jaccard index reaches a value equal to 0.962520. The second best is Fractal Net 16-bit version also obtaining higher values as regards validation performances. For each network is highlighted in red colour the best bit version model.

Table 3.3.: Semantic segmentation results of different ConvNet architectures.

Net	Bit	Accuracy	Precision	Recall	F1-Score
Fractal [59]	8	0.994414	0.991400	0.993120	0.992235
	16	0.994437	0.992667	0.993297	0.992970
U-Net [99]	8	0.992662	0.946475	0.950483	0.948408
	16	0.992569	0.945083	0.948957	0.946938
U-Net2 [98]	8	0.993156	0.970013	0.969206	0.969568
	16	0.993165	0.967884	0.970557	0.969123
U-Net3	8	0.994572	0.990451	0.990387	0.990419
	16	0.994559	0.989382	0.989411	0.989396
SegNet [3]	8	0.992683	0.946304	0.953136	0.949625
	16	0.992699	0.946237	0.953342	0.949658
ResNet [44]	8	0.993789	0.968399	0.968374	0.968359
	16	0.993819	0.968765	0.969256	0.968992

As in the previous case, also in the Table 3.3 the best CNN architecture, in terms of accuracy, is U-Net3 8-bit version. While, Fractal Net 16-bit version exceeds slightly in precision, recall and f1-score metrics.

In Figure 3.9 is shown for each CNN the trend of Jaccard index during the fit procedure. It is easy to see how the Fractal Net and the ResNet reach high values immediately after a few epochs. Instead, the U-Net3 increases its value more slowly. The classic U-Net is always below all other networks.

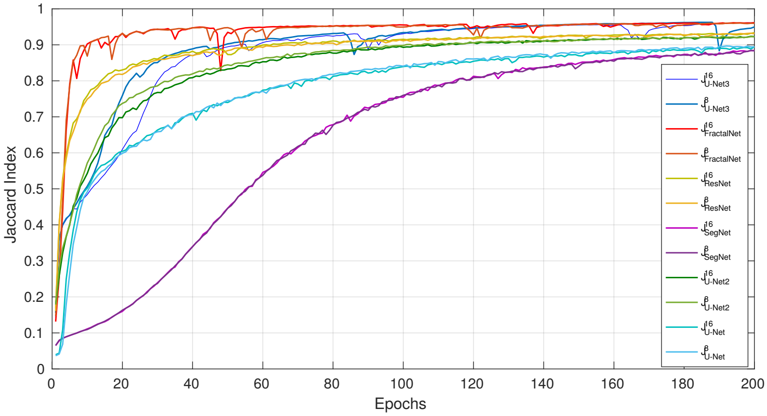


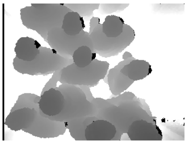
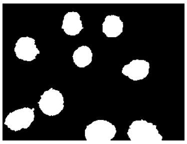
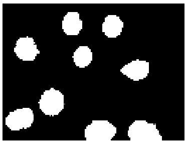
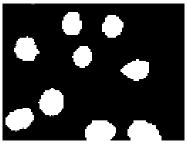
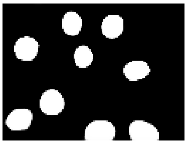
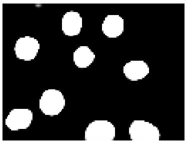
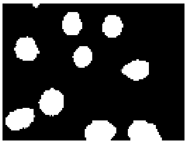
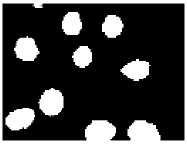
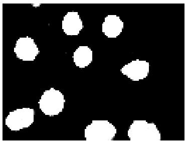
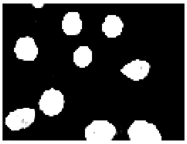
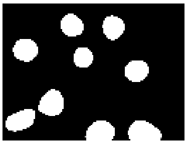
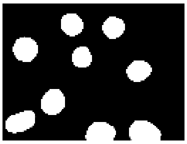
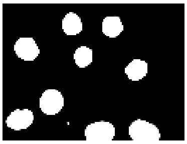
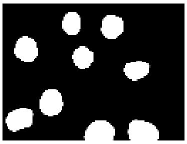
Figure 3.9.: Jaccard index trends.

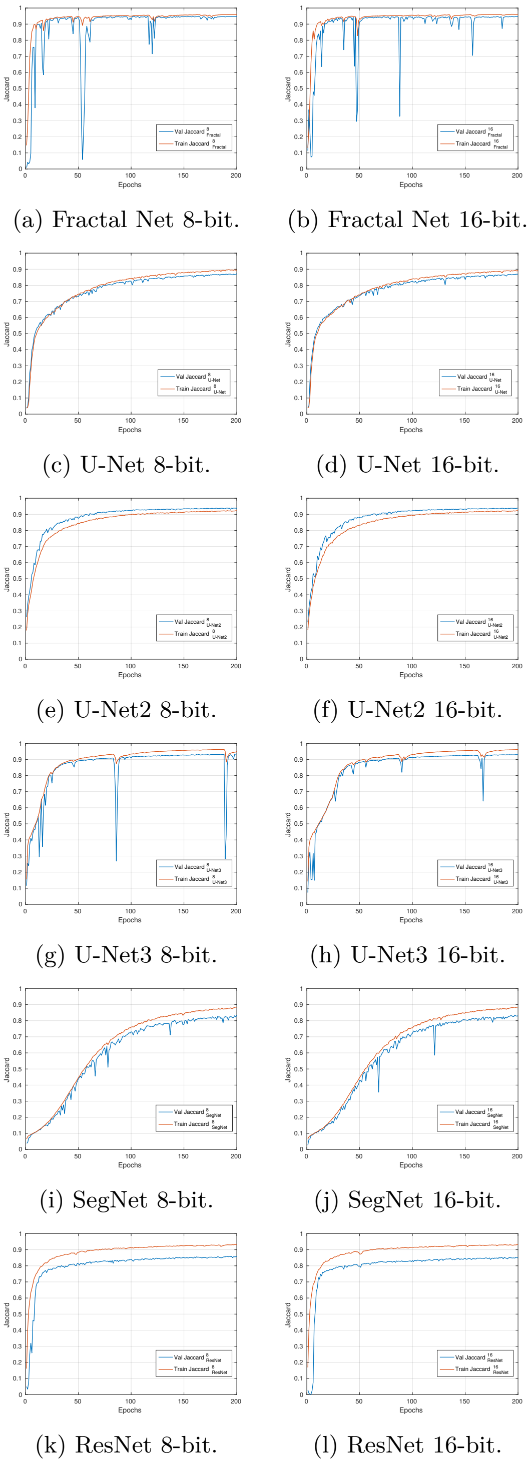
In a similar way, in Table 3.10, for each network is shown the Jaccard index trend also during the validation period.

Qualitative Evaluation

Table 3.4 presents qualitative semantic segmentation results. The table shows the predicted images for each architecture. The best results are obtained by using the U-Net3 (8 and 16 bit). While, typical U-Nets and SegNet provide shapes of heads more smooth and rounded.

Table 3.4.: Qualitative result of prediction.

	8-bit	**16-bit**	**Label**
			
FractalNet [59]			
U-Net [99]			
U-Net2 [98]			
U-Net3			
SegNet [3]			
ResNet [44]			



(a) Fractal Net 8-bit.

(b) Fractal Net 16-bit.

(c) U-Net 8-bit.

(d) U-Net 16-bit.

(e) U-Net2 8-bit.

(f) U-Net2 16-bit.

(g) U-Net3 8-bit.

(h) U-Net3 16-bit.

(i) SegNet 8-bit.

(j) SegNet 16-bit.

(k) ResNet 8-bit.

(l) ResNet 16-bit.

Figure 3.10.: Jaccard index results.

Chapter 4.

RGB-D data for top-view HBU: use cases and results

In this Chapter some cases about HBU of use are presented. The main topics of investigation have been the following: *video surveillance*, described through several applications in Re-id field; *intelligent retail environment*, where a novel shopper analytics system is presented; *activities of daily living* through the case study of an ad-hoc application for environmental monitoring.

4.1. Video surveillance

Re-id represents a valuable task in video surveillance scenarios, where long-term activities have to be modelled within a large and structured environment (e.g., airport, metro station).

In this context, a robust modelling of the entire body appearance of the individual is essential, because other classical biometric cues (face, gait) may not be available, due to sensors' scarce resolution or low frame-rate. Usually, it is assumed that individuals wear the same clothes between the different sightings. The model has to be invariant to pose, viewpoint, illumination changes, and occlusions: these challenges call for specific human-based solutions. For these reasons, in the next subsection a person Re-id approach using top-view RGB-D data is presented.

4.1.1. Re-identification

In the last decades, video analytics has rapidly evolved as autonomous understanding of events occurring in a scene monitored by multiple video cameras. One of the fundamental problems in video surveillance is the person Re-id, which is the process to determine if different instances or images of the same person, recorded in different moments, belong to the same subject. In every day life, this is done by humans without much effort. Our brains are trained to localise and detect people and later to properly re-identify them. In the recent years, this problem has gained a rapid increase in attention in both academic research communities and industrial laboratories.

Person Re-id has many important applications in video surveillance, because it saves human efforts on exhaustively searching for a person from large amounts of video sequences. Identification cameras are widely employed in most of public places like malls, office buildings, airports, stations, and museums. These cameras generally provide enhanced coverage and overlay large geospatial areas because they have non-overlapping fields-of-views. Huge amounts of video data, monitored in real time by law enforcement officers or used after the event for forensic purposes, are provided by these networks. An automated analysis of these data improves significantly the quality of monitoring, in addition to process the data faster [111].

The behaviour characterization of people in a scene and their long term activity can be possible using video analysis, which is required for high-level surveillance tasks in order to alert the security personnel.

Over the past years, in the field of object recognition a significant amount of research has been performed by comparing video sequences. Colour-based features of video sequences are usually described with the use of a set of key frames that characterize well a video sequence. The HSV colour histogram and the RGB colour histogram are robust against the perspective and the variability of resolution [39]. The clothing colour histograms taken over the head, trousers, and shirt regions, together with the approximated height of the person, have been used as discriminative features.

Recently, the person Re-id problem has received a considerable attention, and various reviews and surveys are available, pointing out different aspects of this topic [78]. Research works on person Re-id can be divided into two categories: feature-based and learning-based [114].

The use of anthropometric measures for Re-id was proposed for the first time in [73]. In this case, height was estimated from RGB cameras as a cue for associating tracks of individuals coming from non-overlapping views.

In [36], the authors proposed the use of local motion features to re-identify people across camera views. They obtained correspondence between body parts of different persons through space-time segmentation. On this body parts, colour and edge histograms are extracted. In this approach, person Re-id is performed by matching the body parts based on the features and correspondence.

Shape and appearance context, which computes the co-occurrence of shape words and visual words for person Re-id is proposed in [115]. Human body is partitioned into L parts with the shape context and a learned shape dictionary. Then, these parts are further segmented into M subregions by a spatial kernel. The histogram of visual words is extracted on each subregion. Consequently, for the person Re-id the $L \times M$ histograms are used as visual features.

In [7] the appearance of a pedestrian is represented by combining three kinds of features (sampled according to the symmetry and asymmetry axes obtained from silhouette segmentation): the weighted colour histograms, the maximally stable colour regions, and recurrent highly structured patches.

Another method to face the problem of person Re-id is learning discriminant models on low-level visual features. Adaboost is used to select an optimal ensemble of localized features for pedestrian recognition in [39]. The partial least squares methods is used to perform person Re-id in [101]. Instead, Prosser *et al.* [93] have used ranking SVM to learn the ranking model.

In last years, it is well-known the metric learning for person Re-id. A probabilistic relative distance comparison model has been proposed in [124]. It maximizes the probability that the distance between a pair of true match is smaller than the distance that between an incorrect match pair.

In [89], the authors investigate whether the Re-id accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. They also propose a dissimilarity-based framework for building and fusing the multimodal descriptors of pedestrian images for Re-id tasks, as an alternative to the widely used score-level fusion.

Several datasets used to test Re-id models are available: *VIPeR*[1], *iLIDS*,[2] *ETHZ*[3] and the more recent *CAVIAR4REID*[4]. These datasets cover many aspects of the person Re-id problem, such as shape deformation, occlusions, illumination changes, very low resolution images, image blurring, etc. [38]. Another Re-id dataset is proposed in [5]; this is composed by 79 people and four groups. Data are gathered using RGB-D technology, but are not suitable for my purposes as mentioned above in Table 4.1.

Recent literature about Re-id approaches is mostly focused on appearance-based models. Researchers have paid attention on interest points, structural information, and colour as principal appearance cues [23]. The introduction of RGB-D cameras provides affordable and additional rough depth information coupled with visual images, offering sufficient accuracy and resolution for indoor applications. Due to this fact, this camera has already been successfully applied in the retail field to uniquely identify customers and to analyse behaviours and interactions of shoppers [61].

In the next few paragraphs, a dataset of person Re-id that uses an RGB-D camera in a top-view configuration (TVPR dataset) is presented. An Asus Xtion Pro Live RGB-D camera has been used because it allows to acquire

[1] `https://vision.soe.ucsc.edu`

[2] `http://www.eecs.qmul.ac.uk`

[3] `https://data.vision.ee.ethz.ch/cvl/aess/dataset`

[4] `http://www.lorisbazzani.info/datasets`

Table 4.1.: Main motivations and possible applications of TVPR.

Research Challenges	Applications	Related works
Reliable and occlusion free people counting	Safety and security in crowded environments; people flow analysis; access control and counting	[122, 116, 54, 112, 16]
Interaction detection between people and environment	Intelligent retail environment shelf: Shopper Analytics; Ambient Assisted Living (AAL)	[61, 29, 79]
Fall detection, HBU	High reliability fall detection; occlusion free; HBU at home and AAL	[65, 52]

colour and depth information in an affordable and fast way. The camera is installed on the ceiling above the area to be analysed.

For Re-id evaluation, data of 100 people are collected, acquired across intervals of days and in different times. Each person walked with an average gait within the recording area in one direction, stopping for few seconds just below the camera, then it turned around and repeated the same route in the opposite direction, always stopping under the camera for a while. This choice is due to its greater suitability compared with a front view configuration, usually adopted for gesture recognition or even for video gaming. The top-view configuration reduces the problem of occlusions [65] and has the advantage of being privacy preserving, because the face is not recorded by the camera. Main motivations of our top-view dataset and some related applications/works are described in Table 4.1.

The process of extraction of a high number of significant features derived from both depth and colour information is presented. Among all possible features, we selected the nine features described in following sections as the most interesting ones. After analysing the effectiveness of each feature, selected 9 significant features for the Re-id process are collected. The set of features extracted by the colour and depth images is used to perform in future works the Re-id process.

Setup and acquisition

The 100 people were acquired in several days. The camera is installed on the ceiling of a laboratory at $4\,m$ above the floor and covers an area of $14.66\,m^2$ ($4.43\,m \times 3.31\,m$). The camera is positioned above the surface which has to be analysed (Figure 4.1).

The first step is the processing of the data acquired from the RGB-D camera. The camera captures depth and colour images, both with dimensions of 640×480 pixels, at a rate up to approximately $30\,fps$ and illuminates the

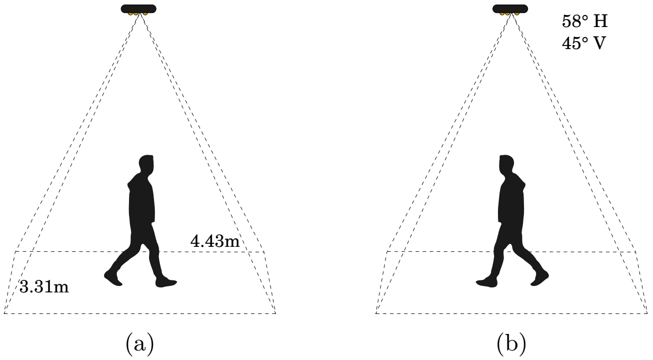


Figure 4.1.: System architecture.

scene/objects with structured light based on infrared patterns.

Seven out of the nine features selected are the *anthropometric features* extracted from the depth image:

- distance between floor and head, d_1;
- distance between floor and shoulders, d_2;
- area of head surface, d_3;
- head circumference, d_4;
- shoulders circumference, d_5;
- shoulders breadth, d_6;
- thoracic anteroposterior depth, d_7.

The remaining two *colour-based features* are acquired by the colour image. I also define *TVH*, *TVD* and *TVDH.*

- *TVH* is the colour descriptor:

$$TVH = \{H_h^p, H_o^p\} \tag{4.1}$$

- *TVD* is the depth descriptor:

$$TVD = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p\} \tag{4.2}$$

- Finally, *TVDH* is the signature of a person defined as:

$$TVDH = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p, H_h^p, H_o^p\} \tag{4.3}$$

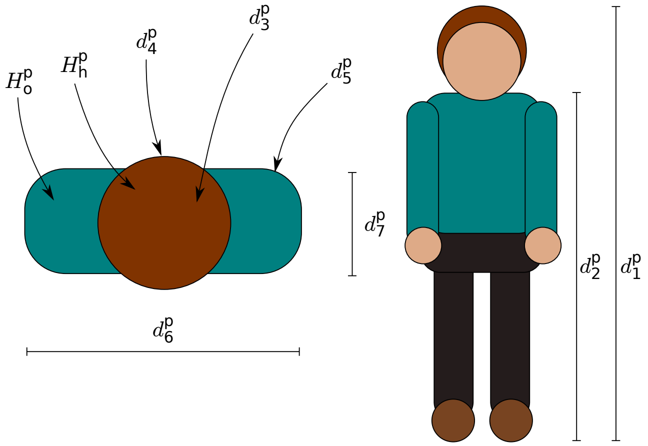


Figure 4.2.: Anthropometric and colour-based features.

Colour is an important visual attribute for both CV and human perception. It is one of the most widely used visual feature in image/video retrieval. To extract this two features we used HSV histograms. Local histograms have proven to be largely adopted and very effective. The signature of a person is also composed by two colour histograms computed for head/hairs and outerwear: H_h^p, H_o^p in 4.3, such as in [4], with $n = 10$ bin quantization, for both H channel and S channel.

Figure 4.2 depicts the set features considered: anthropometric and the colour based ones.

Results evaluation

The 100 people of dataset were acquired in 23 registration session. Each of the 23 folders contains the video of one registration sessions. The recording time $[s]$ for the session and the number of persons of that session are reported in Table 4.2. Acquisitions have been performed in 8 days and the total recording time is about 2000 seconds.

Registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling. Another big issue is environmental illumination. In each recording session, the illumination condition is not constant, because it varies in function of the different hours of the day and it also depends on natural illumination due to weather conditions. The video acquisitions, in our scenario, are depicted in Figure 4.3, which are examples of person registration respectively with sunlight and artificial light. Each person during a registration session walked with an average gait within the recording area in one direction, then it turned back and repeated the same route in the opposite direction. This methodology is used for a better split of TVPR in training set (the first passage

Table 4.2.: Time [s] of registration for each session and the number of people of that session.

Session	**Time [s]**	**# people**	**Session**	**Time [s]**	**# people**
g001	68.765	4	g013	102.283	6
g002	53.253	3	g014	92.028	5
g003	50.968	2	g015	126.446	6
g004	59.551	3	g016	86.197	4
g005	75.571	4	g017	95.817	5
g006	128.827	7	g018	57.903	3
g007	125.044	6	g019	82.908	5
g008	75.972	3	g020	87.228	4
g009	94.336	4	g021	42.624	2
g010	116.861	6	g022	68.394	3
g011	101.614	5	g023	56.966	3
g012	155.338	7			
			Total	**2004.894**	**100**

of the person under the camera) and testing set (when the person passed again under the camera).

The recruited people are aged between $19-36$ years: 43 females and 57 male; 86 with dark hair, 12 with light hair and 2 are hairless. Furthermore, of these people 55 have short hair, 43 have long hair. The subjects were recorded in their everyday clothing like T-shirts/sweatshirts/shirts, loose-fitting trousers, coats, scarves and hats. In particular, 18 subjects wore coats and 7 subjects wore scarves. All videos have fixed dimensions and a frame rate of about 30 fps. Videos are saved in native `.oni` files, but can be converted in any other format. Colour stream is available in a non compressed format.

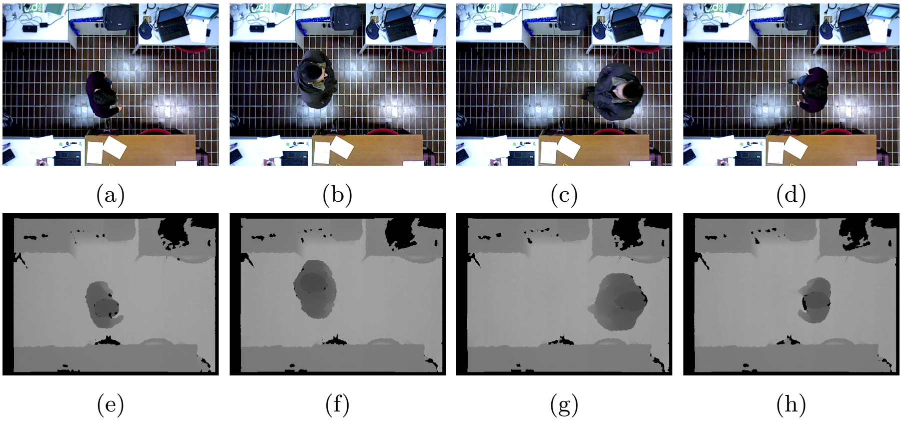
Figure 4.4 reports the histograms of each extracted anthropometric feature. Due to the dissimilarity of the analysed subjects a Gaussian curve is obtained from the data.

Performance validation

The Cumulative Matching Characteristic (CMC) curve represents the expectation of finding the correct match in the top n matches. It is equivalent of the Receiver Operating Characteristic (ROC) curve in detection problems. This performance metric evaluates recognition problems, by some assumptions about the distribution of appearances in a camera network. It is considered the primary measure of identification performance among biometric researchers.

As well-established in recognition and in Re-id tasks, for each testing item we ranked the training gallery elements using standard distance metrics. We examined the effects of 3 distance measures as the matching distance metrics: the $L1$ City block, the Euclidean Distance and the Cosine Distance.

To evaluate the dataset, performance results are reported in terms of recog-



(a) (b) (c) (d)

(e) (f) (g) (h)

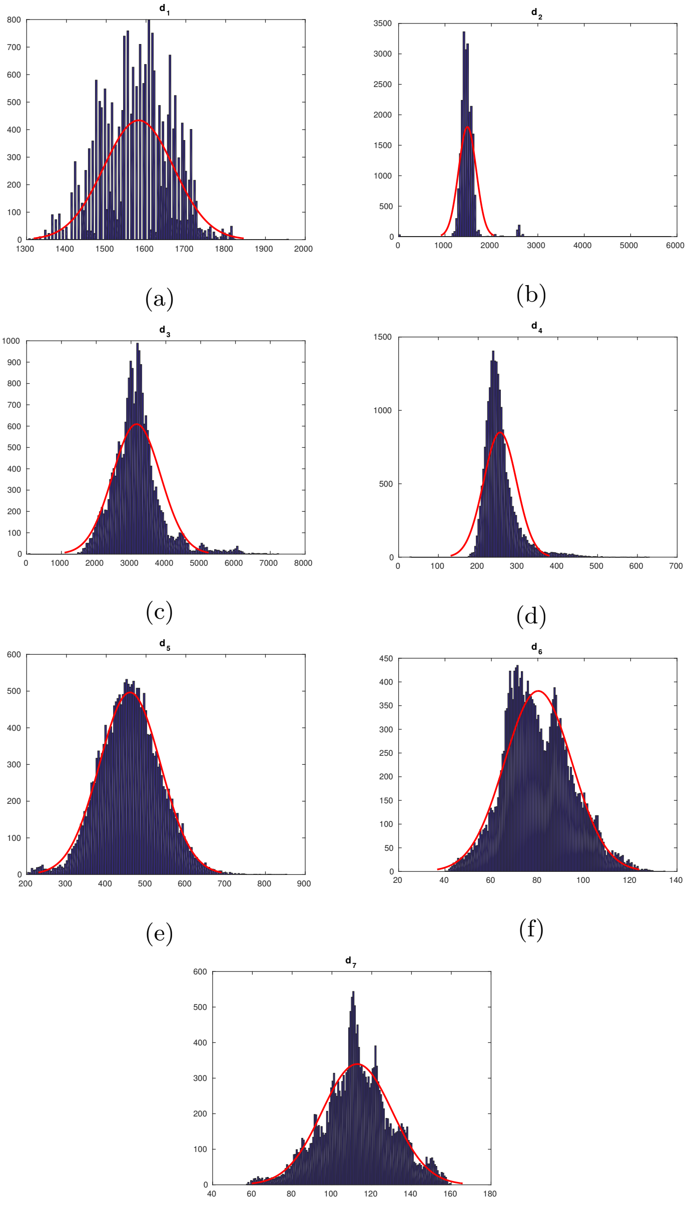
Figure 4.3.: Snapshots of a registration session of the recorded data, in an indoor scenario, with artificial light. People had to pass under the camera installed on the ceiling. The sequence 4.3a-4.3e, 4.3b-4.3f corresponds to the sequence 4.3d-4.3h, 4.3c-4.3g respectively training and testing set of the classes `8-9` for the registration session `g003`.

nition rate, using the CMC curves, illustrated in Figure 4.5. In particular, the horizontal axis is the rank of the matching score, the vertical axis is the probability of correct identification.

Considering the dataset, a comparison among *TVH* and *TVD* in terms of CMC curves are depicted, to compare the ranks returned by using these different descriptors.

Figure 4.5a provides the CMC obtained for *TVH*. Figure 4.5b represents the CMC obtained for *TVD*. We compare these results with the average obtained by *TVH* and *TVD*. The average CMC is displayed in Figure 4.5c.

It is observed that the best performance is achieved by the combination of descriptors. In Figure 4.5d, it can be seen that the combination of descriptors improve the results obtained by each of the descriptor separately. This result is due to the depth contribution that can be more informative. In fact, the depth outperform the colour, giving the best performance for rank values higher than 15 (Figure 4.5b). Its better performance suggests the importance and potential of this descriptor.



(a)

(b)

(c)

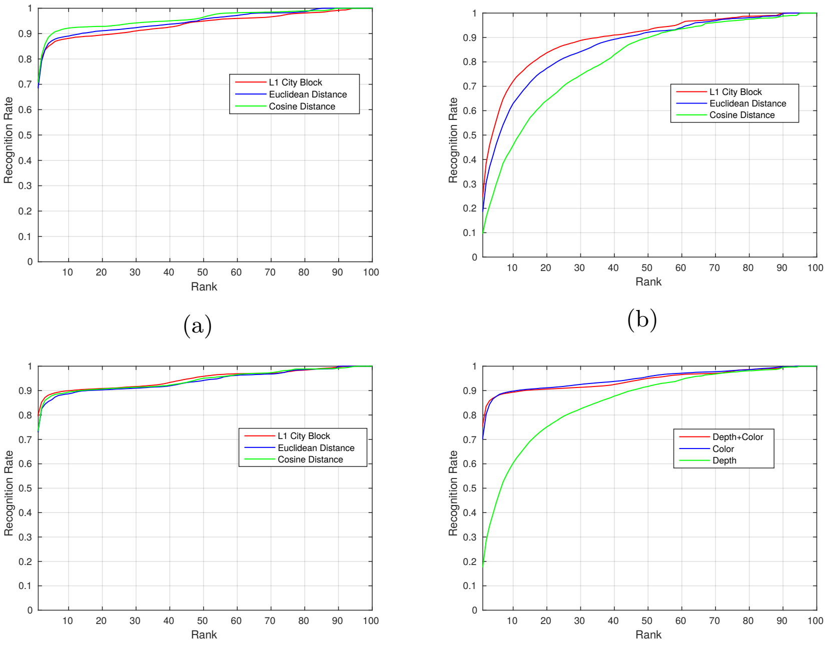
(d)

(e)

(f)

(g)

Figure 4.4.: Statistics histogram for each feature (4.4a d_1 distance between floor and head; 4.4b d_2 distance between floor and shoulders; 4.4c d_3 area of head surface; 4.4d d_4 Head circumference, 4.4e d_5 shoulders circumference, 4.4f d_6 shoulders breadth; 4.4g d_7 thoracic antero-posterior depth). The resultant Gaussian curve (in red) is due to the dissimilarity of the analysed subjects.



(a)

(b)

(c)

(d)

Figure 4.5.: The CMC curves obtained on TVPR Dataset.

4.2. Intelligent retail environment

In the field of IRE, numerous studies to investigate how shoppers behave inside a store and how businesses can change strategies to improve sales are emerging.

In order to analyse the buyer activity and to solve general aspects of these problems, techniques of artificial intelligence are used and, in particular, vision and image processing. In the next subsection an intelligent video system to monitoring the activity of customers is proposed.

4.2.1. Shopper behaviour analysis

In literature, there are several researches that study the behaviour of consumers in retail environments, for example, [55] and [94] and references therein. In particular, Puccinelli *et al.* [94] identified seven topic areas of consumer behavior research in retail environments: (*1*) goals, schema and information processing, (*2*) memory, (*3*) involvement, (*4*) attitudes, (*5*) affect, (*6*) atmospherics and (*7*) consumer attributions and choices. For each topic, they highlighted the most important issues necessary to be further investigated.

A common characteristic of all these studies is to do not use automated approaches for data acquisition and information retrieval. They mainly focus on consumer research and retailing from the social, psychological and marketing point of views. On the contrary, the potential of computing to improve all aspects of retail is firstly studied in deep in [58]. In particular, CV systems appear very useful in retail environments (as well as in other application fields), mainly for the huge amount of data and the possibility of an automatic data collection. Obviously, an increasing number of these applications [113], [81] have been and are possible thanks to the strengthening of information systems, the development of more stable and efficient vision algorithms and also the higher speed and the lower price of current hardware.

Focusing on CV approaches of consumer attributions and choices, Chandon *et al.* [18] and Strandvall [105] both used eye tracking methods for measuring the value of point-of-purchase. Määttä *et al.* [72] classified shopper motion into four behaviour classes, distinction if their movement is neutral or repetitive. Haritaoglu [42] examined the use of CV systems to estimate shoppers' attention to billboard or product promotions, counting the number of people and the time spent observing the display. Another approach for video-based extraction of customer movements at the point of sale is described in [57]: their human behaviour analysis is based on the measurement of the customer trajectories inside the store and on the time spent by each person in each zone of the store. Senior *et al.* [102] proposed a video analytics system for retail that counts the number of customers entering a store and monitors where they go within the store.

According to [82], newer video surveillance applications, not necessarily related to security issues, were developed for shopping, not only to identify anomalous activities, but also to identify people and to analyse consumer behaviour.

At the same time, other pervasive computing approaches were adopted to solve problems in retail environments. ConvenienceProbe [121] examines trajectory data offered by mobile phone users to identify retail trade areas. These data are critical information for planning outdoor advertisement, finding competing stores and determining the optimal store location. Another system developed for the retail store is “SmartStore” [48], which analyses the customer interest immediately, gathers the sensing data from large-scale area and attaches massive tiny sensors to shopping items.

This work focuses on the implementation of a software infrastructure coupled with a hardware technology to build a pervasive computing intelligent system for detecting and analysing the human behaviour in real retail stores.

By means of video cameras and CV algorithms, the pervasive system detects human motion and then describes human behaviour by quantitative parameters. More in detail, the main objective is to analyse the interactions between customers and products on the shelves. Therefore, the system detects and monitors people when they are in front of a shelf, using a distributed video sensor network. This allows us to better detail the activities of consumers when they stop in a zone of the store, e.g., the objects of the shelf that are touched by each person.

The installation of the system in several parts of the store provides large volumes of multidimensional data on which to perform statistics and deduce insights. The analysis of these data offers a unique possibility to better understand several crucial aspects of a retail ambient, e.g., the appealing of a product, a good positioning of different products on a shelf, the human traffic in front of each shelf.

Just because the sensor installation should be repeated in several zones of the shop to collect a significant large amount of data, we developed a system that can be easily scaled, from a single shelf installation to a large widespread grid of sensors.

The system is able to detect all the objects on the shelf that interact with the customer using only one RGB-D sensor for each shelf, while in [55] the activity recognition needs an RFID sensor mounted on each object to be performed.

In [63] has been developed a software infrastructure that is able to automatically detect, measure and store crucial information for a retail ambient. In particular, the pervasive system does not need to interact with customers to retrieve the desired information, as for the cases of the interactive display [106] or the mobile phone [41], where direct customer interactions with the systems are used to exchange information between shoppers and the retail ambient. The information are collected automatically through the CV algorithms that will be presented in 4.2.1.

Summarizing, this subsection presents a smart and low-cost embedded sensor network for IRE able to identify customers and to analyse their behaviour and shelf interactions. Major characteristics of this system are the general and easily scalable architecture really focused on the retail environment application and the very precise and reliable CV algorithms, which are able to run efficiently in low cost hardware and to collect automatically several relevant information.

IRE architecture and application requirements

The hierarchical architecture of the proposed pervasive retail environment is shown in Figure 4.6 along with the information provided at each abstraction layer. Sensor nodes, able to measure autonomously a part of the environment, are logically connected to the concept of shelf, multiple shelves are part of a store and, finally, several stores are part of a retailer chain. The general idea is based on several aggregation layers that provide to the system different information, from raw data to high level data analytics and insights.

At the single camera node level only raw data are available; a first data processing to provide interaction maps occurs at the shelf level. Multi camera analysis is also functional to perform flow comparisons in different areas of the store. At the top level general insights, store comparisons, store optimizations and re-design can be performed by retailers.

The functional requirements of the system, concerning what the system is able to do, its expected behaviour and which are its input/output functions, are:

1. counting the number visitors;
2. storing the path of each person detected by the camera during the visit;
3. storing all type of interactions with the shelf:
   a) who performed the action;
   b) which product (SKU) was touched;
4. sending and saving data in a remote database available for statistical analysis;
5. from any remote location, controlling RGB-D camera parameters:
   a) video streaming visualization;
   b) redefinition of shelf area;
   c) software upgrade;

6. restoring data from a backup.

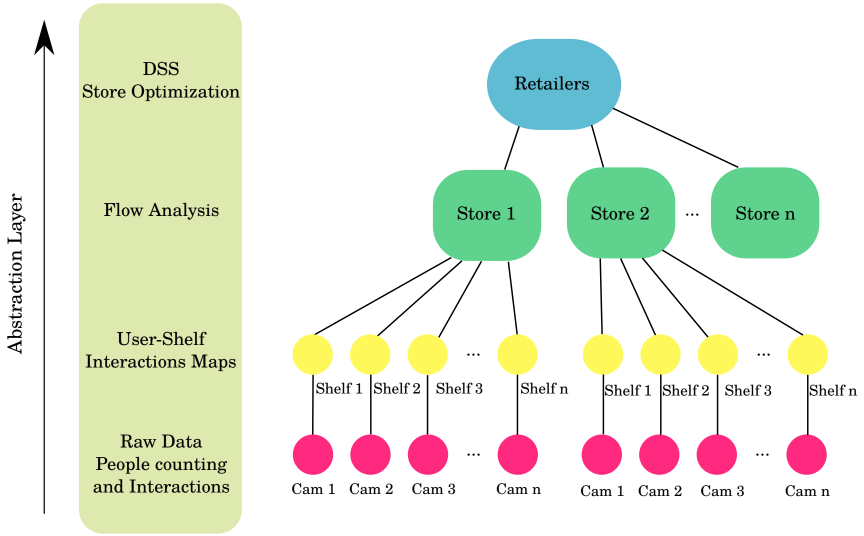


Figure 4.6.: The hierarchical architecture and the information provided at each abstraction layer of an intelligent retail environment.

Computer vision algorithms

The whole system can be seen as a big sensor network where each node is a micro system that analyses the consumer behaviour in front of a specific shelf of a store. Each node consists of an embedded system that includes an RGB-D sensor, ceiling mounted and looking to the scene from the top, and a software component to send the calculated information to the cloud. Adding a new node/shelf does not require structural modifications of the entire system, so that it is possible to install several RGB-D sensors in every store.

Image processing algorithms previously presented was used for this system in order to recognize people monitoring and user-shelf interactions.

Detection of user-shelf interactions

The algorithm for user-shelf interactions is presented in this paragraph. A shelf zone is defined as the part of the store interested by an interaction between the hand of the shopper and a product in the shelf. As described in [30], it is defined by the user during the installation phase and it is characterized by the following tree parameters, written in a configuration file: the maximum distances of the left (x_{dl}), right (x_{dr}) and frontal (y_d) shelf sides from the image borders (see Figure 4.7). This setting is valid for most of the shelves of a store, but it is possible to define other areas of interest (e.g., in the case of a circular island when a cylinder-shaped configuration is needed).

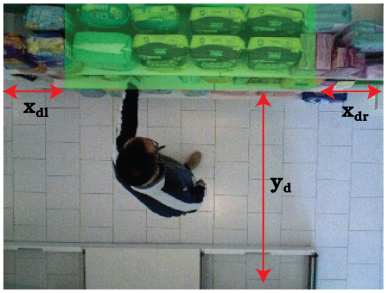


Figure 4.7.: Shelf zone definition using the three parameters (x_{dl}, x_{dr}, y_d).

The image processing algorithms such as multi level segmentation and water filling, allowed to detect the head and the body contours of each person. The contour of the head is used to track the movements of a person within the scene, while the contour of entire body is used to identify interactions with the shelf.

The three vertical planes built at the distances x_{dl}, x_{dr}, y_d from the image border and defining the shelf zone are used to detect interactions. When the contour of entire body intersects at least one of the three planes we establish that a contact occurs and so determine the 3D coordinates of the contact point.

The contact detection algorithm is explained in detail in the pseudo-code Algorithm 3. The FindInteractions function has in input the depth image of the sensor and the contour vector. Each point of each contour is analysed (from line 2 to line 4) to find contact points with *shelfzone*. If a contact occurs the pushBack method inserts in the vector *vec* the 3D contact point, where the third dimension corresponds to the depth value. Finally, the function returns vector *vec* with all the contact points.

Algorithm 3 Contact detection algorithm

```
1:  function FINDINTERACTIONS($d(x, y)$, $peopleVec$)
2:      for each contour $i \in peopleVec$ do
3:          for each point $p \in i$.GETCONTOURS( ) do
4:              if $d(p.x, p.y) \in shelfzone$ then
5:                  $vec$.PUSHBACK($d(p.x, p.y)$)
6:              end if
7:          end for
8:      end for
9:      return $vec$
10: end function
```

Efficiency and reliability of algorithms

The two main requirements/capabilities that the software has to satisfy are: *i*) to monitor people; *ii*) to understand the occurrence of an interaction between a shopper and a shelf.

To compare results we considered the sensitivity or true positive rate $TPR = TP/P$, where TP is the number of true positives, calculated by counting the real number of people "passing by" the camera and $P = (TP+FN)$, where FN is the number of false negatives, corresponding to the "passing by" people that the camera has not detected. The same evaluation method has been applied for establishing the correctness of the user-shelf interaction.

Table 4.3 shows the results of our performance analysis on 4 out of the 7 stores where the system was installed. We have checked the passages and the interactions of consumers measured by the system with the ground truth. More in detail, Table 4.3 corresponds to the confusion matrices of the people detection and contact detection algorithms. Since the system is built to detect only positive events (detection of people), values for true negative events can not provide. Also the number of negative events is unknown. For these reasons, typical confusion matrix parameters (e.g., specificity) are not listed. The sensitivity obtained was 99.02% and 80.52% for the people detection and the hand detection algorithm, respectively.

Table 4.3.: People detection confusion and user-shelf interaction matrices.

	TP	*FN*	*FP*	**Total**	***TPR***
People detection	1110	29	11	1150	0,9902
User-shelf interaction	1050	233	254	1537	0,8052

User-shelf interaction recognition with deep CNNs

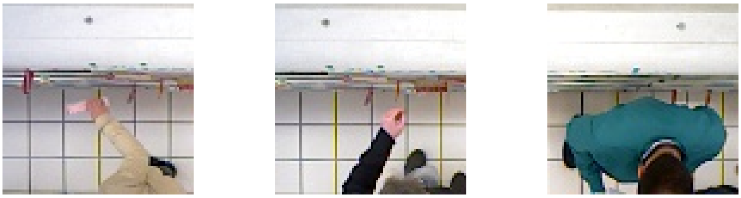
Deep CNNs are made up of different types of layers such as convolutional, pooling, fully connected layers. All of these may have additional hyperparameters such as filter size, padding, stride for the convolutional layers and the number of neurons for the fully connected layers. Note that parameters of the network architecture and training procedure are called hyperparameters to distinguish them from the weights and biases learned by the network during training, which are collectively called parameters. The amount of layer types along with their additional hyperparameters and their effect on training speed and quality make it difficult to choose an architecture.

The idea is to classify using CNN approach the type of interaction the customer has in front of a shelf. By using the previous algorithm 3 are possible to capture the moment when a part of the body comes into contact with the shelf.

In order to solve this challenge, a specific dataset (User-Shelf Interactions Dataset) is built. In particular, colour images of user-shelf interactions from 4 different cameras have been acquired. The goal is to combine each image with one of the following classes:

- *Positive*: the hand has already done the interaction and contains a product or the customer is putting the product back on the shelf.
- *Neutral*: the hand is approaching to the shelf in order to grab a product.
- *Negative*: this class contains the accidental interactions of customers with the shelf.

This distinction can be easily seen in the Figures 4.8 that includes the three types of images described above.



(a) Positive. (b) Neutral (c) Negative

Figure 4.8.

The collected dataset contains a total of 2745 colour images with a dimension of 80 × 80 pixels. The ground truth was manually labelled by four human annotators.

The purpose of this work is to train 4 different types of CNNs. Below, a small description of each network is reported.

CNNs

In a CNN, each neuron is only connected with a few local neurons in the previous layer, and the weight is shared for every neuron in that layer. Convolutional neural networks are effective for image classification problems because the convolution operation produces information on spatially correlated features of the image.

In order to solve the user-shelf interaction classification problem, two architectures are designed, trained and compared in this work. The baseline of the first two models are a classic CNN architecture, whose core structure is essentially the same as that of the LeNet architectures introduced in the late 1980s by LeCun *et al.* [60].

The first is composed by two convolutional layers, and each of them is generated by convolving through a 32 and 63 filters, respectively, and a 3×3 kernel. After each convolution, a ReLU activation function is used. Subsequently, there is a max pulling layer with a 2×2 size. After convolution, these features can be more readily learned by a fully connected neural network. Following, there are two fully connected layers with 128 and 3 nodes, respectively.

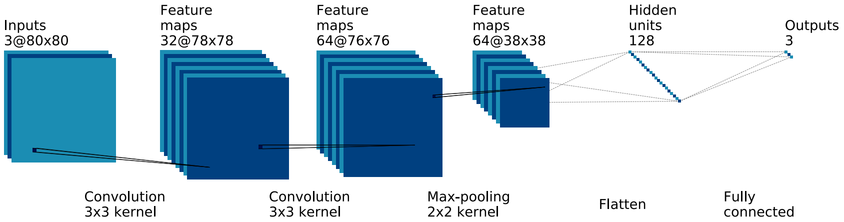


Figure 4.9.: CNN Architecture.

The second CNN architecture is slightly different from the first one. The main block, composed by two convolution and a max pooling layer, is duplicated.

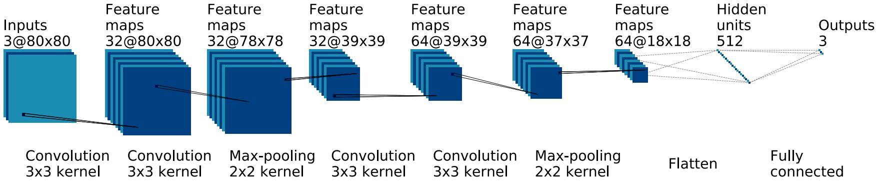


Figure 4.10.: CNN$_2$ Architecture.

AlexNet

AlexNet [56] is the first work that popularized ConvNets in CV. It features convolutional layer stacked on top of each over. AlexNet consists of a total of 8 layers, which are 5 convolutional layers and 3 fully-connected layers (final layer outputs the class labels). Batch-normalisation is applied after the first two convolutional layers. Dropout is applied after each layer during the last two fully connected layers.

CaffeNet

CaffeNet [49] is a modification of AlexNet [56] without relighting data augmentation and where the order of pooling and normalisation layers is switched. CaffeNet is made up of five convolutional layers, three fully connected layers and a softmax output layer. It uses ReLUs as activation functions and employs

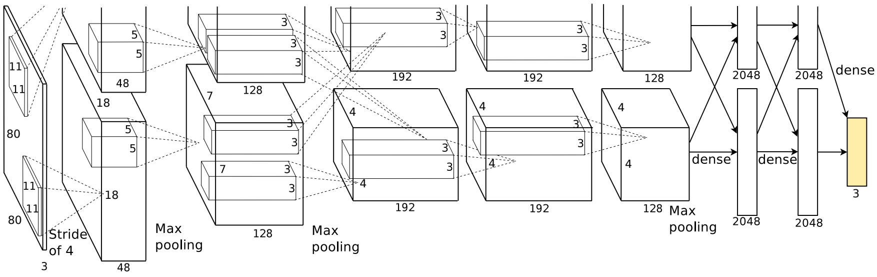


Figure 4.11.: AlexNet Architecture.

dropout in the first two fully connected layers. Additionally, pooling layers are used after the first, second and fifth convolutional layers; furthermore, Local Response Normalization (LRN) layers are used after the first two pooling layers. CaffeNet has the common architecture of convolutional layers followed by fully connected layers. The output softmax function interprets the data as a probability distribution and the result is, therefore, in the range of 0 to 1, summing up to a total of 1.

Classification Evaluation

For the evaluation of different trained networks on various test sets, an evaluation algorithm is implemented using Keras[5] Python library. It is able to load a model, load ground truth data, classify the test set and compare the results with accuracy, recall, precision and, f1-score.

Table 4.4.: User-shelf interaction results on train and validation sets.

Net	**Accuracy** *Train*	**Accuracy** *Validation*	**Loss** *Train*	**Loss** *Validation*
CNN	0.716238	0.809045	0.674015	0.541859
CNN$_2$	0.846936	0.909548	0.391440	0.290128
AlexNet [56]	0.737039	0.809045	0.616384	0.536162
CaffeNet [49]	0.885805	0.919598	0.483898	0.474933

Table 4.5.: User-shelf interaction results on test set.

Net	**Precision**	**Recall**	**F1-Score**
CNN	0.780691	0.622234	0.691118
CNN$_2$	0.873640	0.816821	0.843801
AlexNet [56]	0.771158	0.687371	0.726130
CaffeNet [49]	0.899060	0.873149	0.885705

[5] https://keras.io/

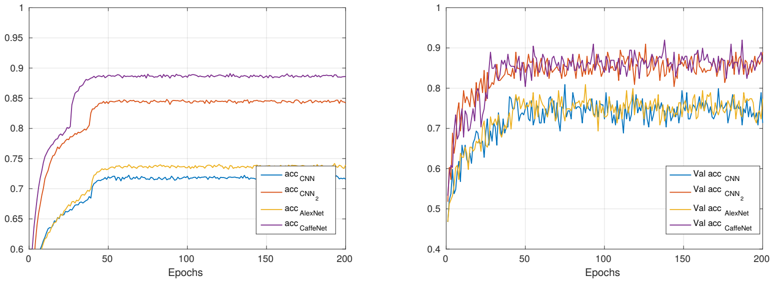
Data augmentation is a method applicable to shallow and deep representations, but that has been but that, so far, mostly applied to the latter [56]. By augmentation, an image I is perturbed by transformations that leave the underlying class unchanged (e.g. cropping and flipping) in order to generate additional examples of the class. The augmented samples can either be taken as-is or combined to form a single feature, e.g. using sum/max-pooling or stacking. The augmentation method has been applied at the training and test time.

The advantages of using several architecture can also be observed from the following two Figures 4.12. As the number of training iterations increased, the validation accuracy of the architectures quickly and smoothly ramped up to 0.7 after 40 iterations. On the other hand, the loss value of the convolutional neural network was lower, which indicated that the gradient descent function inside the nets had a better performance in converging to the local minimum point.

The loss value is calculated by a cost function, which essentially defines how far the model is from the desired output. The gradient descent is attempting to converge on a result that minimizes the cost function by slowly changing the weights.

Table 4.4 shows the performance results of the user-shelf interaction detection based on the training set.

In order to assess the predictive performance of a classification algorithm, it must be evaluated on a test set, i.e. a separate data set containing examples that have neither been used for training the algorithm, nor for choosing hyper-parameters, nor for determining when to stop training. After training the four models with different regularisation methods, the networks were finally evaluated on the official test set, obtaining the classification indicators displayed in Table 4.5.



(a) Accuracy trends on train set.   (b) Accuracy trends on validation set.

Figure 4.12.: Accuracy results.

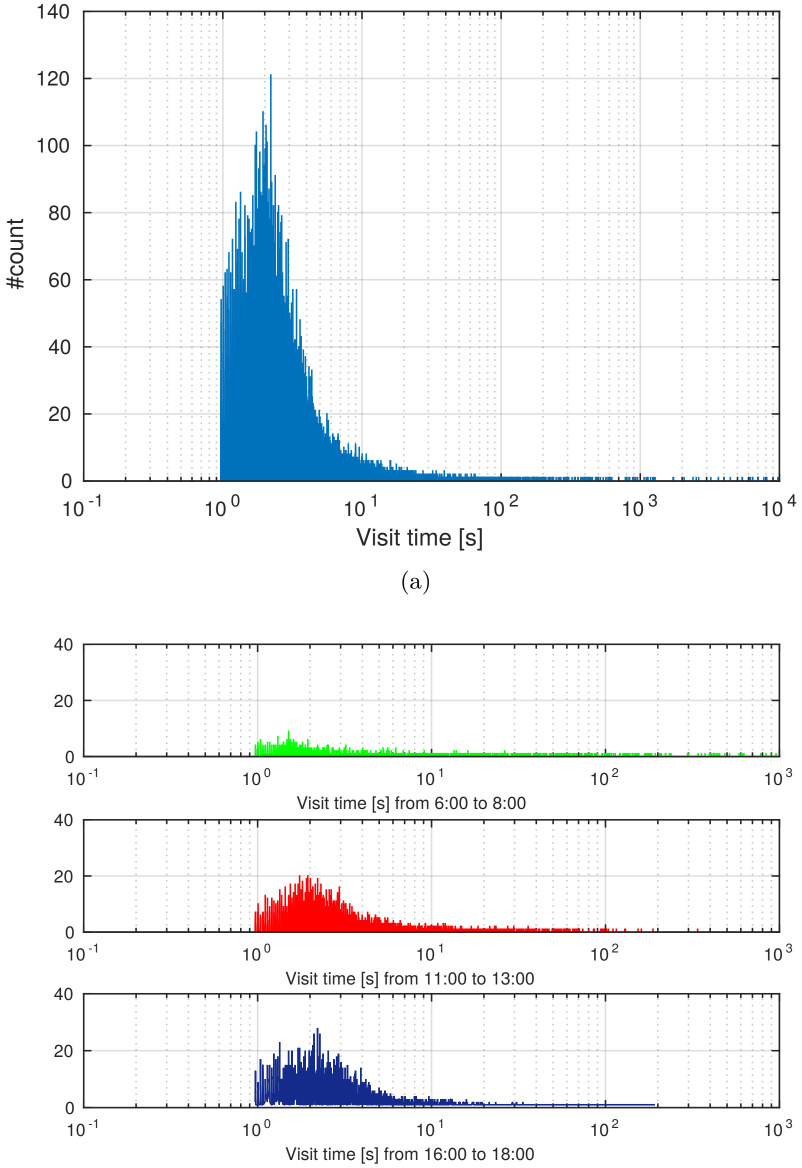
The curves in Figure 4.12 allow a first qualitative analysis, in particular, both models exhibited strong overfitting. The best value of accuracy on validation set is obtained with CaffeNet model and it corresponds to 91.9%.

Results

The main indicators adopted to evaluate shopper behaviour and preferences are:

- Number N_v of *visitors*, that is people crossing camera field of view;
- Number V_z of *visitors in each category*;
- Number V_s of *visitors people interacting with the shelf*, where $V_s \subset N_v$;
- Number N_s of *stopped visitors*, that is the number of people who stops in front of the selected category's shelf (min 5 secs) $V_s \subset N_s$;
- *Conversion rate* $CR = V_s/N_s \ \in [0,1]$ is the relationship between the number of stopped visitors and the number of visitors interacting with selected shelf products;
- Number I_s of *interactions for each person*, with $I_s = I/V_s$, where I is the *number of the interactions*;
- *Average visit time* $\ barT = \sum_{i=1}^{N_v} \Delta t_i/N_v$, where the *visit time* Δt_i is the permanence of each person in the camera view;
- Number P of products touched;
- Number P_{pos} of positive interactions, shopper touches the product and "buys" it (takes it from the shelf without returning);
- Number P_{neu} of neutral touching, shopper just touches the product without holding it.
- Number P_{neg} of negative interactions, shopper touches the product, holds it for a while and returns it to the shelf.
- Duration of interactions $T_I = \sum_{i=1}^{I} \delta t_i$, where $\delta t_i = t_{i,end} - t_{i,init}$ is the difference between final and initial instant of interaction i;
- Average interaction time $\bar{T_I} = T_I/I$.

Table 4.6 summarizes the results obtained by monitoring 7 stores using a total of 15 cameras and for a working period equivalent to 45 months by a single camera. The values in the table refer to the most significant indicators previously introduced. They reveal that the average visit time in front of the shelf is $6.21s$, while the average interaction time is $1.23s$. Moreover, the number



(a)

(b)

Figure 4.13.: Visit time histogram related to the overall studied period (4.13a). Visit time histogram related to three different time slots: *i*) 6.00 to 8.00 (*green line*). *ii*) 11.00 - 13.00 (*red line*) *iii*) 16.00 - 18.00 (*blue line*) (4.13b).

Table 4.6.: Values of indicators in real experiments stores.

Indicator	N_v	V_s	I_s	$\bar{T}$	P	$\bar{T}_I$
Value	87885	17762	1.45	$6.21s$	25710	$1.23s$

of interactions for each person is higher than the number of products touched, corresponding to 1.45 interactions for each person.

For example Figure 4.13a shows the histogram of the visit time Δt_i related to the overall testing period. Each bin corresponds to an elapsed time of $1s$. The plot shows that there are several counts (47% of the total) with a visit time smaller than $3s$, which can be easily interpreted as not-interacting people. The mean visit time higher than $3s$ is equal to $6.46s$. This value, when compared to similar results related to other shelves, can describe the appealing of the shelf to shoppers. Namely, the larger is the mean value higher is the shelf attractiveness. At the same time, the total number of counts indicates if the shelf is located or not in a populated place of the shop.

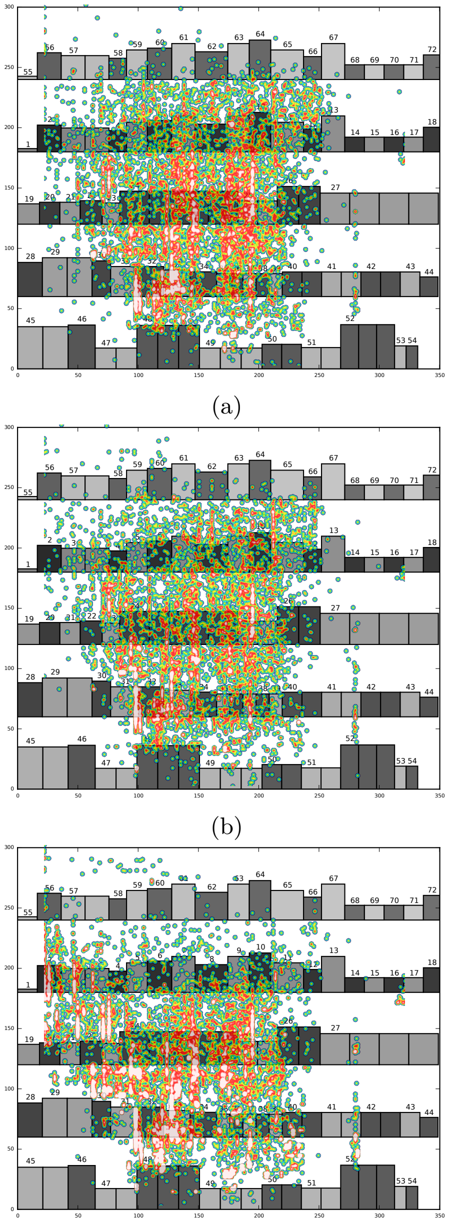
Since the system detects the precise date/time when each shopper appears in the camera field of view, it is possible to make the same histogram of Figure 4.13a for different time slots. Figure 4.13b shows an example of this for three different time slots: *i*) from 6.00 to 8.00; *ii*) 11.00 - 13.00; *iii*) 16.00 - 18.00. During the first slot, the shop is closed, hence the data refer only to shop operators. From the remaining slots we can evince that the morning time slot is more populated than the afternoon one. Such information can be very useful to better organize, for example, the staff of the shop according to the time slot with the maximum flux of buyers.

Going into the detail of the products hosted by the shelf, the system is able to storage all the interactions between the shopper and the shelf. Together with this, the system is able to discriminate among three different types of interaction with CNN approaches: *neutral* if the hand exceeds the threshold without taking anything; emphpositive when the object is picked up; *negative* when the object is put back after a pickup.

As expected the most interacted zone of the shelf is the central one, but looking at the width distribution of the interactions (Figure 4.14a, top panel) it is possible to discriminate at least another peak around shelf width equal to $800mm$, that probably corresponds to an appealing product.

Figures 4.14 show maps of the contact points, identified by coloured zones, generated during the processing. In particular, Figure 4.14a, Figure 4.14b and Figure 4.14c show respectively, positive, negative and neutral interactions.

Furthermore, Figures 4.14 represent some example of a planogram, that is a detailed visual map that establishes the position of the products in a shelf. So to obtain the contact map, the system automatically compares the coordi-



(a)

(b)

(c)

Figure 4.14.: Maps of interactions produced by the software in a test conducted by our research in a real environment. 2D plots showing the shelf along its width (*x axis*) and height (*y axis*) in millimetres.

nates of contact points with ancillary information provided by the planogram management software.

In this work, a pervasive, intelligent vision system for retail applications is developed. It consists in a software infrastructure coupled with a low cost hardware that: *i*) receives images from an RGB-D camera; *ii*) elaborates the images with CV algorithms; *iii*) extracts information and collects them into a database for statistical analysis and for being used by a decision support system.

The implemented CV algorithms: *i*) detect the people in the camera field of view; *ii*) measure the visit time of each person; *iii*) detect occurrences of interactions between shoppers and products on the shelf.

The system has been installed in real retail stores. The long life and real environment tests show the effectiveness of the described system and, in general, the feasibility of the proposed architecture and approach.

The efficiency of the system is defined by its capability in detecting people and shelf interactions. Results show that the people detection algorithm has a very high sensitivity, while the hand detection algorithm shows a good accuracy slightly above 90%.

The collected information can be used for several useful statistical studies, since they enhance, e.g., the knowledge of shopper-shelf interactions and the product appeal.

The system can be used as part of a sensor network focused on retail reality mining, with the purpose of better understanding customer interactions in retail environments.

4.3. Activities of daily living

The ADLs are a series of basic activities performed by individuals on a daily basis necessary for independent living at home or in the community. ADLs include eating, taking medications, getting into and out of bed, bathing, grooming/hygiene, dressing, socializing, cooking, cleaning and walking.

Automated recognition of ADLs is also interesting for the scientific community because of its potential applications in retail and security. Furthermore, monitoring human ADLs is important in order to identify possible health diseases and apply corrective strategies in AAL. ADLs analysis can provide very useful information for elder care and long-term care services.

In the next subsections are presented two approaches for activity recognition.

4.3.1. Activity recognition

As introduced, the ADLs analysis can provide useful information for elder care and long-term care services. This aspect can be observed in the recent appearance of smart environments, such as smart homes. Thanks to these advanced technologies, the assistance, monitoring and housekeeping of chronically ill patients or people with special needs or elderly has been enabled in their own home environments, in order to foster their autonomy in daily life by providing the required services when and where needed.

By using such systems, costs can be reduced considerably, while alleviating some of the pressure on healthcare systems. However, many issues related to this technology are raised such as activity recognition, assistance, monitoring.

For instance, dementia diseases of the elderly have a strong impact on ADLs. In fact, the ageing diseases result in a loss of autonomy. Medical researches [24] have shown that early signs of diseases, such as Alzheimer, can be identified up to ten years before the current diagnostics. Therefore the analysis of possible lack of autonomy in the ADLs is essential to establish the diagnostics and give all the help the patient may need to deal with the disease.

Being able to automatically infer the activity that a person is performing is essential for many disabilities in older adults, which have been associated with functional status based on ADLs in individuals with stroke, Parkinson's disease, traumatic brain injury, and multiple sclerosis.

The way to determine the autonomy of patient is to analyse his ability to execute the ADLs in his own environment. However, it can be complicated for a doctor to come and watch the patient doing these ADLs, as this would be a very time consuming task. An alternative would be to record the patient doing ADLs with a camera.

Previous papers on activity classification have focused on using 2D video [85] [40] or RFID sensors placed on objects and humans [118]. The use of 2D videos leads to low accuracy even when there is no clutter [69]. Moreover, RGB-D cameras are commonly used for the recognition of human actions [65]. Instead, the use of RFID tags is generally too intrusive because it requires RFID tags on the people.

Recognizing ADLs is a potential field where CV can really help, for example, elderly people to improve the quality of their lives [90]. Several research works and several models are proposed to recognize activities with intrusive and non-intrusive approaches. Activity recognition using intrusive approaches requires the use of specific equipment such as cameras.

Previous works on detection of human activities have been developed from still images as well as videos [74] [100] [47]. Many papers have shown that modelling the mutual context between human poses and objects is useful for activity detection [92] [53].

The recent availability of affordable RGB-D cameras, together with depth information, has enabled significant improvement in scene modelling, estimation of human poses and obtaining good action recognition performance [50] [61] [107]. This topic is very challenging and important because understanding and tracking human behaviour through videos has several useful applications. In [84] Nait-Charif *et al.* developed a computer-vision based system to recognize abnormal ADLs in a home environment. The system tracked human activity and summarized frequently active regions to learn a model of normal activity and the system could then detect falling as an abnormal activity.

Activity recognition with non-intrusive systems is a complex task, and it is based on a deep analysis of the data gathered from the environment. The sensors in the environment record the events about the state and any changes that happen within it. Each sequence of events is associated to a particular activity. The same person can perform an activity in several ways. This variation in the behaviour of a person leads to the generation of a set of patterns that characterize this person.

In this light, the variability in the person's behaviour and activity, detecting interesting patterns among many others, is a task of great importance for understanding the general behaviour of the person [2]. In fact, by discovering frequent patterns, the underlying temporal constraints, association rules, progress and changes over time, it is possible to characterize the behaviour of persons and objects and automate tasks such as activity monitoring, assistance and service adaptation [96].

Currently, there are many mathematical models for activity recognition, such as HMMs [95], Bayesian Networks [87], Kalman Filters [11] and Neural Networks [11]. Deep learning approaches on RGB video streams for activity recognition have also been introduced. This creates a system that improves and learns itself by updating the activity models incrementally over time [43].

Traditionally, most activity recognition work has focused on representing and learning the sequential and temporal characteristics in activity sequences. This has led to the widespread use of the HMM. In fact, in [108] HMM is employed with depth images to effectively recognize human activities. An HMM [95] is a finite set of states; each state is linked with a probability distribution. Transitions among these states are governed by a set of probabilities called transition probabilities. In a particular state a possible outcome or observation can be generated, according to the associated observation probability distribution. It is only the outcome, not the state that is visible to an external observer and therefore states are "hidden" to the outside, hence the name HMM.

In earlier exploratory studies the HMM has shown good results thanks to their suitability to model sequential data, which is the case for monitoring human activities. Indeed, acceleration data are measured over time during physical human activities of a person and are therefore sequential over time.

In [19] an approach to activity recognition for indoor environments based on incremental modelling of long-term spatial and temporal context is presented. Even in [20] the authors introduced a simple way to apply qualitative trajectory calculus to model 3D movements of the tracked human body using HMMs. HMMs combined with GMM to model the combination of continuous joint positions over time for activity recognition was introduced in [91].

True daily activities take place in uncontrolled and cluttered households and offices and they do not happen in structured environments (e.g., with closely controlled background). For this reason their detection becomes a much more difficult task. In addition, each person has their own habits in carrying out tasks, and these variations in style and speed create additional difficulties in trying to recognise and to detect activities.

In this work, interest is focused on reliably detecting daily activities that a person performs in the kitchen. In this context, this work proposes an automated RGB-D video analysis system that recognises human ADLs activities, related to classical actions such as making a coffee. The main goal is to classify and predict the probability of an analysed subject action. Activity detection and recognition are performed using an inexpensive RGB-D camera. Human activities, despite their unstructured nature, tend to have a natural hierarchical structure; for instance, generally making a coffee involves a three-step process of turning on the coffee machine, putting sugar in the cup and opening the fridge for milk. Action sequence recognition is then handled using a discriminative HMM. A dataset with RGB-D images and 3D position of each person for training as well as evaluating the HMM has been developed and made publicly available.

Several contributions are made by this work. First of all, the model is generic, so it can be applied to any sequential datasets or sensor types. Second, the model deals with the problem of scalability by taking into account the sequences recorded independently of the environment. Finally, the approach is validated using real data gathered from a real smart kitchen which helps to make our results more confident and the experiments repeatable.

The innovative aspects of this work are in proposing an adequate HMM structure and also the use of head and hands 3D positions to estimate the probability that a certain action will be performed, which has never been done before in ADLs recognition in indoor environments.

In this work, a method for ADLs recognition is proposed. In particular, it focused on using the HMM to facilitate the detection of anomalous sequences in a classical action sequence such as making a coffee.

Design of HMM structure

Let:

$$X = \{x_1, x_2, \dots, x_n\}$$

be a discrete finite activity space and

$$O = \{o_1, o_2, \dots, o_m\}$$

the observation space of a HMM [95]. Let T be the transition matrix of this HMM, with $T_{x,y}$ representing the probability of transitioning from activity $x \in X$ to activity $y \in X$, and $p_x(o)$ be the emission probability of observation $o \in O$ in activity $x \in X$.

The probability that HMM trajectory follows the activity sequence s given the sequence of n observations is denoted as:

$$P(X_{1:n} \in seq_n(s) | o_{1:n})$$

where $seq_n(s)$ is a set of all length n trajectories whose duration free sequence equals to s.

Finding the most probable activity sequence can be seen as a search problem that requires evaluation of probabilities of activity sequences. The Viterbi algorithm based on dynamic programming can be used to efficiently find the most probable trajectory. In fact, it makes use of the Markov property of an HMM (that the next state transition and symbol emission depend only upon the current state) to determine, in linear time with respect to the length of the emission sequence, the most likely path through the states of a model which might have generated a given sequence.

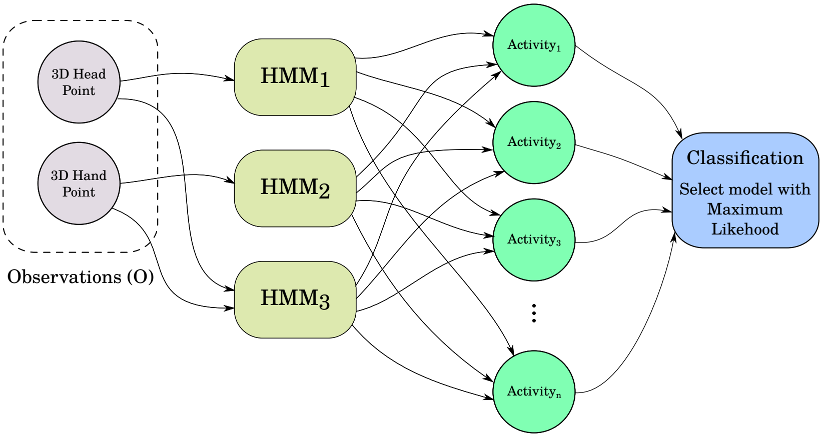


Figure 4.15.: Block diagram of the recognition process.

Head and hands detection algorithms

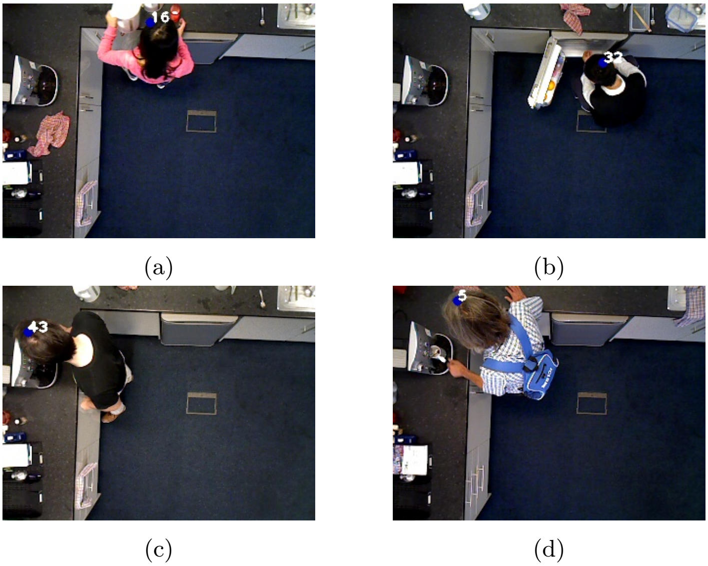
The main goal of this work is to classify different activities that people carry out during their daily life using an RGB-D camera in a top-view configuration. The idea is to extract from depth information the $3D$ position of the person for each frame. In particular, using the multi level segmentation algorithm in [61], the head and the hands of each person are tracked when these are visible. In fact, this algorithm intends to overcome the limitations of the single-level segmentation in the case of collisions among people in the same scene.

In a similar way, to find the hands $3D$ position, this algorithm is applied again to each person blob leaving out the upper part of person profile (head and shoulders) previously found.

ADLs model

In this paragraph, an ADLs model is described. It takes into account both the complexity of data and the lack of a large amount of training data for learning purposes. The problem of recognition of daily activities in the image to its simplest core, can be notice as an equivalence between an activity and a hidden state of an HMM. This could be obtained with the design of a fully connected HMM and training the inherent state-transition probabilities from the labelled data. Regarding these ADLs as very heterogeneous and complex, the suggested equivalence between an activity and a hidden state cannot hold.

Information provided by head and hands detection algorithms can be used as input for a set of HMMs. Each of these recognise different actions sequence. After training the model, an action sequence $s = \{s_1, s_2, \ldots, s_n\}$ is considered and its probability λ is calculated for observation sequence $P(s|\lambda)$. Then the



(a) (b) (c) (d)

Figure 4.16.: Snapshots of RADiAL session registration.

action is classified as the one which has the largest posterior probability.

Figure 4.15 depicts the general scheme of the recognition process. In particular, the three different HMMs are used, which have respectively as observations, the 3D points of::

- the head (HMM_1);
- the hands (HMM_2);
- both head and hands together (HMM_3).

Table 4.7 indicates the number of vertical and horizontal layers used in the quantization step for each HMM and the total number of observations, after the resampling process.

The set of actions includes:

- making a coffee;
- taking the kettle;
- making tea or taking sugar;
- opening the fridge;
- other activities performed in a kitchen environment.

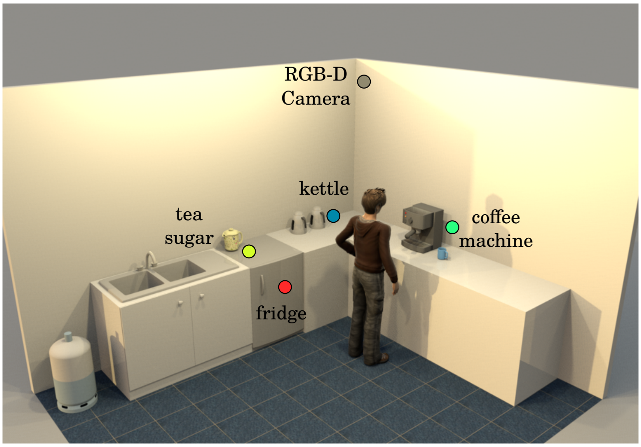


Figure 4.17.: Reconstructed layout of the kitchenette where RGB-D camera is installed.

Finally, the classification module provides the action x_j that maximizes P_{HMM_i}. It is the HMM trajectory probability that follows the activity sequence s given the sequence of n observations, i.e.:

$$x_j = \arg\max_i P_{HMM_i}(X_{1:n} \in seq_n(s)|o_{1:n}) \tag{4.4}$$

Setup and acquisition

To evaluate the usefulness of approach for activity recognition, a new dataset is built. This dataset contains common daily activities such as making coffee, making tea, opening the fridge and using the kettle. The data were collected over a period of 5 days.

The dataset also consists of random activities of each individual that can be performed in a kitchen environment, which are not similar to any other activity done before. The RGB-D camera was installed on the ceiling of L-CAS laboratory at approximately $4m$ above the floor. The camera was positioned above the surface which has to be analysed (Figure 4.17).

Table 4.7.: Number of observations for each HMMs (v: vertical layer, h: horizontal layer).

3D Points	# layers	# observations
head	$v:8,\ h:8$	512
hands	$v:8,\ h:8$	512
head & hands	$v:8,\ h:8;\ v:8,\ h:8$	262144

RADiAL Dataset

The RADiAL dataset[6] was collected in an open-plan office of the Lincoln Centre for Autonomous Systems (L-CAS). The office consists of a kitchenette, resting area, lounge and 20 working places that are occupied by students and post-doctoral researchers. A ceiling RGB-D camera was installed (Figure 4.17) that took a snapshot (with dimensions of 320×240 pixels, Figure 4.16) of the kitchenette area every second for 5 days, and activities of one of the researchers over time were hand-annotated. Furthermore, the RADiAL dataset contains the 3D positions of the head and hands for each person with a minute-by-minute timeline of 5 different activities performed at the kitchen over the course of days. RADiAL contains 100 trials. Each trial includes the actions related to one person.

Results

The experimental results obtained using our approach are presented. An architecture to implement HMMs ADLs recognition is proposed. The architecture uses the 3D points extracted from the head and hands to classify different sequences of actions corresponding to some ADLs.

The standard algorithm for HMM training is the forward-backward, or Baum-Welch algorithm [6]. Baum-Welch is an iterative algorithm that uses an iterative expectation/maximization process to find an HMM which is a local maximum in its likelihood to have generated a set of training' observation sequences. This step is needed because the state paths are hidden, and the equations cannot be solved analytically.

In this study, the Baum–Welch algorithm was employed to estimate a transition probability matrix and an observation emission matrix so that the model best fits the training dataset.

Since the discrete observation density is used in implementing HMMs, a Vector Quantization and clustering step is required to map the continuous observation in order to convert continuous data to discrete data.

A total of five models of activities were built using the method described in Subsection 4.3.1. The models were used to recognize activities in the RADiAL dataset. The five models correspond respectively to the activities "Other" (this action contains all the other activities performed in a kitchen environment), "Coffee" (making a coffee), "Kettle" (taking the kettle), "tea/sugar" (making tea or taking sugar), and "fridge" (opening the fridge). The results were obtained using two different validation techniques.

Below, the results are given at first for the head only (HMM_1), then the hands (HMM_2) and finally, the combination of both (HMM_3). The main

[6] http://vrai.dii.univpm.it/radial-dataset

Table 4.8.: Classification Results Cross Validation HMM_1

	precision	**recall**	**f1-score**
other	0.73	0.57	0.64
coffee	0.67	0.80	0.73
kettle	0.60	0.70	0.65
tea/sugar	0.66	0.70	0.68
fridge	0.74	0.61	0.67
avg / total	0.68	0.68	0.68

goal is to gradually improve the activities recognition.

In the case we applied a k-fold cross-validation approach (with $k = 5$) to test our HMM_1. The resulting confusion matrix is shown in Figure 4.18a. In the confusion matrix most of the actions are detected with high accuracy. Table 4.8 summarises the activity recognition results demonstrating the effectiveness and suitability of approach.

The confusion matrix for HMM_2 is depicted in Figure 4.18b. The activity recognition results, as reported in Table 4.9, prove the effectiveness and suitability in terms of precision, recall and f1-score.

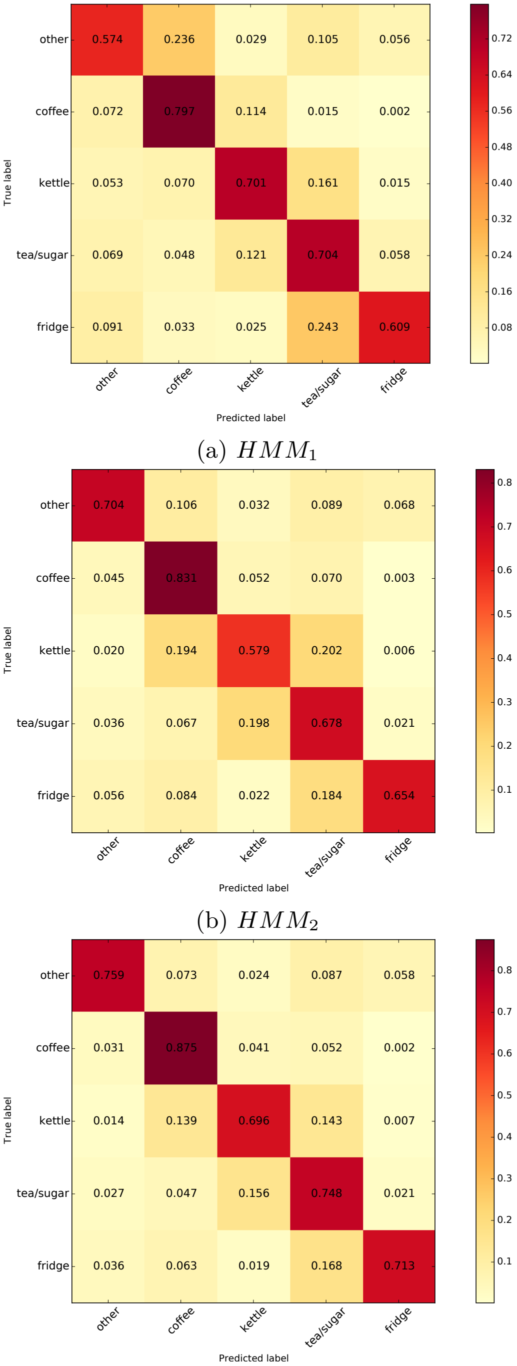
Table 4.9.: Classification Results Cross Validation HMM_2

	precision	**recall**	**f1-score**
other	0.89	0.70	0.79
coffee	0.69	0.83	0.75
kettle	0.47	0.58	0.52
tea/sugar	0.64	0.68	0.66
fridge	0.74	0.65	0.69
avg / total	0.73	0.71	0.71

The confusion matrix of HMM_3 for both the head and hands is shown in Figure 4.18c. The results in Table 4.10 indicate an increase in the metrics for evaluating the performance of our approach.

Table 4.10.: Classification Results Cross Validation HMM_3

	precision	**recall**	**f1-score**
other	0.93	0.76	0.84
coffee	0.76	0.87	0.81
kettle	0.58	0.70	0.63
tea/sugar	0.70	0.75	0.72
fridge	0.78	0.71	0.74
avg / total	0.78	0.77	0.77



(a) HMM_1

(b) HMM_2

(c) HMM_3

Figure 4.18.: k-fold cross-validation confusion matrices.

4.3.2. Ambient assisted living

Recently the society, particularly in the industrialized countries, is moving towards a significant demographic change known as the so-called ageing society. This is due to the increase in life expectancy, which causes the ageing of population. For maintaining the expenses for the health care within the limits of economic possibility, it is necessary to find cost-effective and innovative solutions [104, 35].

Currently, according to the World Health Organization [88], elderly people (i.e. people of 60 years of age and older) in the world is about 650 million and by 2050 will reach the 2 billion. At the same time, even European population will keep on growing older. In fact, in 2008 the population over the age of 65 was over 17%, and in 2060 it will rise to 30%. As regards the population over the age of 85, instead, the rate will rise from 4% to 12% [37]. Furthermore, Eurosat estimates that from this year inwards, deaths will exceed births. Then, in this situation the social behaviour, the lifestyle and the identity of senior people will strongly change. It is important, therefore, to implement smart solutions for elderly care because they should remain independent and able to work for a longer time. This can be achieved thanks to the technology. Falls, in fact, are the main cause of injury death among people 65 years and over and they are a significant obstacle to independent living of the seniors [77]. They are the cause of hospital admissions for loss of independence and traumatic injuries. Among hospitalized patients about 70% of accidents are due to falls. For ensuring user-friendly assistive devices have been devoted many efforts [83]. Generally falls occur in home environment and most of them often happens during the night time [123]. Frequently, at the time of falling, the subjects are getting up from bed or chair to go to the bathroom.

Lately, AAL has attracted a growing attention in scientific community since it involves emerging and innovative technological solutions, providing embedded system in the home environment, that will increase the quality of life and will reduce costs for independent living. In the AAL field, the automatic fall detection is an important issue, because falls affect the elderly people living alone. The aim of AAL is to increase the self-confidence and the autonomy of elderly or ill person for enhancing their security [104].

Currently the most frequently used techniques for fall detection are focused on wearable sensors, such as gyroscopes and accelerometers [86]. However, this devices often generate false alarms, because some ADLs are manifested with fast moving down, that can be classified as a fall from a detector based only on inertial sensor [14]. For this reason, in literature, a lot of research investigates the fall detection using various sensors, for example the survey of Mubashir *et al.* [83]. Other used methods are based on vision systems, however, a lot of ethical issues about the respect of privacy and intimacy, especially in bathroom and in bedroom, are introduced by the continuous monitoring through a vision systems. Moreover, these devices do not work in poor light or night-light conditions, which are among the most frequent situations in which falls happen.

In this context, this work proposes an automated RGB-D video analysis system that recognizes dweller activities that are crucial for assistance purposes, with a particular focus on the detection of falls. The study is based on people detection and tracking algorithms, for mapping the users and for the detection of important events such as falls or sit in a chair. The system allows to extract and collect a lot of statistical data that, properly processed, provide knowledge about the dwellers.

The knowledge can be an aid for customizing the home according to their needs and to adapt the design of the home to their habits. In order to provide such a knowledge, the movements of the dwellers are recorded on line using a database. This way, the physical activity of the subject can be supervised anytime. Furthermore, the RGB-D camera is able to extract the depth images even in dark rooms, respecting the subjects' privacy.

Experimental tests conducted in different domestic scenarios proved the effectiveness of the proposed solution, that is fast, accurate, and able to provide a fall map in-home fall risk assessment.

Physical architecture

The physical architecture of system includes a RGB-D sensor installed in a top-view configuration, as illustrated in Figure 4.19. It is controlled by an embedded system that manages the sensor acquisition and that processes the depth stream extracting measures of the people on the camera view. The sensor is installed $3 \div 4\ m$ above the floor, by covering an area of $8.25 \div 14.70\ m^2$.

Software architecture

Figure 4.20 shows the software architecture of system.

As a first step, the RGB-D sensor acquire the depth stream. After this, GMM algorithm is applied. This way a background model is obtained and the background subtraction procedure can be used. Than, multi level segmentation or water filling algorithms are used in order to find the heads of subjects. These segmentation algorithms provide also a set of features including:

- the height of each person, h_{height};
- the size of each head, A_{head};
- the head-shoulders distance, $d_{h/s}$.

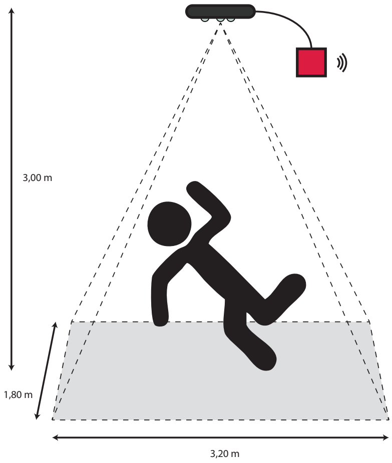


Figure 4.19.: Representation of the home environment in which the system is installed.

Furthermore, people detection algorithm determines if a set of pixel represents a person or an object by using the features extracted by the segmentation procedure.

A tracking algorithm is used to maintain the same *ID* for as long as the person remains in the field of view.

The feature extracted, over the time, of the height of each person, are used for classify the the posture of the person. In particular, three types are analysed:

- when a person is standing up;
- when a person is sitting;
- and when a person is falling.

The last scenario is the most important to be considered for the elderly safety and, therefore, I take into account, the robustness of this measurement and I provide also an output signal that emits an alarm.

The information provided by the last procedures is recorded in a database, where is an Analytical Processing System, i.e. a separate process that accesses the data published on the database and extracts statistics and knowledge about the inhabits.

Results

This paragraph discusses the results of the proposed technique for fall detection problem, based on the processing of the depth information provided by the RGB-D sensor in top-view configuration.

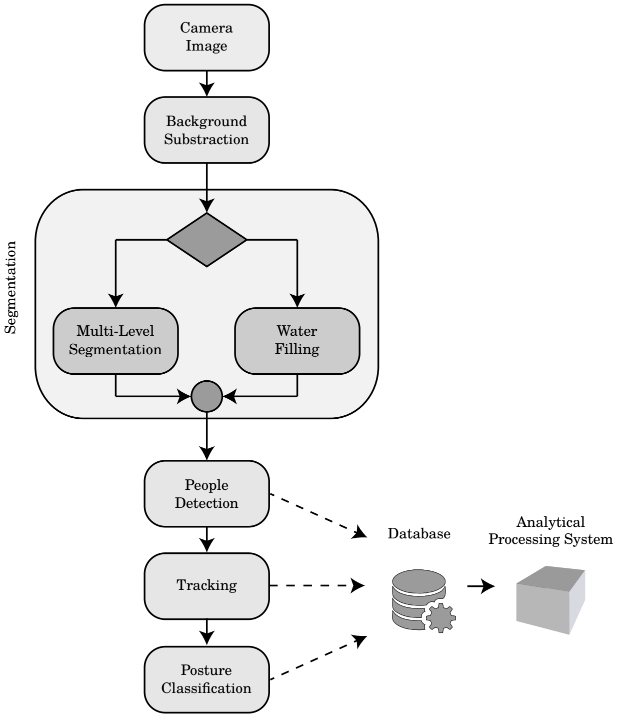


Figure 4.20.: The main steps of algorithms.

The system has been tested quite extensively in laboratory. One major concern is that a fall simulated in laboratory may be significantly different from an actual fall. This could have a large impact on the velocity based algorithm. For example, if actual falls have a shorter duration or lower velocity than those recorded in laboratory, the number of frames and threshold velocity would have to be adjusted. As a first evaluation of the fall detection system, a basic case is considered (Figure 4.21): a single person walks in the scene, and falls to the ground, without interacting with objects (i.e., the person does not intercept any object when walking or falling down).

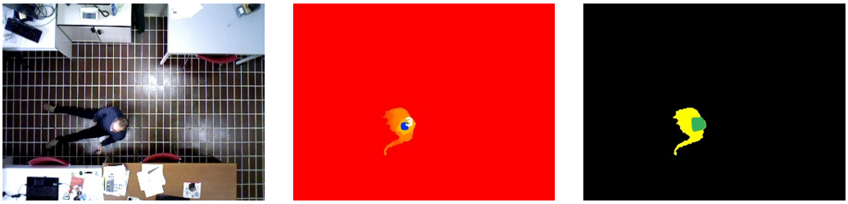


Figure 4.21.: Fall simulation during test phase in laboratory.

Figures 4.22a and 4.22b show the image processing procedure during a fall simulation. These data are used to detect the fall of a person. Instead of using the floor plane equation provided by the max distance of depth image (this is not always detected, particularly on stairs), the following equation is used to calculate the floor plane.

$$Ax + By + Cz + D = 0 \tag{4.5}$$

where,

$$A = 0,$$

$$B = \cos\eta,$$

$$C = \sin\eta,$$

$$D = 3.$$

A, B, and C are simply the vector normal to the floor and D shifts the floor plane 3 meters below the RGB-D sensor. The distance from the floor plane can then be calculated using:

$$d = \frac{Ax + By + Cz + D}{\sqrt{A^2 + B^2 + C^2}} \tag{4.6}$$

For frame i and $i + 1$ the velocity for a particular joint normal to the floor is then:

$$v_i = \frac{d_{i+1} - d_i}{t_{i+1} - t_i} \tag{4.7}$$

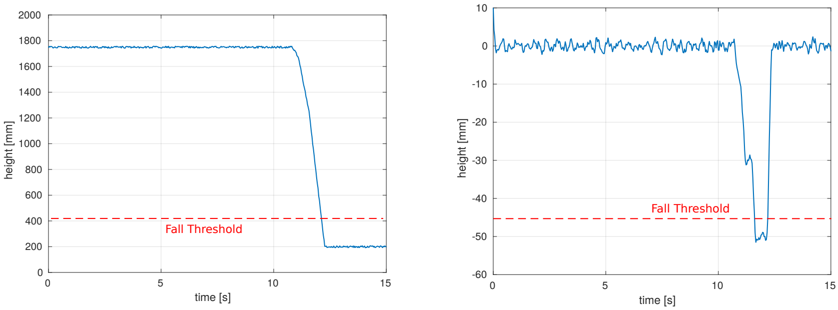
Where t is the timestamp in milliseconds.

Since the 3D head point trajectory has been tracking, the head motion can be analysed by the physics mechanics principle. As recommended by the author of [10], during the falling phase, the joint motion can be seen as a free fall body. The free fall body is described as aa equation:

$$h(t) = h_0 + \frac{1}{2}a(t - t_0)^2 \tag{4.8}$$

where $h(t)$ is the height at the time t, h_0 represents the height at the beginning of fall, a is the acceleration, t is the current time, and t_0 is the starting time. This formula can be used to simulate the point fall motion to generate fall patterns. Figure 4.22 shows a fall head trajectory of a person. This curve fits well with previous function. The difference can be considered as Gaussian white noises. In order to improve the robustness, Gaussian white noises are added into the free fall body curve. Based on the free fall body simulation, a large fall and non-fall patterns dataset can be built up.

In order to confirm the fall detection, the recover motion analysis after the



(a) Height of person.

(b) Derived height function.

Figure 4.22.: Position 4.22a and velocity 4.22b over time for standing up and falling.

fall motion is required. There are two recover metrics:

1. the heights of shoulders and head are higher than a recover threshold value $T^1_{recover}$ for a certain time;
2. the height of the head is higher than a high recover threshold value $T^2_{recover}$ for a certain time.

If one of these two metrics is satisfied, it means that the person is recovered.

In order to verify the system capabilities in term of frame rate and accuracy, the system was tested in a simulated environment, considering a bathroom, a kitchen and an hallway instrumented with the installed sensors. The image acquisition and processing allows a processing rate of about 25 frames per second, meeting the system requirements.

The same considerations are also valid to understand if the person is sitting.

Movements map

The system can shows in real time a heat map of the movements of the people inside the field of view of RGB-D sensor (figure 4.23). The heat map is a graphical representation of data where the values contained in the matrix are represented as colours. This map is a valid instrument for the real time visual monitoring. Moreover, it is possible to establish, with high confidence, the percentage of time spent in a certain area and detect prolonged stay in the fall position.

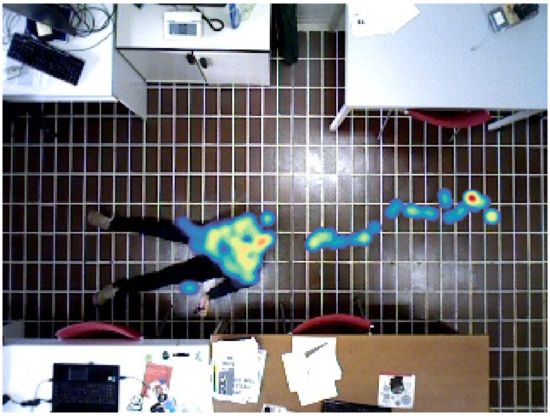


Figure 4.23.: Example of "*Movements Map*" generated by the software.

Chapter 5.

Conclusions and future works

In this Thesis, different algorithms and applications based on RGB-D data in top-view configuration for HBU have been proposed. In particular, the implemented algorithms can be used for people detection and tracking and, for human interactions understanding. This Chapter starts presenting a discussion summarizing the main results achieved in the Thesis. Then, the main contribution is highlighted. Finally, the open issues and future research directions are presented.

5.1. Discussion

After the introduction, a review of the literature on the two main topics addressed in this Thesis, i.e. HBU and RGB-D data form top-view has been provided. Chapter 2 included also an overview of main public available datasets.

Two approaches for people detection are proposed in Chapter 3. In particular, the objective of these approaches is to find the heads of people present in the depth image.

Image processing techniques, such as water filling and multi level segmentation, provide the shape of heads by evaluating depth data local minima. In this way, the number of false positives increases when the subjects gesticulate with their hands. In fact, these could be confused for heads. The introduction of particular constraints on the shape could reduce this number, but the risk is that the algorithm may become too specific for each setup.

Approaches based on semantic segmentation techniques allow a neural network to distinguish a particular class, which in this case corresponds to the class "head". Five different types of CNNs were tested and achieved significantly better results than image processing approaches.

In Chapter 4 are presented some applications for different use cases. Previous algorithms for people detection have been used to monitor their movements within a particular area. In fact, three macro-research fields were analysed.

In the context of video surveillance, it was necessary to extract different characteristics of the subject in order to re-identify the latter when it reappears a second time. A dedicated dataset has been built to solve this problem and anthropometric and colour based features have been extracted. CMC curves evidence the robustness and good ability of descriptors to recognize the various subjects.

Another application addressed has been developed in an IRE. Several RGB-D sensors have been installed in a real store in order to monitor the behaviour of consumers in front of different shelves. Moreover, through depth data, the system detects all type of interaction that the costumers has with the shelf. Four different CNNs have been developed to understand the type of interaction, i.e. whether or not the customer has taken a product from the shelf. This system is useful because it is able to extract a series of indicators that describe the performance of store up to the single shelf in real time.

The last application scenario, where top-view RGB-D data for people detection algorithms has been used, was home environments. Two different types of problems have been addressed: ADLs and fall detection. The first was HMM approached. 3D points of head and hands were used as input of model (observation) in order to predict the activity of the user (state). The model was validated with a dataset and five types distinct of activity were considered. Instead, fall detection problem was solved by monitoring the person's height value.

5.2. Thesis contributions

The main contributions of this Thesis can be summarized as follows:

- design and implementation of two novel algorithms for people detection from top-view configuration with RGB-D data using image processing approaches. In particular, a performance improvement of water filling algorithm is proposed in terms of computational complexity. Furthermore, a new algorithm, called multi level segmentation, has been developed. It carries out several segmentations on different levels of height in order to find all the heads of people.

  This work has been published in [61, 65, 63];

- development of semantic segmentation CNNs for heads detection, in particular, U-Net, SegNet, FractalNet, and ResNet are used in this work. By introducing changes on different layers of these nets, the performances are significantly improved;

- proposal and validation of new descriptors for Re-id task in top-view configuration. Descriptors are composed of anthropometric and colour-based features.

This work has been published in [66];

- design and implementation of several CNNs for user-shelf interaction recognition. Through a manually annotated dataset made up of images representing interactions between user and shelf, four different types of CNNs have been trained.
- creation of four public available datasets:
  - TVPR Dataset
  - TVHeads Dataset
  - RADiAL Dataset
  - User-Shelf Interactions Dataset

  Some of these datasets have been published in [66, 64];

In this Thesis, the potential of top-view configuration for detection and tracking applications in several sub-domains has been demonstrate, to outline key limitations and to indicate areas of technology where solutions for remaining challenges may be found. The success of RGB-D cameras can be closely linked to their affordability and to the additional depth information coupled with visual images that this approach provides. These cameras have already been successfully applied in the several field to identify people and to analyse behaviours and interactions. The choice of the RGB-D camera in a top view configuration is due to its greater suitability compared with a front view configuration, usually adopted for gesture recognition or even for video gaming. The top-view configuration reduces the problem of occlusions and has the advantage of being more privacy preserving, because a person's face is not recorded by the camera. Starting from this, further investigation could be devoted to explore approaches more accurate and effective such as Convolutional Neural Networks or U-Net [99].

5.3. Open issues and future works

This section analyses the open issues in the proposed algorithms and applications, identifying some future research directions.

Novel algorithms for head detection are based on deep learning approach. The CNNs used in this Thesis are the best techniques of semantic segmentation. Indeed, this approach is better than traditional image processing techniques because it is not based on geometric or colour constraints, but rather allows to identify the heads of individuals based on previous learning.

Further investigation on Re-id task and video surveillance field, will be devoted to the study of more sophisticated features. The CMC curves have suggested that for the different distance metric approaches the depth descriptor has strong discriminative power. The integration of more features in the model seems to improve the identity discrimination. This aspect is of great importance, in order to perform a classification model. Future works would include the use of other types of RGB-D sensors, such as time of flight (TOF) ones. The system can additionally be integrated as a source of high semantic level information in a networked ambient intelligence scenario, to provide cues for different problems, such as detecting abnormal speed and dimension outliers, that can alert of a possible uncontrolled circumstance.

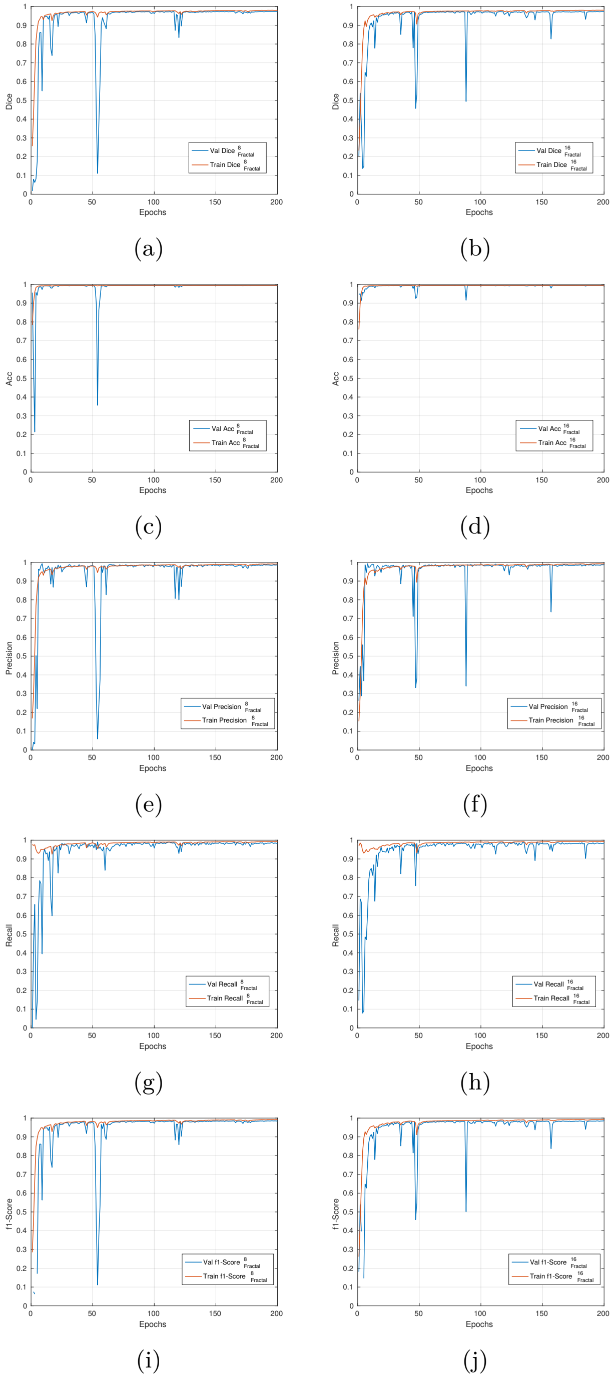
Future projects about IRE are directed towards a detailed study of person Re-id using top-view RGB-D data from several cameras, a task necessary to assign a single and robust identifier to each buyer. Among several other information (i.e. audio recognition system, carts tracking), this will allow us to better describe the client behaviour inside the shop and not only in front of a single shelf.

Future efforts in the field of assistive technology are expected on the integration of video and audio systems. In this way, the identification of abnormal events, such as the strange activity of the user, an intrusion or a object breakage, can can be detected using audio microphones. Further investigation will be devoted to extend HMM approach to select human joints that provide the most informative spatio-temporal relations for ADLs classification. The long term goal in this field is to develop a mobile robot that searches for the best location to observe and successfully recognise ADLs in domestic environments.

Appendix A.

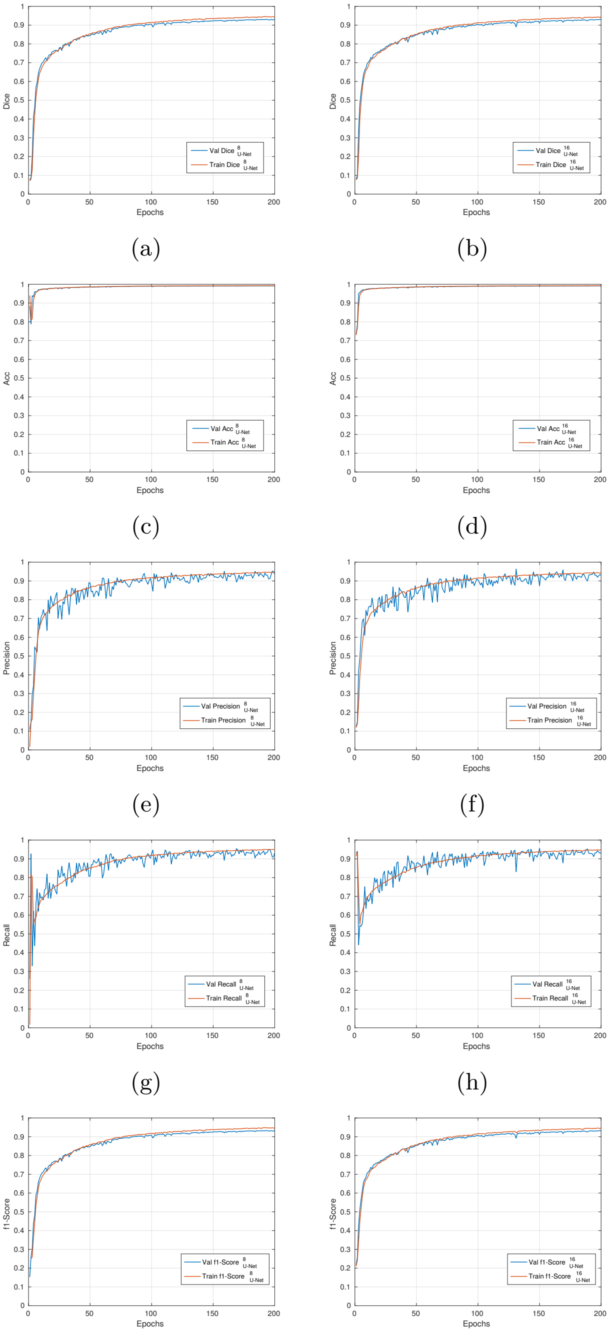
Appendix

A.1. Semantic segmentation results



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure A.1.: Fractal results.



(a) (b)

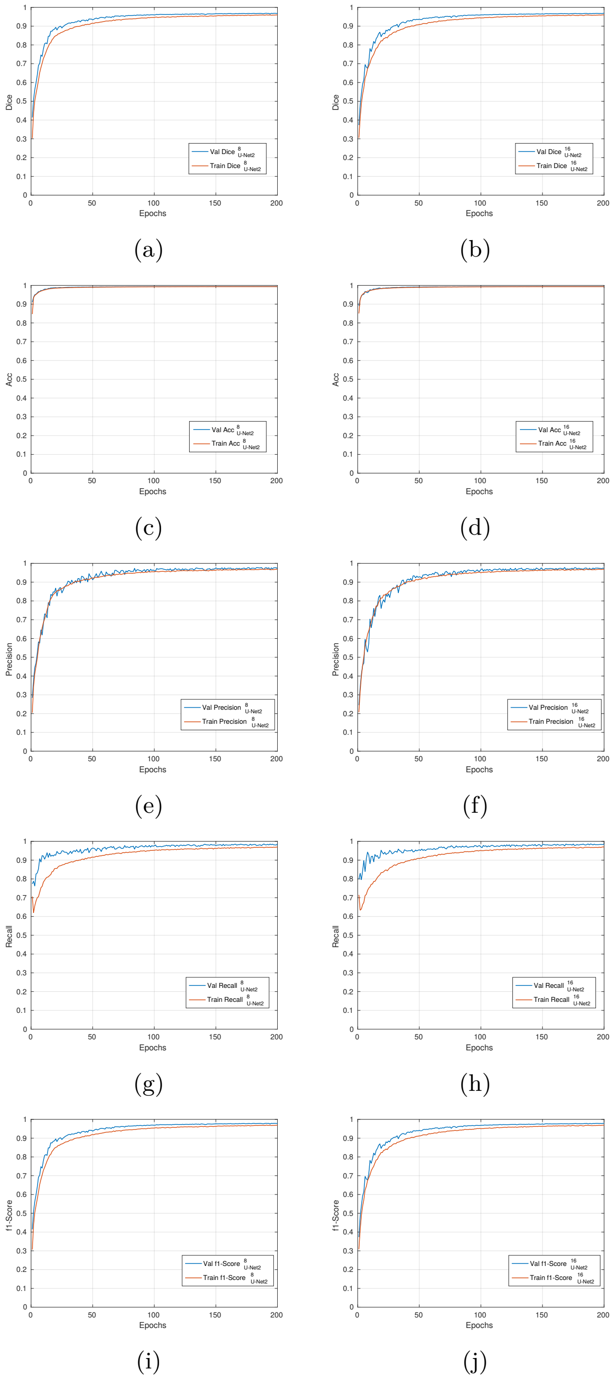
(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure A.2.: U-Net results.



(a) (b)

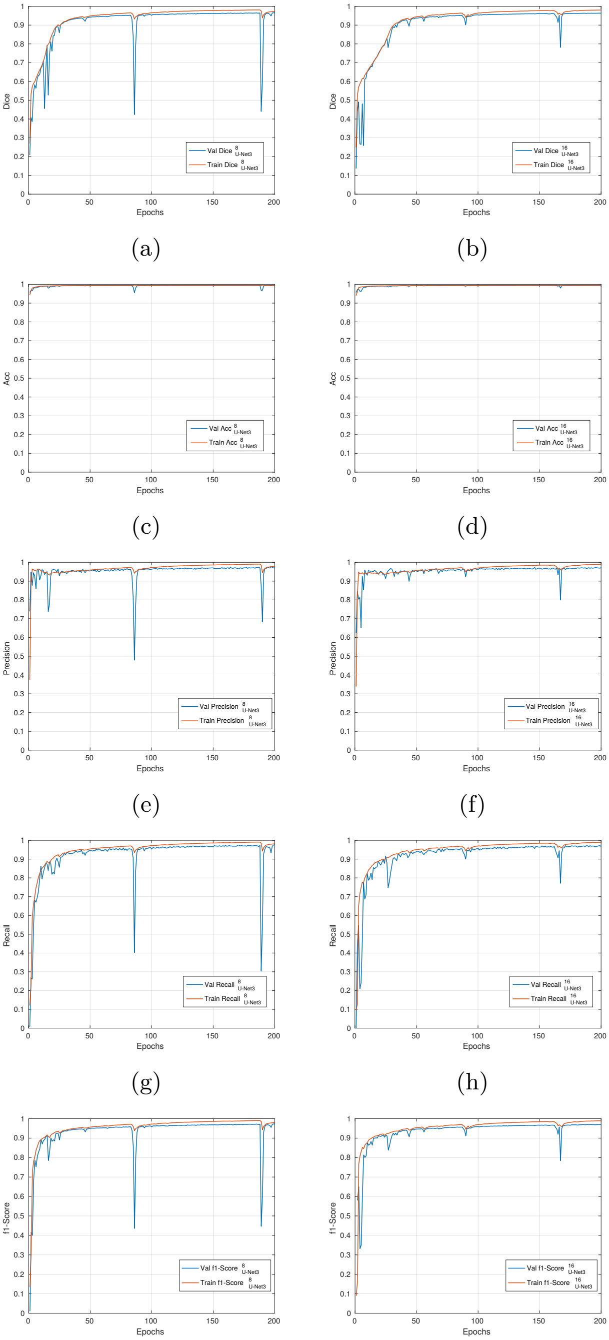
(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure A.3.: U-Net2 results.



(a) (b)

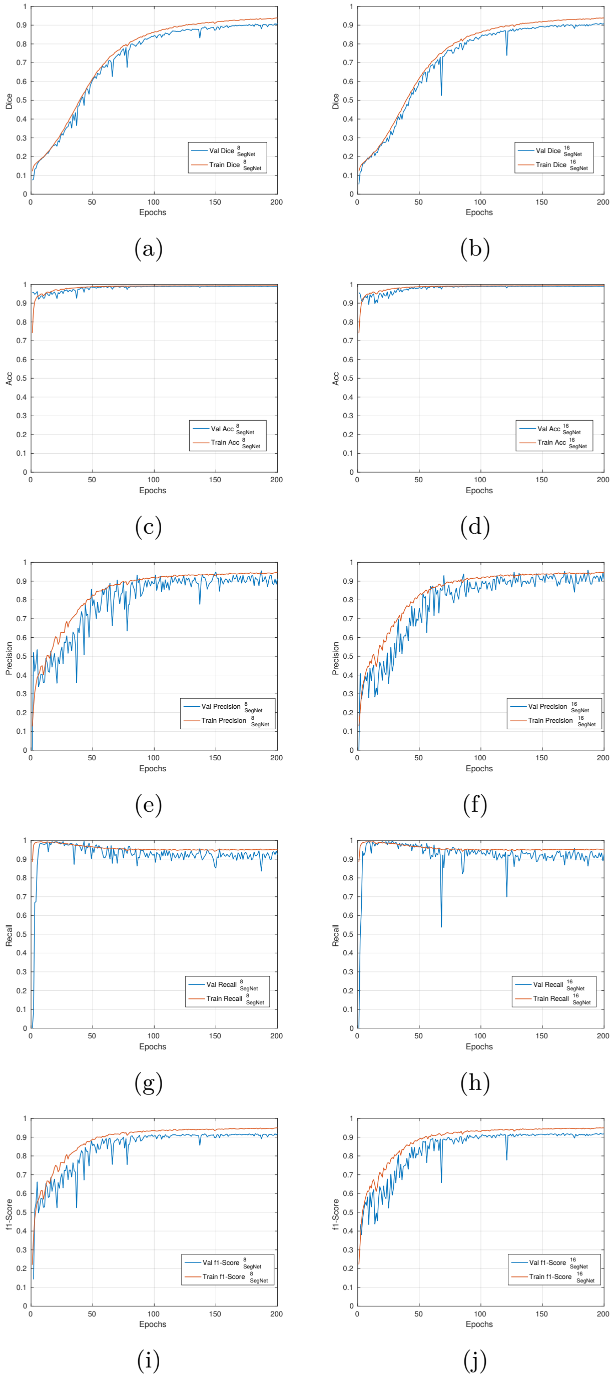
(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure A.4.: U-Net3 results.



(a) (b)

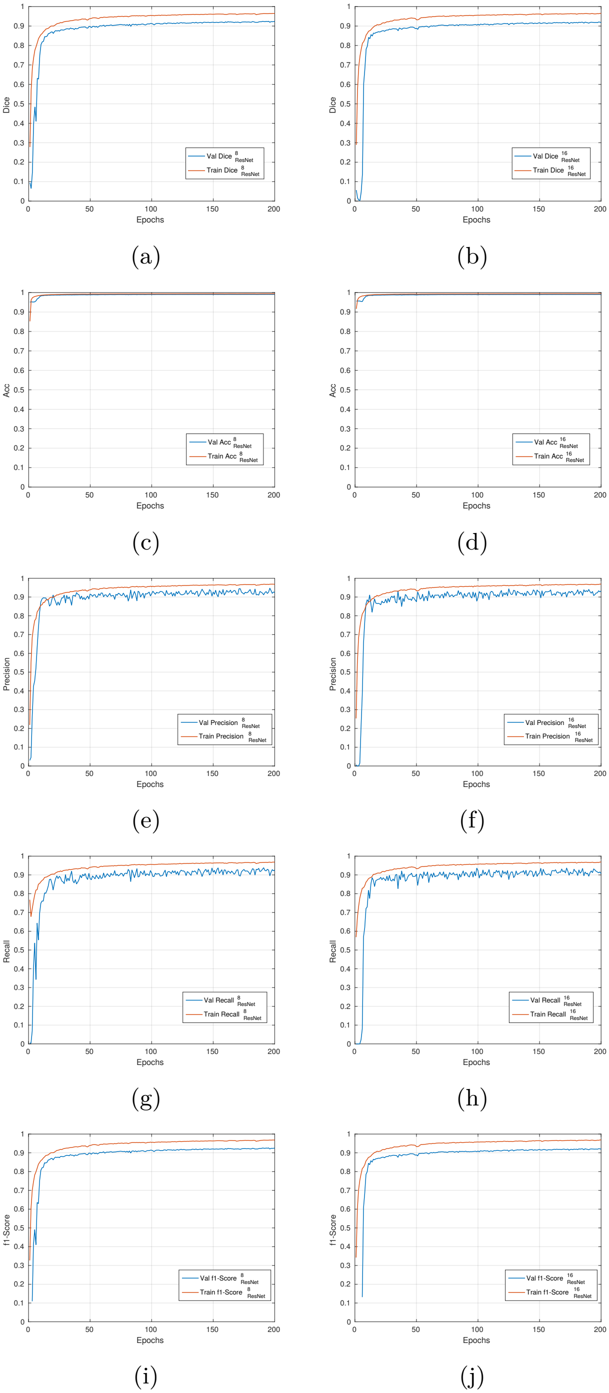
(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure A.5.: SegNet results.



(a) (b)

(c) (d)

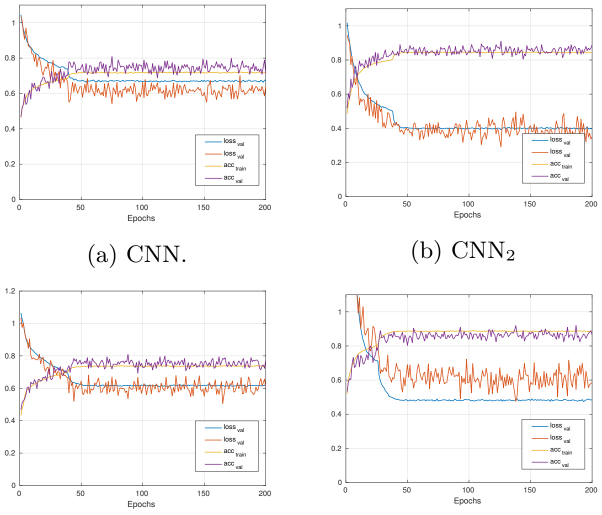
(e) (f)

(g) (h)

(i) (j)

Figure A.6.: ResNet results.

A.2. User-shelf interaction results



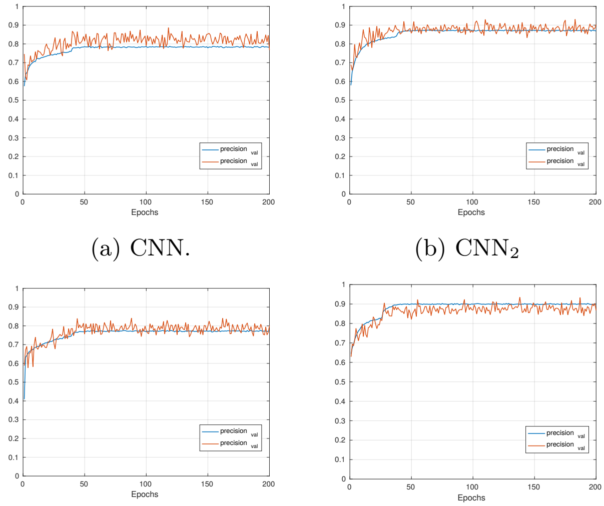
(a) CNN.

(b) CNN$_2$

(c) AlexNet.

(d) CaffeNet.

Figure A.7.: Accuracy and loss results.



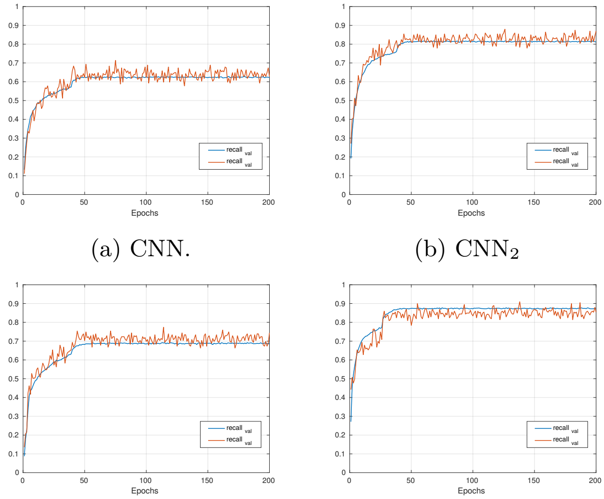
(a) CNN.

(b) CNN_2

(c) AlexNet.

(d) CaffeNet.

Figure A.8.: Precision results.



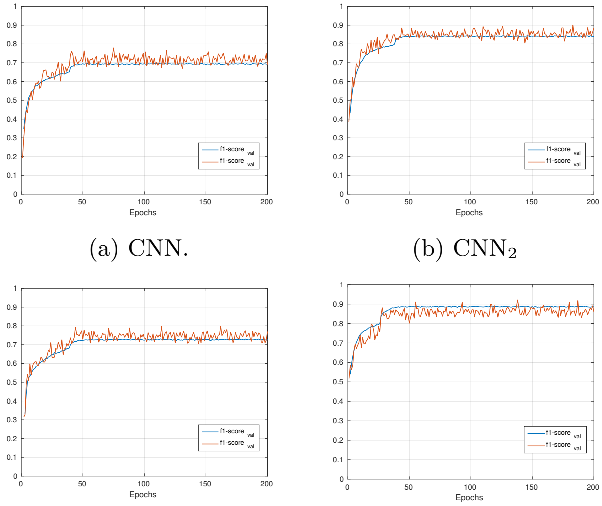
(a) CNN.

(b) CNN_2

(c) AlexNet.

(d) CaffeNet.

Figure A.9.: Recall results.



(a) CNN.

(b) CNN$_2$

(c) AlexNet.

(d) CaffeNet.

Figure A.10.: F1-Score results.

Bibliography

[1] B. A. Y. Agusta, P. Mittrapiyanuruk, and P. Kaewtrakulpong. Field seeding algorithm for people counting using kinect depth image. *Indian Journal of Science and Technology*, 9(48), 2016.

[2] R. Ali, M. ElHelw, L. Atallah, B. Lo, and G.-Z. Yang. Pattern mining for routine behaviour discovery in pervasive healthcare environments. In *2008 International Conference on Information Technology and Applications in Biomedicine*, pages 241–244. IEEE, 2008.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, 2015.

[4] D. Baltieri, R. Vezzani, and R. Cucchiara. Learning articulated body models for people re-identification. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 557–560. ACM, 2013.

[5] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 433–442. Springer, 2012.

[6] L. E. Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[7] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013.

[8] J. Bednarık and D. Herman. Human gesture recognition using top view depth data obtained from kinect sensor. *Proc. of Excel@ FIT*, 2015.

[9] A. Bevilacqua, L. Di Stefano, and P. Azzari. People tracking using a time-of-flight depth sensor. In *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*, pages 89–89. IEEE, 2006.

[10] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann. Fall detection based on body part tracking using a depth camera. *IEEE journal of biomedical and health informatics*, 19(2):430–439, 2015.

[11] R. Bodor, B. Jackson, and N. Papanikolopoulos. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, volume 1. Citeseer, 2003.

[12] A. Bonnin, R. Borràs, and J. Vitrià. A cluster-based strategy for active learning of rgb-d object detectors. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1215–1220. IEEE, 2011.

[13] P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology*, 23(11):1993–2008, 2013.

[14] A. Bourke, J. O'brien, and G. Lyons. Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. *Gait & posture*, 26(2):194–199, 2007.

[15] A. Burbano, S. Bouaziz, and M. Vasiliu. 3d-sensing distributed embedded system for people tracking and counting. In *Computational Science and Computational Intelligence (CSCI), 2015 International Conference on*, pages 470–475. IEEE, 2015.

[16] M. Castrillón-Santana, J. Lorenzo-Navarro, and D. Hernández-Sosa. People semantic description and re-identification from point cloud geometry. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4702–4707. IEEE, 2014.

[17] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[18] P. Chandon, J. Hutchinson, E. Bradlow, and S. H. Young. Measuring the value of point-of-purchase marketing with commercial eye-tracking data. *INSEAD Business School Research Paper*, (2007/22), 2006.

[19] C. Coppola, T. Krajnık, T. Duckett, and N. Bellotto. Learning temporal context for activity recognition. In *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands-Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, volume 285, page 107. IOS Press, 2016.

[20] C. Coppola, O. Martinez Mozos, N. Bellotto, et al. Applying a 3d qualitative trajectory calculus to human action recognition using depth cameras. In *IEEE/RSJ IROS Workshop on Assistance and Service Robotics in a Human Environment*, 2015.

[21] A. Coşkun, A. Kara, M. Parlaktuna, M. Ozkan, and O. Parlaktuna. People counting system by using kinect sensor. In *Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on*, pages 1–7. IEEE, 2015.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[23] A. D'Angelo and J.-L. Dugelay. People re-identification in camera networks based on probabilistic color histograms. In *IS&T/SPIE Electronic Imaging*, pages 78820K–78820K. International Society for Optics and Photonics, 2011.

[24] J. Dartigues. [methodological problems in clinical and epidemiological research on ageing]. *Revue d'épidémiologie et de santé publique*, 53(3):243–249, 2005.

[25] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento. Counting people by rgb or depth overhead cameras. *Pattern Recognition Letters*, 81:41–50, 2016.

[26] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[27] F. Dittrich, H. Woern, V. Sharma, and S. Yayilgan. Pixelwise object class segmentation based on synthetic data using an optimized training strategy. In *Networks & Soft Computing (ICNSC), 2014 First International Conference on*, pages 388–394. IEEE, 2014.

[28] P. F. Felzenszwalb. Learning models for object recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[29] E. Frontoni, A. Mancini, and P. Zingaretti. Rgbd sensors for human activity detection in aal environments. In *Ambient Assisted Living*, pages 127–135. Springer, 2014.

[30] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi. Customers' activity recognition in intelligent retail environments. In *New Trends in Image Analysis and Processing–ICIAP 2013*, pages 509–516. Springer, 2013.

[31] H. Fu, H. Ma, and H. Xiao. Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. *Multimedia Tools and Applications*, 73(1):273, 2014.

[32] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[33] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, and F. Flórez-Revuelta. Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using kinect. 2015.

[34] S. Gasparrini, E. Cippitelli, S. Spinsante, and E. Gambi. A depth-based fall detection system using a kinect® sensor. *Sensors*, 14(2):2756–2775, 2014.

[35] P. Georgieff. *Ambient Assisted Living: Marktpotenziale IT-unterstützter Pflege für ein selbstbestimmtes Altern.* MFG-Stiftung Baden-Württemberg, 2008.

[36] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1528–1535. IEEE, 2006.

[37] K. Giannakouris. Eurostat: Statistics in focus population and social conditions 72/2008, 2008.

[38] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014.

[39] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision–ECCV 2008*, pages 262–275. Springer, 2008.

[40] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2012–2019. IEEE, 2009.

[41] S. Güven, O. Oda, M. Podlaseck, H. Stavropoulos, S. Kolluri, and G. Pingali. Social mobile augmented reality for retail. In *PerCom*, pages 1–3. IEEE Computer Society, 2009.

[42] I. Haritaoglu and M. Flickner. Attentive billboards: Towards to video based customer behavior understanding. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 127–131. IEEE, 2002.

[43] M. Hasan and A. K. Roy-Chowdhury. A continuous learning framework for activity recognition using deep hybrid feature models. *IEEE Transactions on Multimedia*, 17(11):1909–1922, 2015.

[44] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] K. Heath and L. Guibas. Multi-person tracking from sparse 3d trajectories in a camera sensor network. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–9. IEEE, 2008.

[46] D. Hernandez, M. Castrillon, and J. Lorenzo. People counting with re-identification using depth cameras. 2011.

[47] M. Hoai and F. De la Torre. Max-margin early event detectors. *International Journal of Computer Vision*, 107(2):191–202, 2014.

[48] M. Iwai, M. Mori, and H. Touda. A marketing analysis using massive tiny sensor nodes. In *Networked Sensing Systems (INSS), 2009 Sixth International Conference on*, pages 1–4. IEEE, 2009.

[49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[50] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *Robotics: Science and Systems*, 2013.

[51] M. Kepski and B. Kwolek. Detecting human falls with 3-axis accelerometer and depth sensor. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 770–773. IEEE, 2014.

[52] M. Kepski and B. Kwolek. Fall detection using ceiling-mounted 3d depth camera. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, pages 640–647. IEEE, 2014.

[53] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[54] D. Kouno, K. Shimada, and T. Endo. Person identification using top-view image with depth information. In *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on*, pages 140–145. IEEE, 2012.

[55] P. Kourouthanassis and G. Roussos. Developing consumer-friendly pervasive retail systems. *IEEE Pervasive Computing*, 2(2):32–39, 2003.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[57] J. Krockel and F. Bodendorf. Customer tracking and tracing data as a basis for service innovations at the point of sale. In *SRII Global Conference (SRII), 2012 Annual*, pages 691–696. IEEE, 2012.

[58] A. Krüger, J. Schöning, and P. Olivier. How computing will change the face of retail. *IEEE Computer*, 44(4):84–87, 2011.

[59] G. Larsson, M. Maire, and G. Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

[60] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[61] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, and V. Placidi. Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network. In *International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage*, pages 146–157. Springer, Cham, 2014.

[62] D. Liciotti, G. Ferroni, E. Frontoni, S. Squartini, E. Principi, R. Bonfigli, P. Zingaretti, and F. Piazza. Advanced integration of multimedia assistive technologies: A prospective outlook. In *Mechatronic and Embedded Systems and Applications (MESA), 2014 IEEE/ASME 10th International Conference on*, pages 1–6. IEEE, 2014.

[63] D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti. Pervasive system for consumer behaviour analysis in retail environments. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, volume 2. 2017.

[64] D. Liciotti, E. Frontoni, P. Zingaretti, N. Bellotto, and T. Duckett. Hmm-based activity recognition with a ceiling rgb-d camera. In *ICPRAM (International Conference on Pattern Recognition Applications and Methods)*, 2017.

[65] D. Liciotti, G. Massi, E. Frontoni, A. Mancini, and P. Zingaretti. Human activity analysis for in-home fall risk assessment. In *Communication Workshop (ICCW), 2015 IEEE International Conference on*, pages 284–289. IEEE, 2015.

[66] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti. Person re-identification dataset with rgb-d camera in a top-view configuration. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 1–11. Springer, 2016.

[67] D. Liciotti, P. Zingaretti, and V. Placidi. An automatic analysis of shoppers behaviour using a distributed rgb-d cameras system. In *Mechatronic and Embedded Systems and Applications (MESA), 2014 IEEE/ASME 10th International Conference on*, pages 1–6. IEEE, 2014.

[68] S.-C. Lin, A.-S. Liu, T.-W. Hsu, and L.-C. Fu. Representative body points on top-view depth sequences for daily activity recognition. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 2968–2973. IEEE, 2015.

[69] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[70] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen. Detecting and tracking people in real time with rgb-d camera. *Pattern Recognition Letters*, 53:16–23, 2015.

[71] J. Lorenzo-Navarro, M. C. Santana, and D. Hernández-Sosa. An study on re-identification in rgb-d imagery. In *IWAAL*, pages 200–207. Springer, 2012.

[72] T. Määttä, A. Härmä, H. Aghajan, and H. Corporaal. Collaborative detection of repetitive behavior by multiple uncalibrated cameras. *Information Fusion*, 21:68–81, 2015.

[73] C. Madden and M. Piccardi. Height measurement as a session-based biometric for people matching across disjoint camera views. In *Image and Vision Computing New Zealand*, pages 282–286. Citeseer, 2005.

[74] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3177–3184. IEEE, 2011.

[75] F. Malawski. Top-view people counting in public transportation using kinect. *Challenges of Modern Technology*, 5, 2014.

[76] N. Marquardt, K. Hinckley, and S. Greenberg. Cross-device interaction via micro-mobility and f-formations. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 13–22. ACM, 2012.

[77] S. W. Marshall, C. W. Runyan, J. Yang, T. Coyne-Beasley, A. E. Waller, R. M. Johnson, and D. Perkis. Prevalence of selected risk and protective factors for falls in the home. *American journal of preventive medicine*, 28(1):95–101, 2005.

[78] S. Messelodi and C. M. Modena. Boosting fisher vector based scoring functions for person re-identification. *Image and Vision Computing*, 44:44–58, 2015.

[79] C. Migniot and F. Ababsa. 3d human tracking from depth cue in a buying behavior analysis context. In *Computer Analysis of Images and Patterns*, pages 482–489. Springer, 2013.

[80] C. Migniot and F. Ababsa. Hybrid 3d—2d human tracking in a top view. *Journal of Real-Time Image Processing*, 11(4):769–784, 2016.

[81] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer vision and image understanding*, 81(3):231–268, 2001.

[82] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126, 2006.

[83] M. Mubashir, L. Shao, and L. Seed. A survey on fall detection: Principles and approaches. *Neurocomputing*, 100:144–152, 2013.

[84] H. Nait-Charif and S. J. McKenna. Activity summarisation and fall detection in a supportive home environment. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 323–326. IEEE, 2004.

[85] H. Ning, T. X. Han, D. B. Walther, M. Liu, and T. S. Huang. Hierarchical space-time model enabling efficient search for human actions. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(6):808–820, 2009.

[86] N. Noury, A. Fleury, P. Rumeau, A. Bourke, G. Laighin, V. Rialle, and J. Lundy. Fall detection-principles and methods. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 1663–1666. IEEE, 2007.

[87] N. Oliver and E. Horvitz. A comparison of hmms and dynamic bayesian networks for recognizing office activities. In *International conference on user modeling*, pages 199–209. Springer, 2005.

[88] W. H. Organization. 10 facts on ageing and the life course., 2007. Accessed: 2015-01-30.

[89] F. Pala, R. Satta, G. Fumera, and F. Roli. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):788–799, 2016.

[90] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[91] L. Piyathilaka and S. Kodagoda. Human activity recognition for domestic robots. In *Field and Service Robotics*, pages 395–408. Springer, 2015.

[92] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, 2012.

[93] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person reidentification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[94] N. M. Puccinelli, R. C. Goodstein, D. Grewal, R. Price, P. Raghubir, and D. Stewart. Customer experience management in retailing: understanding the buying process. *Journal of retailing*, 85(1):15–30, 2009.

[95] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[96] P. Rashidi and D. J. Cook. An adaptive sensor mining framework for pervasive computing applications. In *Knowledge Discovery from Sensor Data*, pages 154–174. Springer, 2010.

[97] M. Rauter. Reliable human detection and tracking in top-view depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 529–534, 2013.

[98] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya. Learning and incorporating shape models for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 203–211. Springer, 2017.

[99] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[100] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043. IEEE, 2011.

[101] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*, pages 322–329. IEEE, 2009.

[102] A. W. Senior, L. Brown, A. Hampapur, C.-F. Shu, Y. Zhai, R. S. Feris, Y.-L. Tian, S. Borger, and C. Carlson. Video analytics for retail. 2007.

[103] D. Siegmund, A. Wainakh, and A. Braun. Verification of single-person access in a mantrap portal using rgb-d images. In *XII Workshop de Visao Computacional (WVC)*, 2016.

[104] H. Steg, H. Strese, C. Loroff, J. Hull, and S. Schmidt. Europe is facing a demographic challenge ambient assisted living offers solutions. *IST project report on ambient assisted living*, 2006.

[105] T. Strandvall. Eye tracking as a tool in package and shelf testing. *White Paper*, (Toby technology), 2008.

[106] M. Strohbach and M. Martin. Toward a platform for pervasive display applications in retail environments. *IEEE Pervasive Computing*, 10(2):19–27, 2011.

[107] M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini, M. Contigiani, and P. Zingaretti. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognition Letters*, 2016.

[108] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[109] Q. Tian, B. Zhou, W.-h. Zhao, Y. Wei, and W.-w. Fei. Human detection using hog features of head and shoulder based on depth map. *JSW*, 8(9):2223–2230, 2013.

[110] T.-E. Tseng, A.-S. Liu, P.-H. Hsiao, C.-M. Huang, and L.-C. Fu. Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 4077–4082. IEEE, 2014.

[111] P. H. Tu, G. Doretto, N. O. Krahnstoever, A. A. Perera, F. W. Wheeler, X. Liu, J. Rittscher, T. B. Sebastian, T. Yu, and K. G. Harding. An intelligent video framework for homeland protection. In *Defense and Security Symposium*, pages 65620C–65620C. International Society for Optics and Photonics, 2007.

[112] P. Vera, S. Monjaraz, and J. Salas. Counting pedestrians with a zenithal arrangement of depth cameras. *Machine Vision and Applications*, 27(2):303–315, 2016.

[113] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern recognition*, 36(3):585–601, 2003.

[114] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[115] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[116] C. Wateosot and N. Suvonvorn. Top-view based people counting using mixture of depth and color information. In *The Second Asian Conference on Information Systems, ACIS*, 2013.

[117] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[118] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[119] T. Yahiaoui, C. Meurie, L. Khoudour, and F. Cabestaing. A people counting system based on dense and close stereovision. *Image and Signal Processing*, pages 59–66, 2008.

[120] J. Yamamoto, K. Inoue, and M. Yoshioka. Investigation of customer behavior analysis based on top-view depth camera. In *Applications of Computer Vision Workshops (WACVW), 2017 IEEE Winter*, pages 67–74. IEEE, 2017.

[121] C.-W. You, H.-L. C. Kao, B.-J. Ho, Y.-H. T. Chen, W.-F. Wang, L.-T. Bei, H.-H. Chu, and M.-S. Chen. Convenienceprobe: a phone-based system for retail trade-area analysis. *Pervasive Computing, IEEE*, 13(1):64–71, 2014.

[122] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li. Water filling: Unsupervised people counting via vertical kinect sensor. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 215–220. IEEE, 2012.

[123] Z. Zhang, U. Kapoor, M. Narayanan, N. H. Lovell, and S. J. Redmond. Design of an unobtrusive wireless sensor network for nighttime falls detection. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 5275–5278. IEEE, 2011.

[124] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 649–656. IEEE, 2011.