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

Deep Understanding of Shopper Behaviours and Interactions in Intelligent Retail Environment

Ph.D. Dissertation of:
Rocco Pietrini

Advisor:
Prof. Emanuele Frontoni

Curriculum Supervisor:
Prof. Francesco Piazza

XVIII edition - new series

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

Deep Understanding of Shopper Behaviours and Interactions in Intelligent Retail Environment

Ph.D. Dissertation of:
Rocco Pietrini

Advisor:
Prof. Emanuele Frontoni

Curriculum Supervisor:
Prof. Francesco Piazza

XVIII 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 family.

“Not all bits have equal value.”
(Carl Sagan)

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Prof. Emanuele Frontoni for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

A very special gratitude goes to my colleague Marina, who has always supported me along the way, without her precious support it would not have been possible to conduct this research.

I am grateful to the VRAI (Vision Robotics and Artificial Intelligence) research group, present and past members, who welcomed me since the very first moment with precious feedback, cooperation and of course friendship. I wish in particular to express my gratitude to Prof. Primo Zingaretti and Prof. Adriano Mancini who have been precious collaborators and source of inspiration. A special thanks goes to Daniele, I learned so much from him since the very first moment. Thanks to Roberto, because a brief chat three years ago really convinced me to start this journey. I would like to express my gratitude to my office buddies Lucia, Michele and Sara for the good moments spent together. Thanks to Massimo, without his collaboration the results presented in this thesis would not have been achieved. I would also like to thank Paola, for her advices in the writing of this thesis.

My sincere thanks also go to Prof. Mubarak Shah, who provided me an opportunity to join the research team at the Center for Research in Computer Vision in the University of Central Florida for a visiting period, and who gave me access to the laboratory and research facilities. It's been a truly open-minding experience of research and life in general.

I would also like to express my sincere gratitude to Grottini Lab for co-funding my scholarship and letting me conduct this research with real applications all around the world. Thanks to Valerio for always trusting and giving me this opportunity. My sincere thanks goes to Luigi and Marco, I learned so much from them and we spent unforgettable moments around the world installing the systems described in this thesis. Thanks to Francesco, my lunch buddy, and the rest of the team: Mauro, Davide, Marco R., Simona, Erica, Lorenzo, Matteo, Omar, and Saeed for being part of my daily life in Grottini

Lab.

I am also pleased to say thank you to all the people I met in Orlando during my visiting period, everyone of them gave me a different perspective about life and contributed to making my visit enjoyable and productive. I would always remember Hugo, Marli and Kelly who welcomed me in their home. Alessandro, Antonio, Rudy, Jorge, Ivilina, Liza, Ayana and Alen: thank you guys for the great moments spent together.

Many thanks to the Fenix team for the good time spent on the futsal field, I always think about our motto *Post fata resurgo*!.

I would like to thank my friends Daniele, Giovanni, Sergio, Valeriano, Chiara, Lucia, Gianluca, Sara, Giulia and Francesca for the best moments spent outside my Ph.D life, thank you guys for being part of my life.

Last but not the least, I would like to thank my family: my parents and my sister for supporting me spiritually throughout writing this thesis and my life in general.

Ancona, Novembre 2019

Rocco Pietrini

Abstract

In retail environments, understanding how shoppers move in the store's spaces and interact with products is very valuable. While the retail environment has several favourable characteristics that support computer vision, such as reasonable lighting, the large number and diversity of products sold, as well as the potential ambiguity of shoppers' movements, mean that accurately measuring shopper behaviour is still challenging. Over the past years, machine-learning and feature-based tools for people counting as well as interactions analytics and re-identification were developed with the aim of learning shopper behaviors based on occlusion-free RGB-D cameras in a top-view configuration. However, after moving into the era of multimedia big data, machine-learning approaches evolved into deep learning approaches, which are a more powerful and efficient way of dealing with the complexities of the human behaviour.

Starting from such a premise, this thesis addresses the evolution process of 3 real systems such as: People Counting, Shopper Analytics and Re-Identification. The main goal is to develop Deep Learning architectures especially designed for the retail environment. A novel VRAI deep learning framework is described for this purpose. In particular, it uses 3 Convolutional Neural Networks (CNNs) to count the number of people passing or stopping in the camera area, perform top-view re-identification and measure shopper-shelf interactions from a single RGB-D video flow with near real-time performances. The VRAI framework is evaluated on the following 3 new datasets that are publicly available: TVHeads for people counting, HaDa for shopper-shelf interactions and TVPR2 for people re-identification.

The proposed applications open up a wealth of novel and important opportunities for the machine vision community. The newly datasets collected as well as the complex areas taken into exam, make the research challenging. In fact, it is crucial to evaluate the performance of state of the art methods to demonstrate their strengths and weaknesses and help identify future research to design more robust algorithms. For a comprehensive performance evaluation, it is of great importance to develop benchmarks to gauge the state of the art because methods designed for specific domains do not work properly on others. Furthermore, the dataset selection is needed in order to offer the user the opportunity to prove the validity of the proposed methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction.

Understanding the consumer behaviour is of great importance and one of the keys to success for retailers [2]. Many efforts have been devoted in particular toward monitoring how shoppers move through the retail space and interact with products. This challenge is still open due to several serious problems, which include occlusions, appearance changes and dynamic and complex backgrounds. Popular sensors that are used for this task are RGB-D cameras because of their 3-D scene modeling capability. The great value (both in accuracy and efficiency) of using depth cameras in coping with severe occlusions among humans and complex backgrounds has been demonstrated in several studies. Additionally, while the retail environment has several favourable characteristics for computer vision (such as reasonable lighting), the large number and diversity of products sold and the potential ambiguity of shopper movements mean that accurately measuring shopper behaviours is still challenging.

The advent of low-cost RGB-D devices, such as Microsoft's Kinect and Asus's Xtion Pro Live sensors, has led to a revolution in computer vision and vision-related research. The combination of depth and color information has led to new challenges and opportunities for action detection and people tracking in many retail applications based on human-environment interactions.

Several research manuscripts show that the top-view configuration was adopted to tackle these challenges because it facilitates tasks making easier to extract different features. This setup choice also increases the robustness, because it minimizes occlusions among individuals, it has also the advantage of being non intrusive in the installation in a retail environment. Reliable depth maps can provide valuable additional information that can significantly improve detection and tracking results [3]. Top-view RGB-D applications are the most accurate (up to 99% accuracy) in people counting applications, especially in very crowded scenarios (defined as more than 3 people per square metre).

Over the past years, machine-learning and feature-based tools were developed with the aim of learning shopper skills in intelligent retail environments. Each application uses RGB-D cameras in a top-view configuration that are installed in different locations of a given store, providing large volumes of multidimensional data that can be used to gather and deduce insights [1, 4, 5]. These data are analysed with the aim of examining the *attraction* (the level of attraction that the shopper is showing for a store category based on the rate between the total amount of shoppers that entered the store and those who passed by the category), the *attention* (the amount of time that shoppers spend in front of a brand display) and the *action* (the consumers visit the store and interact with the products, either buying or not a product). People's re-identification (re-id) in different categories is also crucial to understand the shopping journey of every customer. Based on these insights, new store layouts could be designed to improve product exposure and placement and to promote products by actively acquiring and maintaining users' attention [6].

However, after moving into the era of multimedia big data, machine-learning approaches evolved into deep learning approaches, which are a more powerful and efficient way of dealing with the massive amounts of data generated from modern approaches and coping with the complexities of understanding human behaviour. Deep learning has taken key features of the machine-learning model and has even taken it one step further by constantly teaching itself new abilities and adjusting existing ones [7].

This thesis presents a novel VRAI[1] deep learning framework with the goal of improving existing solutions previously developed by the aforementioned research group [1, 4, 5], this evolution of machine intelligence also provides a solid guide for discovering powerful insights in the current big data era.

According to the Pareto principle, stores are mapped with a focus on targeted Stock Keeping Units (SKUs) that offer greater profit margins.

A camera installation layout in one of the stores where the experiments were performed is depicted in Figure 1.1. This store has an area of about 1500 m^2, and 24 RGB-D (Asus Xtion Pro Live) cameras were installed in a top-view configuration without any overlapping area. A typical camera installation at a store entrance is depicted in Figure 1.2.

In order to maximize the space coverage using this relatively small number of cameras, 2 RGB-D cameras were placed at the store's entrances to identify and count the shoppers, and, to measure the shoppers' attractions, attentions and interactions, the other 22 cameras were placed on the ceiling above the shelves of interest, counting the number of people and re-identifying them in every top-seller category.

This test installation, together with 4 other stores located in Italy, China, Indonesia and the US, became the basis for the datasets and results presented in this thesis, based on a 3-year experience that measured 16 million shoppers and about 2 million interactions.

[1] This name of the framework is related to the Vision Robotics and Artificial Intelligence (VRAI) research group of Universitá Politecnica delle Marche.

In order to conduct a comprehensive performance evaluation, it is critical to collect representative datasets. While much progress has been made in recent years regarding efforts in sharing codes and datasets, it is of great importance to develop libraries and benchmarks to gauge state-of-the-art datasets.

Newly challenging datasets were specifically designed for the tasks described in this study. In fact, each described application involved the collection of one dataset, which was used as the input. Thus, the learning methods described were evaluated according to the following proposed datasets: the Top-View Heads (TVHeads) dataset, the Hands dataset (HaDa) and the Top-View Person Re-Identification 2 (TVPR2) dataset.

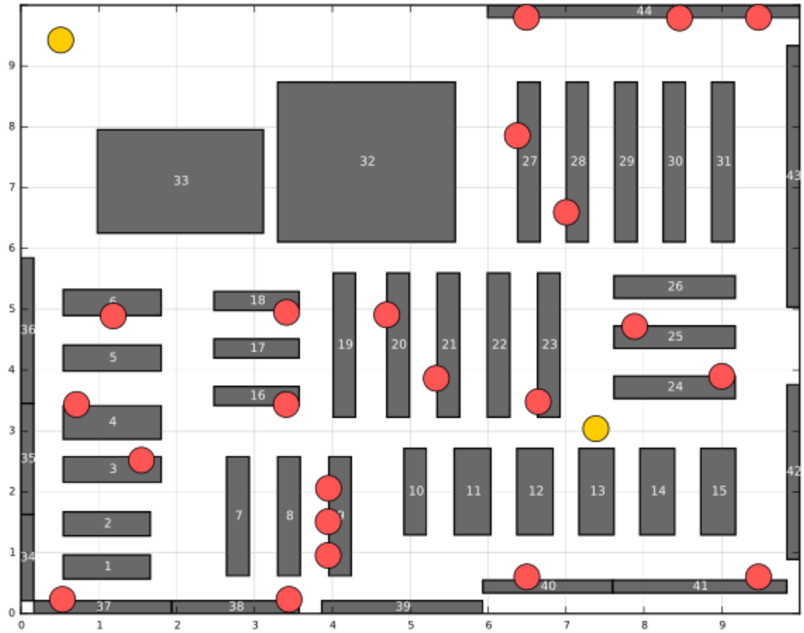


Figure 1.1: Camera installations in the target store where the experiments were performed. This store has an area of about 1500 m^2 and was covered with a total of 24 RGB-D cameras that were installed in a top-view configuration. In particular, 2 RGB-D cameras were used for counting and identifying shoppers at the store's entrances (marked in yellow), and the other 22 cameras, in order to measure shoppers' attractions, attentions and interactions, were installed above the shelves, counting the people and re-identifying them in every top-seller category (marked in red).

Based on these evaluation configurations and datasets, this thesis introduce a novel VRAI deep learning framework that uses 3 CNNs to count people passing by the camera area, perform top-view re-id and measure shopper-shelf interactions in a single RGB-D frame simultaneously. The system is able to process data at 10 frames per second, ensuring high performances even in cases of very brief shopper-shelf interactions.

Experimental results showed that the proposed VRAI networks significantly outperformed all competitive state-of-the-art methods with an accuracy of 99.5% on people counting, 92.6% on interaction classification and 74.5% on re-id.

This thesis, presents the first study on understanding shoppers' behaviours using an RGB-D camera installed in a top-view configuration. As discussed in this chapter, the choice to use a top-view configuration was because of its greater suitability than a front-view configuration, as the former reduces the problem of occlusions and has the advantage of a non intrusive installation.

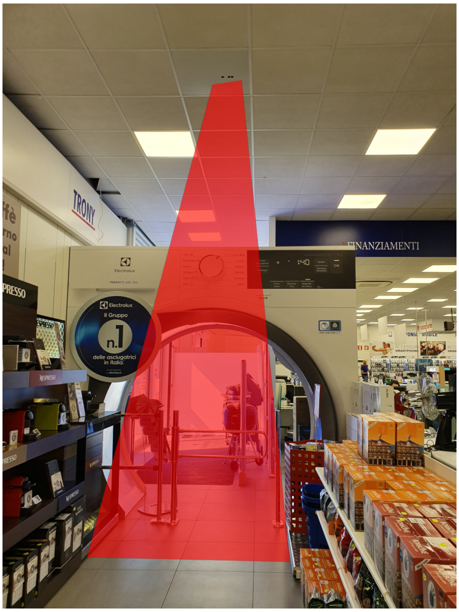


Figure 1.2: Typical entrance camera installation for people counting and re-identification.

The research presented in this thesis has been conducted in cooperation and collaboration with the company Grottini Lab[2], which makes the retail intelligence and the shopper behavior understanding the core of its business, with running installations in different countries around the world.

The thesis is organized as follows. Chapter 2 provides an overview of the state of the art of deep learning applications in retail environments. Chapter 3 describes the approach to evolve systems toward VRAI deep Learning, and

[2] http://www.grottinilab.com

offers details on “VRAI datasets”, 3 new, challenging datasets that are publicly available. In Chapter 4, limitations, challenges and lesson learnt are discussed. Finally, in Section 5, conclusions and future directions for this field of research are introduced.

Chapter 2

From a Geometric and Features-based Approach to Deep Learning in Retail Environment: State of art and Perspectives.

Studying the customer behaviour within a retail store is a very important topic for academic research. Understanding how shoppers move, which shelves and products they interact with and what choices they make during the purchase phase, can lead to an improvement of the different marketing strategies. Based on this knowledge, for example, retailers can choose which products to promote and where to place them, they can modify the store layout by pushing the choice towards certain products and outline the possible trajectories for shopping. All this in order to offer the shopper a more attractive but never intrusive shopping experience [8].

Over time, various models have been applied for predicting human behaviour and interactions in retail environment. Considering the current era of big data combined with the development of advanced systems, the amount of data is increasing. The availability of huge datasets and their processing using graphics processing units (GPU) have promoted the development of new modelling approaches. In 2006, Hinton et al. introduced the deep belief networks that made it possible to construct nets with many hidden layers [9]. This resulted in a new theory and caught the attention of many researchers and several companies. The advantages and disadvantages of these algorithms along with their functions are reported in Tables 2.1 and 2.2.

Table 2.1: Machine Learning Approaches for customer understanding, along with their function, advantages, disadvantages, and examples.

Approach	Function	Advantages	Disadvantages
Association (e.g. statistics and apriori algorithms [10])	Establishing relationships between items which exist together in a given record.	Greatly compress the candidate item sets and the size of the frequent item sets, and obtain good performance.	Requires many database scan.
Classification (e.g., neural networks, DT and if-then-else rules [11])	Building a model to predict future customer behaviours through classifying database records into a number of predefined classes based on certain criteria.	Ability to implicitly detect complex nonlinear relationships between dependent and independent variables.	Proneness to overfitting.
Clustering (e.g. neural networks and discrimination analysis) [11])	Segmenting a heterogeneous population into a number of more homogenous clusters.	Easy to implement.	Need to define many channels.
Forecasting (e.g. neural networks and survival analysis [11])	Estimating the future value based on a record's patterns.	The projections rely on the strength of past data.	Some forecasting methods may use the same data but deliver widely different forecasts.
Regression (e.g. linear regression and logistic regression [12])	Statistical estimation technique used to map each data object to a real value provide prediction value.	Large amounts of potential predictor variables management, fine-tuning the model to choose the best predictor variables from the available options.	Overfitting the Model.
Sequence discovery (e.g. statistics and set theory [13])	Identification of associations or patterns over time.	Maximize either the precision or the recall and limit the degradation of the other criterion.	

Table 2.2: Deep Learning Approaches for customer understanding, along with their function, advantages, disadvantages, and examples.

Deep Learning Approach	**Function**	**Advantages**	**Disadvantages**
Convolutional Neural Networks (e.g. VGG16, AlexNet and ResNet [14])	Image Classification.	They take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme.	Lack of ability to be spatially invariant to the input data.
Recurrent Neural Networks (e.g., LSTM [15])	Memorizing previous inputs in memory, When a huge set of Sequential data is given to it.	Ability to model sequence of data (i.e. time series) so that each sample can be assumed to be dependent on previous ones.	It cannot process very long sequences if using tanh or relu as an activation function.
Generative Adversarial Networks [16]	Generating data that is similar to real data.	Learning density distributions of data.	The samples generated are of relatively low quality.

For this purpose, in the following Sections this evolution will be summarized and the insights will be provided in the domain taken into exam in this thesis. In particular, Section 2.1 describes the features-based approaches adopted for the development of applications in retail; Section 2.2 presents an overview of the retail applications that use deep learning algorithms. Last Section will be devoted to the description of the recent technologies in industry (Section 2.3).

2.1 Geometric and Features-based approaches in retail environment

The analysis of consumers behavior within a retail store includes several aspects: human interaction, behaviour understanding, modeling of the environment, people detection and tracking. The monitoring of people is an active area of research in the computer vision community as explained by Smeulders et al. in their survey [17]. Several problems must be taken into account by the algorithms to study the human behaviour: illumination changes, occlusion, cumber and privacy preserving just to name a few [18].

The first important work that use RGB-D devices to analyse shopper behaviour is that reported by Vildjiounaite et al. in [19]. The authors predict the future customer locations, developed for environments where items are frequently re-located and customer routes change accordingly. The tracking system uses an adaptive background model, that is able to distinguish between moving humans and re-positioned objects. The approach intends to solve also the occlusion problems.

The work by Liu et al. [20] has the aim to automatically detect and track people in a dynamic and indoor retail environment using a single RGB-D camera. In this paper, the authors propose an innovative approach that uses a single RGB-D camera to automatically detect and track people, that have different poses in dynamic environments. They develop a novel point ensemble image (PEI) representation by transforming RGB-D pixels in order to overcome segmentation problems met using the RGB-D image. Experimental results are obtained using a purposely dedicated real-world clothing store dataset.

More recently, the quality of depth maps has greatly improved through the use of depth cameras such as Kinect [21], Xtion, Orbbec to mention structured light technology as well as active stereo cameras such as Intel RealSense series, which are available at affordable prices. These cameras have demonstrated great value (efficiency-wise and accuracy-wise) in coping with severe occlusions among humans and complex backgrounds as described by Sturari et al. in [22]. According to the authors, the use of reliable and precise indoor localization systems allow to monitor customer behavior inside retail stores. They use a Kalman filter to obtain an integrated system of active beacon sensors and RGB-D cameras for this purpose. The main reason is that machine vision approaches provide a high level of accuracy while beacons have a large coverage area. Experimental results have demonstrated that this combined system increases the localization performance.

To overcome occlusion problems, in the work of Liciotti et al. [3], the authors present a literature review on the use of RGB-D camera for detecting and monitoring people. Their intention is to demonstrate that the top-view configuration is the best for people monitoring mostly in crowded situations and in the presence of occlusions. Moreover, this kind of camera configuration is also privacy preserving.

Also in the work of Dan et al. [23] the cameras are positioned on the ceiling and they demonstrate that the top-view configuration is robust when the environment is crowded and there are changes in illumination. To solve problems related to optical noise and data loss, a morphological operator processes the depth image. The subject is then extracted using a human model of the depth image obtained after processing. Experimental results have shown very high precision.

From the point of horizontal view, the work of Han et al. [24] deals with the problem of detection and localization of human beings in a domestic environment, also exploiting the combination of depth and color provided by an RGB-D camera. The cameras have a horizontal configuration that is optimal in situations where environments, such as domestic ones, are not crowded and where privacy is not required.

The work of Ravnik et al. [25] has the task to predict the consumer behaviour using machine learning methods on real-world digital signage viewership data to predict consumer behaviour in a retail environment, especially oriented towards the purchase decision process and the roles in purchasing situations. They compare the performance of different machine learning algorithms, obtaining the best performance using the SVM classifier [26].

To determine the position of the shopping cart inside a store, RFID tags are used in the work of Li et al. [27]. Information about the movement inside the store are given by the tags and the knowledge about the most visited areas allows to create an optimization model.

In literature there are many works that evaluate the relationship between shoppers and products in the retail store. These approaches can be divided in those focusing on the physical interactions and others related to the product detection.

The work of Liciotti et al. [28] uses an integrated system consisted of an RGB-D camera and a software able to recognize the interaction of the shopper with the products on the shelves. They classify the interaction in 3 classes (positive, negative and neutral) on the base of the result of the interaction. Experimental results have demonstrated that the low cost and easy to install system provides good results in recognizing the action.

The use of wireless embedded sensors, i.e. a series of smaller beacons, is proposed by the work of Pierdicca et al. [29]. The complex infrastructure aims to analyze the interactions of shoppers with products that are in the shelves.

Melià-Seguí and Pous [30] analyse a system that studies the real-time interactions of the shopper with products based on a combination of 3 RFID sensors. The features extracted are used as input to supervised machine learning approaches, obtaining encouraging results in the experimental phase.

In the research of Yamamoto et al. [31] to analyze the human interactions an environment with book shelves is utilized. The system uses a top-view configuration with depth cameras positioned on the ceiling. The method is based on 2 behavior estimators: the first one is based on the height of the hand, exploiting depth information to distinguish the shelf level and the second is based on SVM with depth and PSA (Pixel State Analysis) based features to detect the human silhouette.

George and Floerkemeier [32] proposed an efficient approach for per-exemplar multi-label image classification, which targets the recognition and localization of products in retail store images. They used discriminative random forests, deformable dense pixel matching and genetic algorithm optimization, achieving promising results in terms of both accuracy and runtime efficiency. They also provide an available novel dataset and labeling tool for products image search, with 8350 different images of single products on a white background, 680 annotated images and 885 test images.

In another work Varol and Kuzu [33], present a low time complexity technique to recognize retail product on grocery shelf images. The problem is divided in detection and recognition tasks. A generic product detection module, trained only on a particular class of products, is presented. The system is realized based on HOG descriptors [34] and the Cascade object detection framework [35]. To recognize the brand inside the detected region they use SVM. Shape and color information is obtained and then the fusion is applied starting from 2 separate descriptors calculated with the bag of words approach. An available dataset of over 5000 images composed of 10 tobacco brands was produced. The authors pointed out that detection and classification can be achieved with satisfactory results even on devices with low computational power. In the future, using the same methodology, the authors intend to implement applications that include planogram compliance control, inventory management and assistance to blind people during shopping. The system has improved the identification of products through an approach based on SIFT features [36] for the description of the shape and on the HSV values for the description of the colors.

2.2 Deep learning approaches in retail environment

In the context of retail store applications, the work of Cruz et al. [37] presents a system that combines deep learning and augmented reality techniques to increase the shoppers experience. Through a deep learning architecture, the implemented system learns the visual appearance of different sectors in the store. Then, the shoppers using their mobile devices take a picture of the area and the system can identify the area where the customer is located. Using the location information and the augmented reality it is possible to provide to the shopper useful information about: route to another area where a product is available, 3D product visualization, user location and analytics. The system gives high performance in terms of location accuracy and deep learning combined with augmented reality techniques optimize the customer experience.

In the work of Nogueira et al. [38] a low-cost deep learning approach to estimate in real-time the number of people in retail stores and to detect and visualize hot spots is presented. The architecture employs an RGB camera and to solve the people counting problem, a deep learning approach based on a Convolutional Neural Network (CNN) is used. They use a four channel image representation named RGBP image, composed of the conventional RGB image and an extra binary image P representing whether there is a visible person in each pixel of the image. The images are used to detect the hot spots of the store. Several experimental results are presented to evaluate the performance of the system using a surveillance camera placed in a real retail store.

Song et al. [39] have developed an accurate, real-time and fully automated system that uses machine vision and deep learning approach to identify returning customers, filter sales and customer information. The system has been divided into 2 sub-modules to reduce the problems related to the limited computing resources and to the accuracy of the recognition. The local computing resources are used to perform customer tracking and monitoring and for feature extraction. Other statistics such as age and gender estimation and customer recognition are made on cloud resources.

The research of Kim et al. [40] is focused on person detection in an indoor retail environment. The authors evaluate and compare deep neural network (DNN) person detection models in cost-effective, end-to-end embedded platforms such as the Nvidia Jetson TX2 and the Intel Movidius. They use a proprietary image dataset captured by conventional security cameras in retail environments to compare the performance of state of the art DNN models including YOLO, SSD, RCNN, R-FCN and SqueezeDet. These images were manually annotated to form the ground truth for training and evaluation of the deep models. Experimental results show that neither of these models nail the tradeoff between speed and accuracy demonstrating the complexity of the problem.

In the work Liu et al. [41], the authors propose a system called TAR that learns the interest of buyers in a retail store by monitoring and identifying people, using different cameras. The combined use of the visual information and Bluetooth information from smartphones allows to accurately track and identify the buyers in the shop. TAR uses visual tracking based on a deep neural network and is able to identify and recognize the identity of the shopper. The experiments were carried out in a real environment. Evaluation results demonstrated that TAR delivers high accuracy (90%) and serves as a practical solution for people tracking and identification.

Karim et al. [42] presented an automated approach to obtain statistics about customer satisfaction using image processing and deep learning approach. The system combines different module for face detection and tracking, best view estimation, customer identification, blacklisted customer warnings, and facial sentiment classification with the aim to increase the shopper satisfaction. However the author don't give details on the accuracy achieved by the system.

A method based on deep learning has been presented by Generosi et al. [43]. The authors propose a system capable of processing images extracted from biometric data and facial expressions to evaluate the shopping experience in different monitoring points of a store. To evaluate the performance of the implemented method, a preliminary test was carried out on a real context and then the results obtained were compared with those of a traditional video analysis. The aim of the work is to support and improve the customer experience in a retail context.

The work of Allegrino et al. [44] proposes an innovative architecture capable of integrating the face detection and the movement to model consumers in real time. The vending machine and the decision support system process data to suitably respond with dynamic prices and product proposals to the particular shopper. A convolutional neural network is trained for the people detection phase.

The work presented by Karlinsky et al. [45] aims to quickly detect and recognize thousands of categories of products by training the system with a limited number of examples (1 per category). Their task is to detect retail products, with only 1 image available per product, to detect brand logos and finally to detect 3D objects and the position within a static 2D image. This work is a challenge as a deep learning detector requires a large amount of data in order to correctly generalize. Synthetic data are used to train the network. The model they use is probabilistic, not parametric, the improvement is based on convolutional neural networks and in some cases temporal integration is carried out.

The work of Qiao et al. [46] introduces the problem of object proposal generation in supermarket images and other natural images. They believe that estimating the scale of objects in images is useful for generating proposals for objects, especially in supermarket images where object scales are limited to a small range. For this reason they evaluate the dimension of the objects in the images before the object proposal. The method is named ScaleNet, a variation of the ResNet architecture, that is trained on a synthetic dataset and tested on 2 real datasets. A scale prediction phase is added to the known object proposal methods, increasing the performance. Moreover, they make publicly available both the synthetic and the real datasets.

The problem of visual recognition of packaged grocery products is still an open problem for modern artificial vision systems. This is due to the fact that the number of elements to be recognized is enormous and changes rapidly over time. To solve this problem, the authors of the work Tonioni et al. [47] propose an end-to-end architecture that uses generative adversarial networks (GAN) that consider the change of domain and a convolutional neural network trained on the generated samples that learns the images of the products with a hierarchy between product categories. The positive experimental results accounted both the recognition of the products present in the training data sets and the different ones not seen during the training phase.

The work presented by Franco et al. [48] identifies and recognizes grocery products on the shelf. They compare a classical Bag of Words technique with a more recent approach to deep neural networks, obtaining rather interesting results. Furthermore, they believe that the available training images are acquired under controlled conditions and therefore are unable to generalize the problem well. To solve this problem and therefore to have complex and robust recognition techniques, they exploit the information on color and texture in a multi-step process: pre-selection, fine selection and post processing.

In the work of Sharif et al. [49], the authors assume that the descriptors extracted from the convolutional neural networks are very powerful. They implement the OverFeat network trained for classifying objects on the ILSVRC13 dataset. The extracted features are used as representation of generic images to be used for image classification activities, considering categories of different images. The categories were chosen considering datasets that gradually move away from the dataset that the network is trained to recognize. The classification results outperformed the algorithms trained and tested on the same category. A linear SVM classifier is used on a representation of features of size 4096, extracted from a layer of the network. According to the authors, the experimental results demonstrated that deep learning-based algorithms provides the best performance for most visual recognition activities.

The study conducted in Tonioni et al. [50] presents a deep learning object detection pipeline based on YOLO [51] and the KNN algorithm [52] to distinguish the class of the product and the associated brand. To recognize the products in the shelf, they initially implement state-of-the-art object detectors to achieve agnostic product detection, then they search similar features through global descriptors calculated on the reference image and the one acquired. They train the convolutional neural network on reference images based on an image embedding loss with the aim to optimize the recognition performance.

The authors of the work Chong et al. [53] use 3 models of convolutional neural networks to classify images of products on the shelves of retail stores. Moreover, to determine which image is the most suitable, they make a preliminary evaluation of different type of images. They use 3 types of training sets using a collection of images performed by Navii, a robot developed by *Fellow Robots*, and images from the internet.

In their work, Cotter et al. [54] present significant performance improvements considering machine learning techniques and using features similar to HOG. The problem is that the system has a large number of training images for each product, making the pipeline rather slow at the time of the test. Furthermore, the addition of a new product causes a further slowdown of the entire system. To improve the performance, the approach has been extended to Advani et al. [55], using a product correlation.

In the work of Kong et al. [56], the authors use deep neural networks for video item removal detection in retail environments. Unlike the Amazon Go application which uses both weight measurement and computer vision, in this paper they consider only computer vision with deep learning to help customers to explore and shop more efficiently. The input of the network is a video stream while the output is a prediction about an item added or removed from the shelf. They also compare the performance of 2 video classification algorithms (late fusion and 3D convolutional network) in the prediction accuracy, in order to evaluate the effectiveness of both algorithms.

In the work Ribeiro et al. [57], the authors propose an adaptive system based on deep learning for the recognition of food packages through information such as the name, the list of ingredients and the use by date. They believe that this information is essential to ensure correct product distribution in retail. The system is based on deep learning algorithms and specifically uses a methodology that exploits some characteristics of a convolutional neural network. They demonstrate the greater precision of the new methodology compared to the original deep neural network.

Loureiro et al. [58] consider the fashion retail industry increasingly competitive and so the companies must adequate products features to increase the customer satisfaction and the loyalty. Although the lifecycles of fashion products are very short, the definition of inventory and purchasing strategies can be supported by the large amounts of historical data which are collected and stored in companies' databases. In this context, the aim of this work is to implement a deep learning approach to predict the sales of individual products in the fashion sector for future seasons. The variables considered for the development of this model are different from each other and concerns both the physical characteristics of the product and the opinions of the experts. Furthermore, to evaluate the performance of the implemented method, the performances obtained with deep learning techniques are compared with a set of shallow techniques (Decision Trees, Random Forest, Support Vector Regression, Artificial Neural Networks and Linear Regression). The deep learning method has better results for predicting the retail sales of a fashion product.

The approach proposed by Femling et al. [59] intends to create a system that recognizes fruit and vegetables in a retail market using the images acquired by a video camera connected to the system. To obtain the best performance, 2 deep learning architectures were compared, in particular the 2 convolutional neural networks, Inception and MobileNet, trained to recognize 10 different types of fruits and vegetables. The aim of the work is to minimize the human-machine interactions in the phase of labeling the desired product, based on weight and typology. The usability of the system was tested in the experimental phase.

In the work of Agnihotram et al. [60], the system proposed intends to provide a solution to automate the activities that are typically performed by employees of a retail store. The time-consuming and labor-intensive activities involve keeping track of the quantity of products on the shelves, replenishing them when out of stock and moving products out of place. The system is based on classification techniques, deep learning techniques and computer vision algorithms. Specifically, they use a double robot that monitors the retail store, moving along a fixed path and acquiring images of the shelves in real time. Furthermore, to notify that a product is out of place or out of stock, there is a mechanism that generates an alert, so that the employee act in a timely manner.

The study of Paolanti et al. [61] has the aim to automatically detect the absence of a product in the shelf, known as SOOS (shelf-out-of-stock). They use a deep learning algorithm, in particular 2 convolutional neural networks trained to recognize the textual and visual features of a product, then identified through a machine learning classifier. Several machine learning algorithms are compared.

Most recent is the work of Hu et al. [62] where they propose a method to recognize different objects in a pair of images of a retail store. They implement a single deep convolutional neural network named DiffNet that has in input the pair of images and gives in output the bounding boxes of different products. DiffNet is trained using a file that contains the label of different objects, not all objects of the input images, reducing in this way human efforts.

2.3 Retail industry

In addition to the academic research, the retail industry is demonstrating a growing interest in those new technologies to understand the shopper behaviors. The retail's ongoing transformation feature a shift in focus from the point of sale to the point of experience. Retailers that will succeed in the digital economy will be those that think beyond the products they sell to provide e personalized shopping experience that surprise and delight the consumer at each interaction, regardless of channel or touchpoint. With an increasing focus on convenience, community, curation and immersion, these retail experiences will become synonymous with the retailer's differentiated brand promise and, therefore, core to building consumer loyalty and advocacy. In these perspective many startups came into the market recently, one of the most well known is RetailNext. RetailNext uses video analytics, Wi-Fi detection of mobile devices (including Bluetooth), data from point-of-sale systems, and other sources to provide insights to retailers about how customers engage with their stores. Their proprietary software focus on the people traffic and runs on the edge on custom devices called *Aurora*. In addition it is worth to mention Amazon, that with its concept store *Amazon Go* proposed the first ever cashierless convenience store. Although there are no much information about the employed technologies, according to promotional video published by Amazon and public speeches[1], the store concept uses several technologies, including computer vision, deep learning algorithms, and sensor fusion to automate much of the purchase. Amazon tackled six core problems: Sensor Fusion, Calibration, Person detection (re-identification), Object Recognition, Pose estimation, Activity Analysis.

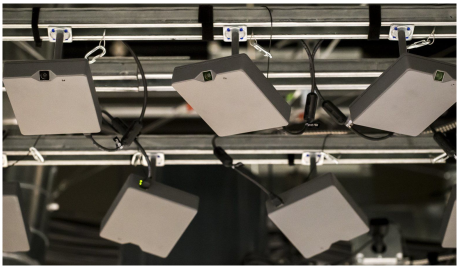


Figure 2.1: Amazon Go ceiling.[2]

Amazon uses thousands of sensors to accomplish these tasks in relatively small stores (100 square metre on average). A ceiling of an Amazon Go store is depicted in figure 2.1 while a shelf is reported in figure 2.2 and can be easily noted the complexity and intrusiveness of the installation, with such high number of sensors.

High level results obtained by Amazon does not necessarily mean this solution is ready for all kinds of products and store formats. That setup may be difficult to replicate in other settings, such as a fashion store with lots of hanging items on racks. In fact Amazon builds from scratch the entire store designing every single aspect to be functional to the technology employed. In this thesis the described systems can be applied to existing stores.

In retail analytics cash flow data must be also mentioned, although it may seem a sufficient source of information to derive the customers behavior, they

[1] re:Mars conference 2019

[2] https://www.pinterest.com/pin/689684130408741795/

[3] http://www.clresearch.com/research/detail.cfm?guid=6A608036-3048-78A9-2FB3-4E6295D65919

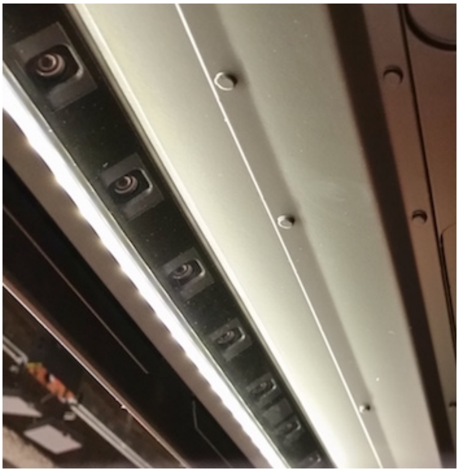


Figure 2.2: Amazon Go shelf cameras.[3]

are not. Cash flow data reports only what shoppers buy, but doesn't answer a fundamental question: “why ?”. It is important to know what the purchase path was, what factors drove the shopper to the purchase. Maybe a new attractive packaging, or maybe the store/shelf layout were not optimal for that product. That is the reason for the need of a more complex analysis.

2.4 Contributions

In this work, novel VRAI deep learning approaches are introduced into existing applications [1, 4, 5] to evolve the machine-learning and features-based approaches into deep learning approaches, which are a more powerful and efficient way of dealing with the massive amounts of data generated from modern approaches. The technologies discussed in this thesis have great relevance in the marketing field. In particular, they offer relevant contributions in the field of behavioural science and, more precisely, to consumer behaviour studies by using innovative methodologies and tools. This marketing aspect will be also discussed and analysed in the following Chapters. The main contributions of this thesis with respect to the state-of-the-art approach can be summarized as follows:

- solutions, for real retail environments with a great variability in data acquired, derived from a large experience over 16 million shoppers observed in 3 years in different types of stores and in different countries;
- an initial integrated framework for the deep understanding of shopper behaviours in crowded environments with high accuracy, actually limited to

count and re-identify people passing by and to analyse their interactions with shelves;

- 3 relevant datasets from real scenarios that are publicly available to the scientific community for testing and comparing different approaches;
- 3 concurrent CNNs for processing the same frame to: i) segment and count people with high accuracy (more than 99%) even in crowded environments); ii) measure and classify interactions between shoppers and shelves classifying positive, negative, neutral and refill actions with a good accuracy also compared with cashier sell-out; iii) perform a re-identification over contemporary shoppers (up to 1000 people in the same area at the same time) with a good accuracy to detect massive behavioural data on the best performing categories (more than 80% with 100 or 250 contemporary shoppers in the area).

Chapter 3

Use cases and Results. From a Geometric and Features-based Approach to VRAI Deep Learning.

This section describes this evolution process as well as the datasets used for the evaluation. The framework is depicted in Figure 3.1 and comprises 3 main systems: People Counting [4], interaction classification [1] and Re-id [5]. Three specially designed new VRAI-Nets are presented: *VRAI-Net 1*, *VRAI-Net 2* and *VRAI-Net 3*, which are applied to every frame coming from every RGB-D camera in the store in order to move these systems toward deep learning.

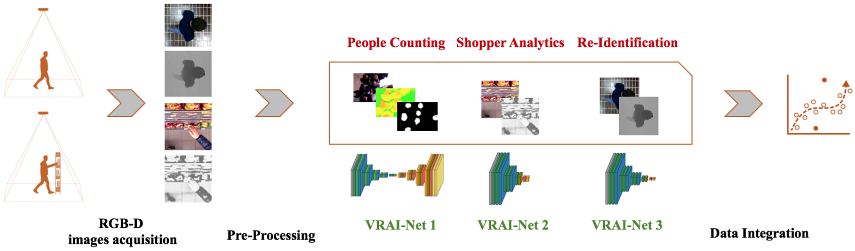


Figure 3.1: VRAI framework for the deep understanding of shoppers' behaviour.

In fact, to address increasingly complex problems and when dealing with big data, deep learning approaches can provide a powerful framework for supervised learning. For example, when measuring what happens in front of the shelf, many fake interactions could occur because of unintentional interactions or changes in the background.

Three new datasets from images and videos acquired by RGB-D cameras that were installed in a top-view configuration in different areas of the target store will be presented. The "VRAI datasets" comprises the following 3 datasets:

- *TVHeads*[1] Dataset;
- HaDa[2] Dataset;
- TVPR2[3] Dataset.

The VRAI framework is comprehensively evaluated on the new "VRAI datasets", collected for this work. The details of the data collection and ground truth labelling are discussed in the following sections.

Systems based on the functionality described here have been deployed at a number of stores around the world, and many have been in operation for over 3 years. Several days of video data were acquired from 24 cameras (2 used solely as counters and 22 used for counting, interaction classification and re-id) and processed by the system. In the following sections, the results of VRAI deep learning framework are evaluated and compared with the state-of-the-art framework.

3.1 People Counting

The methods dealing with people counting problems can be divided into 2 groups: detection-based methods and mapping-based methods. The first ones refer to running a detector, counting or clustering the output. The different features can include body, head, skin, hair and etc. For effective detection algorithms, they can have a high output accuracy for not highly crowded environments, but are not scalable for large crowds [63]. The mapping-based methods are referring to features extraction and mapping them to a value. They use edge points, background, texture, optical flow, etc. as the features. Compared to detection-based methods, these methods can be scalable to large crowds. To address people detection and tracking problems, sensors widely adopted are RGB-D cameras. Compared to conventional cameras, their performance results in increased reliability, availability and affordability. The efficiency and accuracy of depth cameras have been proven to be elevated in cases with severe occlusions among humans and complex background [22]. The combination of high-resolution depth and visual information opens up new opportunities in many applications in the field of activity recognition and people tracking. Tracking and detection results can be significantly improved by the use of reliable depth maps [3].

In the literature several datasets using RGB-D technology exist for the study of person re-id mainly in the front view configuration such as VIPeR [64], the

[1] http://vrai.dii.univpm.it/tvheads-dataset
[2] http://vrai.dii.univpm.it/content/hada-dataset
[3] http://vrai.dii.univpm.it/content/tvpr2-dataset

iLIDS multi-camera tracking scenario [65], ETHZ [66], CAVIAR4REID [67] and that presented by B. I. Barbosa et al. in [68]. They cover many aspects of the existing problems such as shape deformation, occlusions, illumination changes, very low resolution images and image blurring.

There have been many vision techniques and algorithms proposed in the literature in last years for person detection and tracking. In general, we can distinguish the following: segmentation using background subtraction, water filling, statistical algorithms, machine learning and finally deep learning techniques.

3.1.1 VRAI-Net 1 for semantic Heads Segmentation using Top-View Depth Data in Crowded Environment

Semantic segmentation has proven to be a high-level task when dealing with 2D images, 2D videos, and even 3D data [69]. It paves the way toward the complete understanding of scenes and is being tackled using deep learning architecture, primarily Deep Convolutional Neural Networks (DCNNs), because they perform more accurately and sometimes even more efficiently compared to machine-learning and features-based approaches [70]. An efficient segmentation leads to the complete understanding of a scene; moreover, since the segmentation of an image takes place at the pixel level, each object will have to be assigned to a class. Thus, its boundaries will be uniquely defined. To obtain high quality output, this thesis presents the design of a novel *VRAI-Net 1* starting from the U-NET 3 proposed by Liciotti et al. in [4].

VRAI-NET 1 presents a batch normalisation layer at the end of each layer after the first Rectified Linear Unit (ReLU) activation function and after each max-pooling and upsampling function. In this way, it obtains a better training performance and yields more precise segmentations. Furthermore, the classification layer is modified. In fact, it is composed of 2 convolutional layers with hard sigmoid functions. This block is faster to compute than simple sigmoid activation, and it maps the features of the previous layer according to the desired number of classes. Compared to the U-Net [71], the number of filters of the first convolution block were halved in the current study. A simpler network, going from 7.8 million parameters to 2 million is obtained.

The *VRAI-Net 1* architecture is shown in Figure 3.2.

VRAI-NET 1 is even more robust and efficient than the work proposed in [4]. In the current work, the expansive path of the network was modified after being replaced by a refinement procedure. This procedure is composed of 4 blocks, and each block combines 2 types of the features map. It is basically formed by 2 branches: the first uses an up-convolution layer to up-sample the activations of the previous layer and a ReLU function to avoid the vanishing gradient problem, and the second branch joins the corresponding layer of the contracting path with a dropout layer to prevent the overfitting problem. These 2 branches are merged through an Equivalent Layer Thickness (ELT) layer, which determines the element-wise sum of the outputs. The output of each refinement layer is the input of the first branch of the next refinement layer. Also added to this network was the use of a particular dropout technique instead of the standard technique, based on the random zeroing of certain activations. The spatial dropout method of [72] is used, performing standard Bernoulli trials on the training phase then propagating the dropout value on the entire feature map. In this way, a dropout with a 50% ratio zeroes half the channels of the previous layer. The dropout of spatial correlations was aimed toward increasing the robustness of the network in a shorter amount of time than the standard method. Finally, the channels of the layers of the contraction part were increased by a factor of 4 compared to the expansive part. Then, a 1×1 convolutional layer with a single channel was added both between the 2 sides and at the end of the expansive part. A good tradeoff can be achieved between computational efficiency and better segmentation predictions since the first part of the network processes large enough feature maps compared to the second part, but the latter still maintains a suitable number of parameters to perform a good up-sampling.

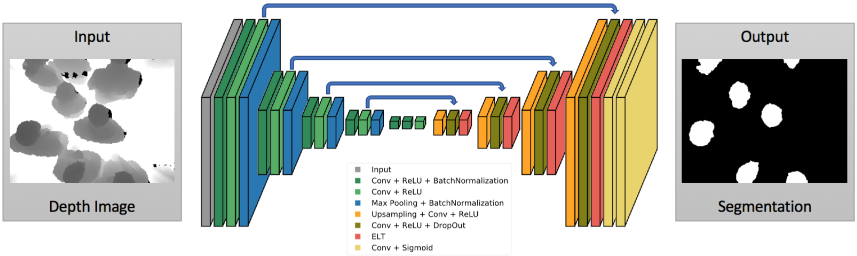
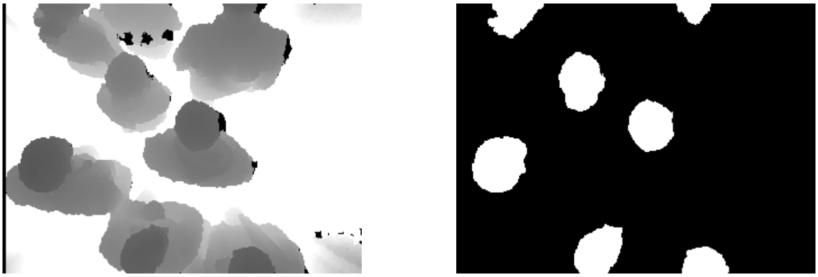


Figure 3.2: VRAI-Net 1 architecture. It is composed of 2 main parts: a contracting path (left side) and an expansive path (right side). Its main contribution is its use of a refinement process in the expansive path; each step is a combination of 2 branches, one from the upsampling and the other from the corresponding layer of the contracting path. The combination is performed using an element-wise sum. Another improvement is the use of spatial dropout layers instead of standard ones, which are aimed toward improving the robustness of the network in a shorter amount of time.

3.1.2 TVHeads Dataset

The *TVHeads* dataset contains 1815 depth images (16 bit) with size of 320 × 240 pixels captured from an RGB-D camera in a top-view configuration. The images collected in this dataset represented a crowded retail environment with at least 3 people per square metre and physical contact between them. Following the pre-processing phase, a suitable scaling method was applied to the images, in order to switch to 8 bits per pixel instead of the original 16. In this way, a more highlighted profile of the heads is obtained, improving the contrast and brightness of the image. The ground truth, for the head detection, was manually labelled by human operators.

Figure 3.3 shows an example of a dataset instance that includes the 2 images described above (8-bit depth image and the corresponding ground truth).



(a) 8-bit depth image.

(b) Ground truth.

Figure 3.3: TVHeads dataset. It consists of an 8-bit scaled depth image 3.3a and the corresponding ground truth 3.3b.

3.1.3 Performance evaluation and Results

In this subsection, the results of the experiments conducted using the TVHeads dataset will be reported. In addition to the performance of *VRAI-Net 1*, also presented here is the performance of the different approaches taken from the literature based on CNNs such as SegNet [73], ResNet [74], FractalNet [75], U-Net, U-Net 2 [71, 76] and U-Net 3 [4] in the attempt to solve the problem of head image segmentations.

Each CNN is trained using 2 types of depth images to highlight the head silhouettes: 16-bit (original depth image) and 8-bit (scaled image). In this way, image contrast and brightness are increased. In the training phase, the dataset was split into training and validation sets with a ratio of 10%. The learning process was conducted for 200 epochs using a learning rate equal to 0.001 and an Adam optimization algorithm. Semantic segmentation performances are shown in Table 3.1, which also reports the Jaccard [77] and Dice [78] indices for training and validation, respectively, as well the results in term of accuracy, precision, recall and F1-score. As we can infer, the new *VRAI-Net 1* network outperformed the state of the art networks in terms of the Jaccard and Dice indices and in terms of accuracy. The *VRAI-Net 1* reached 0.9381 for the Jaccard index and 0.9691 for the Dice index. The accuracy of the network instead reached a value of 0.9951, thus demonstrating the effectiveness and suitability of the proposed approach. The comparison shows that *VRAI-Net 1* performed better than the previous U-Nets design. Among the various tests performed, we can see that the best performance used mainly images scaled to 8 bits.

Table 3.1: Jaccard and Dice indice comparison and segmentation results obtained for different DCNN architectures.

CNN	Bit	Jac.	Dice	Acc.	Prec.	Rec.	F1-Score
Fractal [75]	8	0.9480	0.9733	0.9944	0.9914	0.9931	0.9922
	16	0.9477	0.9732	0.9944	0.9927	0.9933	0.9930
SegNet [73]	8	0.8237	0.9033	0.9927	0.9463	0.9531	0.9496
	16	0.8277	0.9058	0.9927	0.9462	0.9533	0.9497
ResNet [74]	8	0.8563	0.9226	0.9938	0.9684	0.9684	0.9684
	16	0.8482	0.9179	0.9938	0.9688	0.9693	0.9690
U-Net [71]	8	0.8694	0.9301	0.9927	0.9465	0.9505	0.9484
	16	0.8695	0.9302	0.9926	0.9450	0.9490	0.9469
U-Net2 [76]	8	0.9391	0.9686	0.9931	0.9700	0.9692	0.9696
	16	0.9382	0.9681	0.9932	0.9679	0.9706	0.9691
U-Net3 [4]	8	0.9314	0.9645	0.9946	0.9905	0.9904	0.9904
	16	0.9299	0.9637	0.9946	0.9894	0.9894	0.9894
VRAI-Net 1	**8**	**0.9381**	**0.9691**	**0.9951**	**0.9918**	**0.9921**	**0.9918**
	16	**0.9290**	**0.9642**	**0.9946**	**0.9893**	**0.9895**	**0.9894**

Main applications of accurate people counting in crowded scenario related to the shopper marketing area are: i) the accurate funnel evaluation at store and category level, starting from people entering the space; ii) the store and category A/B testing for performance comparison; iii) the store flow modelling also for high traffic areas (e.g. promo areas). To better understand the applications of the proposed method and the high impact on the shopper marketing area other aggregated results of the proposed framework in section 4.3.1 are reported.

3.2 Shopper interactions classification

Understanding shoppers behavior has been always of interest for retailers and one of the most interesting aspect is to detect their activities with the products on the shelves. Analyzing customers using existing methods such as interviews or using cash flow data cannot give a comprehensive view of their behavior. Interviews requires a constant human presence at the point of sale and they allow only to sample a small fraction of the customers, resulting in a time-consuming and inefficient method due to the intrusiveness of the approach.

Mistery Shopping, which is another commonly used traditional method, consists in evaluating the overall shopping experience or other shoppers while remaining discreet and pretending to be a regular customer. This method also suffer form the same limitations described above for interviews.

The aforementioned limitations motivated the need of a way to study the purchase-path of a customer in a non-intrusive and automatic manner, as described in this thesis.

Action detection in retail environment has been investigated in a recent research by Singh et al.[79], where the authors collected a dataset in a laboratory environment simulating a retail space, annotated for action detection specifically for retail (this dataset will be discussed and used later on in this chapter). They proposed a multistream bi-directional recurrent neural network using the RGB color stream. While this approach seems promising, lacks of a fundamental information for retail: the matching between shopper actions and products, which is crucial and complex without relying on the 3-D coordinates given by depth sensors or product recognition techniques. In addition, the dataset was collected in laboratory and not in a real environment and thus doesn't represent the real world complexity.

The action detection problem in retail has been also tackled by Frontoni et al. in [80] using top-view RGB-D sensors and introducing a primary classification about interactions:

- *Positive*: shopper take a product form the shelf;
- *Negative*: shopper put a product back in the shelf;

A “virtual wall” (threshold) is considered to be in front of the shelf with the help of the depth sensor 3D coordinates system. When the shopper crosses this wall, ideally with his hand, forward and backward to interact with a product, a region of interest (ROI) is cropped from the color frame and analysed. People detection and tracking is performed in the depth stream, while the final step of the interaction analysis use the color frame. The analysis thus involves 2 images per interaction: the first entering the “virtual wall” and the last exiting. A template matching method is used for this analysis.

Later on Liciotti et al. in [1] improved this method using the depth frame also for the interaction classification. After a background subtraction the classification was made using geometric features, calculating the difference in area between the ROIs (ideally the hand with or without a product). The authors also introduced a new class, the *neutral* one, to identify a touching interaction (product neither taken nor put in the shelf). Figure 3.4 depict these setup. In order to clarify this concept: if the ROI of the hand exiting the wall has a bigger area than the entering one the shopper has taken something from the shelf, thus the interaction is positive.

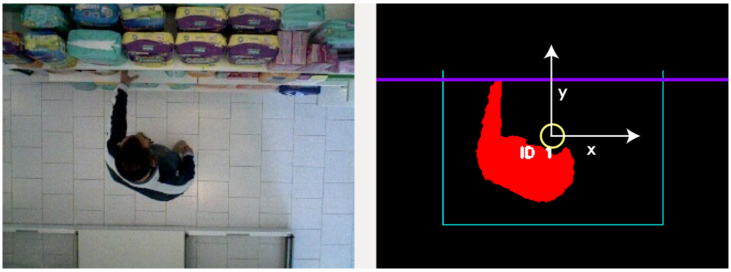


Figure 3.4: Camera setup for interaction analysis. Image from [1]

.

Further improvements of these approaches was necessary, because retail environment is challenging and classical computer vision feature-based method cannot cope with such complexity. The aforementioned methods cannot, for example, distinguish between real and unintentional interactions (i.e. shopper unintentionally cross the “virtual wall” with his body or even a shopping cart, Figure 3.5).

Another issue with these methods is that they cannot distinguish between real customers and store staff. In order to understand the shopper behavior, the interactions performed by the store employees, for example while refilling a shelf, must be filtered out.

Firstly a new interaction class is introduced: the “Refill”, which indicates an interaction performed by a staff employee instead of a customer. The depth information is also crucial in this approach, because allow to easily match a shopper interaction with a product, looking at the real world coordinates of the action. A step into deep learning approach was necessary and will be discussed in the next subsection.

3.2.1 VRAI-Net 2 for Shoppers Analytics

The key idea is to rely on the aforementioned methods and classify the interaction color images ROIs independently with a deep learning approach and

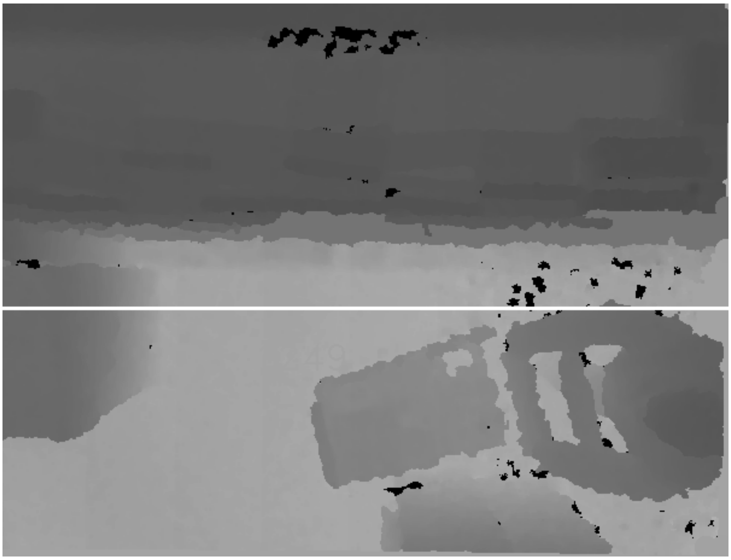


Figure 3.5: Depth frame with a shopper unintentionally interacting with the shelf.

combine the predictions into the interaction classification. Since there were no public datasets in the literature available with this scope, the new HaDa dataset has been collected in real stores with the aim to train a new neural network.

VRAI-Net 2 was designed to have an efficient architecture for classifying shopper interactions. This architecture can be adapted to classify either 3 (negative, neutral and positive) or 4 (negative, neutral, positive and refill) classes by simply modifying the last layer. The design of the network is based on the key idea of the inception module defined in [81], and also uses the improvements described in [82] and [83]. An attempt to scale up the network has been made, but at the same time, the number of parameters and the computational power required has been reduced. In a typical CNN layer, we chose either to have a stack of 3 × 3 filters, a stack of 5 × 5 filters or a max pooling layer. Generally, all of these are beneficial for the modelling power of the network. The inception module suggested the use of all of them. Then, the outputs of all these filters were concatenated and passed on as an input to the next layer.

This thesis follow the main idea of the typical architecture of a convolutional network in which going deep into the net also means making a subsample. A max pooling layer of 2 × 2 (stride 2) is used after every 2 inception modules. At the same time increases the feature map learned in each module. In this way, the output is halved but the number of feature channels is doubled.

The main architecture of the *VRAI-Net 2* is composed of 2 inception modules followed by a max pooling layer of 2 × 2 (stride 2) and 1 inception module.

Choosing the number of modules was designed to optimally process the images of the new dataset; starting from 80 × 80 pixels images, feature maps has been extracted with growing volumes step by step, up to dimensions (width and height) that were neither too small nor too large, until the classification layer has been reached. In this way, a number of learned parameters that were not too high is maintained (2.3 millions).

The last block of the network was used to map the features learned in the desired number of classes. Usually, only fully connected layers are used; however, these are very expensive in terms of learned parameters. Thus, a Global Average Pooling (GAP) layer is used after the last module, as in [84]. The GAP layer calculates the average of each feature map, and these values are fed directly into a softmax layer. This can remove the need for fully connected layers in a CNN-based classifier. It is considered to be a structural regulariser of CNNs, transforming feature activations into confidence maps by creating correspondences between features and classes. It also allows for a significant reduction of the parameters learned, compared to the parameters of the fully connected layers.

To speed up the learning and increase the stability of the neural network, a batch normalisation after each layer is added. This technique normalises the output of the previous activation layer by subtracting the batch's mean and dividing by the batch's standard deviation. The advantages are manifold. First, higher learning rates can be used because batch normalisation ensures that no activation can go extremely high or extremely low. Second, batch normalisation reduces overfitting because it has a slight regularising effect.

However, it is important not to depend solely on batch normalisation for regularisation; it should be used together with a dropout. Thus, a dropout layer (rate 50 %) is inserted before the classification layer. The *VRAI-Net 2* architecture is shown and described in Figure 3.6.

3.2.2 HaDa Dataset

The HaDa dataset comprises of 13856 manually labelled frames. These frames are of the same type as those used in the aforementioned features-based approach; thus, for each interaction, there are a total of 4 images (first RGB, first depth, last RGB and last depth). This dataset was collected in a real retail environment over a period of 3 months using 7 different cameras located in 4 different shelf categories (Chips, Women's Care, Baby Care and Spirits) above a total of 10 shelves.

Frames are labelled in the following 4 classes:

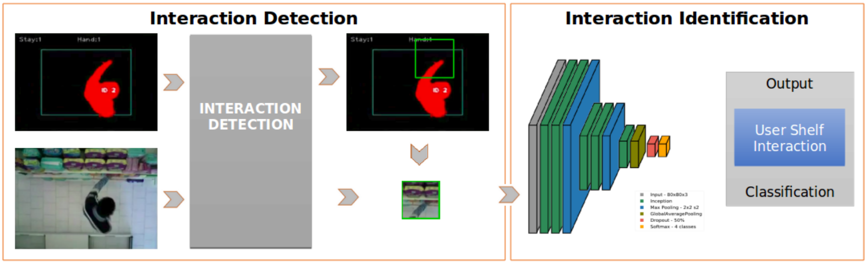


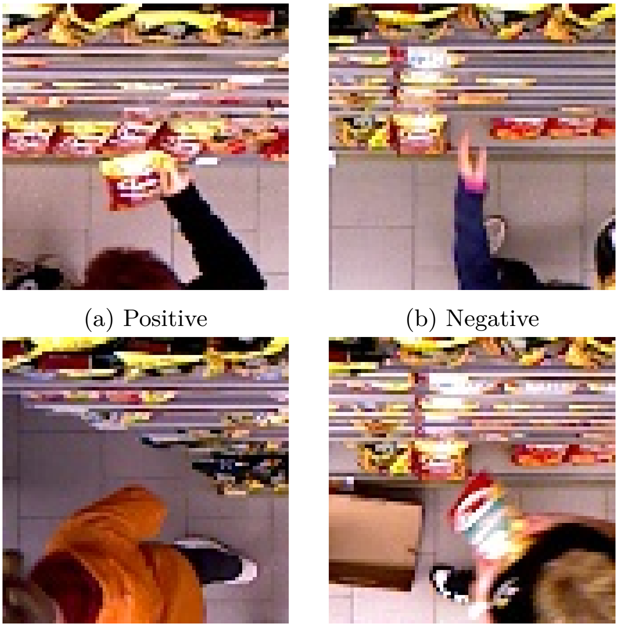
Figure 3.6: VRAI-Net 2 architecture. It is composed of 2 inception modules followed by a max pooling layer of 2×2 (stride 2) and 1 inception module. The last block of the network should be used to map the features learned in the desired number of classes. A GAP layer is used after the last module. These values are fed directly into a softmax layer. This can remove the need for fully connected layers in a CNN-based classifier.

- *Positive*: images that show a hand holding a product;
- *Negative*: images that show only a hand;
- *Neutral*: images in which the customer is not interacting with the shelf;
- *Refill*: images that indicate a refill action, which happens every time a box filled with products is visible. This class has a 'priority' over the others (for instance, if there is a hand holding a product and a box containing the same products in the same image, the class is deemed "refill" and not "positive").

Figure 3.7 depicts 4 samples of HaDa dataset classes.

3.2.3 Performance evaluation and Results

These dataset classes have not to be confused with the interaction classes previously introduced, because an interaction type is now given by the combination of 2 images classification. To evaluate the HaDa dataset, is necessary to independently classify the interaction frames (the first and last ones) of each interaction and combine these predictions to obtain the aforementioned interaction type. Four different networks were tested to determine the best results. These networks included a classic CNN in which the core structure was essentially the same as that of the LeNet architectures introduced in the late 1980s by LeCun et al. [85], AlexNet [86], CaffeNet [87], NASNet [88] and Xception [89]. Then, the CNN structure was modified, deepening it by duplicating the main block, which was composed of 2 convolution layers and a max pooling layer, CNN^2.



(a) Positive

(b) Negative

(c) Neutral

(d) Refill

Figure 3.7: HaDa Dataset.

Table 3.2 outlines the classification results for the interactions frame according to the classes defined in Section 3.2.1. From the same test set each type of interaction is extracted and compared with the features-based approach, using the test set labels as the ground truth. The test set was composed of 784 images; 624 of them were paired leading to an ultimate total of 312 interactions. By combining the labels given to the frames involved in each pair, the type of each interaction is determined. If at least one frame in a pair was labeled as “neutral”, the interaction can be excluded, as it was not a real interaction (fake interactions as described in Section 3.2.1). This led to the first important result: of a total of 312 interactions, 69 (22%) were fake and could at that point be excluded by the deep learning approach while had earlier been misinterpreted by the features-based approaches. After excluding the fake interactions and the interactions labelled as "refill" (in the features-based approach, the refill operation was misclassified in one of the other categories), an accuracy for the features-based approach of 70% is achieved. This value represented the accuracy on the real interactions performed by the customers; however, it represented only 16% of the total interactions. In terms of accuracy, *VRAI-Net 2*, which was the best CNN in this test, achieved 92% for the entire test set, and thus the same accuracy for the interaction type classification, outperforming the previous features-based results.

Looking at the interaction classification accuracy from a different point of view, positive interaction classifications were compared with sell out data over a period of 4 weeks on a real store in Italy. The assumptions behind are that a positive interaction (taking out a product from the shelf) is a final buy action for the shopper and that there are no other secondary placements for the analyzed category (i.e. diapers). A total of 1353 positive interactions in 4 week with opening time from 9 a.m. to 10 p.m. on the diapers category were measured and compared with the sellout provided by the store cashier system with a final accuracy of 96,72%. This final real test confirms again the high quality of the proposed approach on a real scenario. To better understand the applications of the proposed method and the high impact on the shopper marketing area we reported other aggregated results of the proposed framework in section 4.3.1.

Table 3.2: Shopper Interaction Results.

Approach	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
CNN	0.3775	0.9186	0.8395	0.8340	0.8367
CNN2	0.7773	0.8611	0.8620	0.8611	0.8616
AlexNet [86]	0.6115	0.7993	0.8164	0.7711	0.7928
CaffeNet [87]	0.7608	0.8731	0.8768	0.8720	0.8743
NASNet [88]	0.3316	0.9089	0.9124	0.9078	0.9300
Xception [89]	0.3362	0.9002	0.9066	0.8959	0.9011
VRAI-Net 2	**0.2251**	**0.9260**	**0.9347**	**0.9254**	**0.9300**

3.3 Hands Detection for Shopper Beaviour

A further step to improve the shopper-shelf interaction analysis described in section 3.2.1 can be made through a semantic segmentation of the scene. Detecting the shopper hands can minimize ambiguities and increase the accuracy of the action detection. The state-the-art in action detection in retail environments feature the *MERL Shopping Dataset*[4] that consists of 106 videos, each of which is a sequence about 2 minutes long. The videos are from a fixed overhead camera looking down at people shopping in a grocery store setting. Each video contains several instances of the following 5 actions: "Reach To Shelf" (reach hand into shelf), "Retract From Shelf " (retract hand from shelf), "Hand In Shelf" (extended period with hand in the shelf), "Inspect Product" (inspect product while holding it in hand), and "Inspect Shelf" (look at shelf while not touching or reaching for the shelf). This dataset however lacks of the depth in-

[4] http://www.merl.com/demos/merl-shopping-dataset

formation making impossible its use with the aforementioned approaches, but can be used to test semantic segmentation approaches, given the fact that is unique in the retail field.

3.3.1 RefineNet for hands semantic segmentation

In a previous section ResNet architecture has been mentioned and used to benchmark people counting segmentation approaches, but this kind of network suffers from downscaling of the feature maps which is not optimal for a good semantic segmentation. Atrous Convolution [90], which is another popular method in semantic segmentation used for example by DeepLab [91] architecture is computationally expensive to train and quickly reach memory limits even on modern GPUs. For the above reasons RefineNet has been introduced in [92].

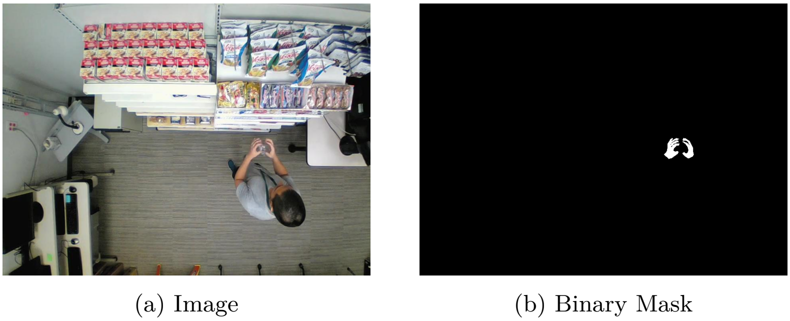
RefineNet is a generic multi-path refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. It uses a ResNet backbone and the deeper layers that capture high-level semantic features can be directly refined using fine-grained features from earlier convolutions. A chained residual pooling is also introduced which captures rich background context in an efficient manner. The key idea is to use the MERL Shopping Dataset on a pre-trained RefineNet in a transfer learning process to detect the shoppers hand to further apply the methods previously described to interaction classification.

3.3.2 MERL-HS

MERL Shopping Dataset is annotated for action detection without any segmentation information at a frame level. A new MERL-HS (MERL Hand Segmentation) dataset to test the hand segmentation approach has been generated from that. A total of 1000 frames from 10 different videos in the dataset have been manually semantically annotated using 2 classes: hand and background. Figure 3.8 shows a sample from the dataset with its binary mask.

3.3.3 Performance evaluation and Results

A RefineNet-Res101 model pre-trained on Pascal Person-Part [93] dataset was used for this experiment. The model was fine-tuned and adapted to account for 2 classes (hand and background) following the methodology used by Urooj et al. in [94], then was trained on the MERL-HS Dataset for 50 epochs. Results are depicted in figure 3.9 and reveal the feasibility of the approach with a training time limited to 12 hours on a single GPU model NVIDIA Tesla K80 using only 1000 images for training. The mean intersection over union (mIoU), reached


(a) Image

(b) Binary Mask

Figure 3.8: MERL-HS Dataset.

a peak of 0.87 in 50 epochs. A sample from the test set is depicted in 3.10 showing a good accuracy for the hand segmentation even for a person holding a shopping basket thus demonstrating the good generalization capability of the network and the feasibility of the approach, considering that Pascal Person-Part dataset is not specifically top-view.

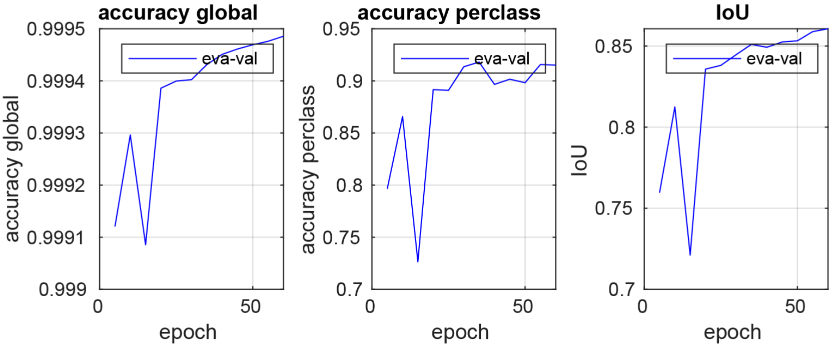


Figure 3.9: RefineNet results on MERL-HS.

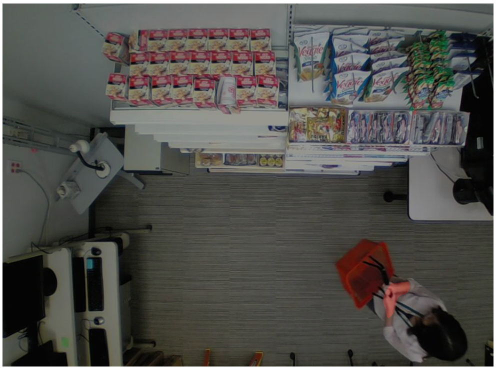


Figure 3.10: RefineNet qualitative result on MERL-HS. Hands highlighted in red.

3.4 Re-Identification

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.

In this context, 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.

Recently CNNs are being widely employed to solve the problem of person re-id. Deep Learning models in the person re-id problem are still suffering from the lack of training data samples. The reason for this is that most of the datasets provide only 2 images per individual [95]. Several CNN models have been proposed in the literature to improve the performance of person re-id. Specifically, 2 models have been employed in re-id area: a classification model and a Siamese model based on either pair or triplet comparisons.

The model based on classification requires determining the individual identity. In [96] a novel feature extraction model called Feature Fusion Net (FFN) is proposed for pedestrian image representation. The presented model makes use of both CNN feature and hand-crafted features. The authors utilise both color histogram features and texture features. The extracted features are followed by a buffer layer and a fully connected layer which are acting as the fusion layer.

The effectiveness was demonstrated on the 3 challenging datasets. In [97] a hybrid deep architecture for person re-id is presented, composed of Fisher vectors and multiple supervised layers. The network has been trained employing the linear discriminative analysis (LDA) as an objective function, with the goal of maximizing margin between classes. The authors in [98] propose a method based on learning deep feature representations from multiple domains by using CNNs with the aim to discover effective neurons for each training data set. The authors propose Domain Guided Dropout algorithm in order to improve the feature learning process by discarding useless neurons. They evaluate on various datasets, with the limitation that some neurons are effective only for a specific data set and useless for another one. The authors in [99] designed a multi-scale context aware network. The network is learning powerful features over the body and body parts. It can capture knowledge of the local context by stacking convolutions of multiple scales in each layer. They also propose to learn and locate deformable pedestrian parts through networks of spatial transformers with new spatial restrictions, instead of using predefined rigid parts. Since the person re-id research area lacks training instances, Siamese network models have been widely and viably employed. Siamese neural network is a type of neural network architectures which contains 2 or more identical sub-networks. A Siamese network is employed as pairwise (in the case of 2 sub-networks), or triplet (the case of 3 sub-networks). Some examples of pairwise research can be found in [100, 101, 102]. The authors in [100] combined 4 CNNs, each of them embedding images from different scale or different body part. Each of sub-CNN is trained with adaptive listwise loss function. In addition, they adopted sharpness parameter and an adaptive margin parameter to automatically focus more on the hard negative samples in the training process. In [101] a Siamese neural network has been proposed to learn pairwise similarity. The method can learn at the same time the color feature, texture feature and metric in a unified framework. The network is a symmetrical structure containing 2 sub-networks which are connected by Cosine function. Binomial deviance is also used to deal with the big variations of person images [102]. The authors proposes novel type of features based on covariance descriptors - the convolutional covariance features. There are 3 steps, firstly a hybrid network is trained for person recognition, next another hybrid network is employed to discriminate the gender, and finally the output of the 2 networks are passed through the coarse-to-fine transfer learning method to a pairwise Siamese network in order to accomplish the final person re-id. In [103], the authors presented a scalable distance driven feature learning framework based on the deep neural network in order to produce feature representation from a raw person images. A CNN network is trained by a set of triplets to produce features that can satisfy the relative distance constraints. In [104], a supervised learning framework is proposed to generate compact and bit-scalable hashing codes from raw images. Training images were organized into a batch of triplet samples, 2 images with the same label and one with a different label. The deep convolutional neural network is utilized to train the model in an end-to-end fashion, with the simultaneous optimization of the discriminative image features and hash functions. In [105], a 3-stage training is proposed: a deep convolutional neural network is first trained on an independent dataset labelled with attributes, then it is fine-tuned on another dataset that is only labelled with person IDs using a particular triplet loss they define, and finally, the updated network predicts attribute labels for the target dataset.

3.4.1 VRAI-Net3 for Top-View Re-identification

The RGB-D cameras installed in the stores were devoted not only to count and classify the interactions but also to re-identify the customers. Shopper re-identification allow a transition from a fully covered area to a sparse camera system and thus reduce the number of deployed sensor tracking customers only in areas of interest. The literature mostly cover frontal view approaches using color information from RGB cameras. Here the challenge is to use a top view configuration, which is the most acceptable way to deal with occlusion and intrusiveness concerns.

In a previous work [5], a Top-View Person Re-id (TVPR) dataset was built using videos of 100 persons recorded from a RGB-D camera in a top-view configuration. An Asus Xtion Pro Live RGB-D camera was chosen because this camera allows for the acquirement of color and depth information in an affordable and fast way. The camera was installed on the ceiling above the area to be analysed. The current work followed the same procedure adopted in [5] for re-identifying customers in the store. In particular, this methodology allowed to extract the important statistics of the shoppers, which included the time spent in the store, the products chosen by the same customer and the shelf attraction times. The approach recognises people from RGB-D images and consists of 2 steps: person detection and person identification. The detection of person is carried out by the Depth Channel and it uses an algorithm to locate people within frames, making a crop of the person through a 150 × 150 pixel bounding box, with a threshold on the minimum height of people. In this way, it is possible to remove the noise produced by the frame background and focus only on the interested details for every single image, i.e. the person. The 150 × 150 size is chosen experimentally, since we found that the people in our dataset had average dimensions between 80 × 80 and 125 × 125 pixels.

In the second step, a novel architecture called *VRAI-Net 3* was designed to carry out the identification of the people. This network is based on a type of classic DCNN architecture used for classification tasks, which in turn is based on the same concepts as those in VRAI-Net2. The network is adapted to process 150 × 150 pixel images by adding several inception layers followed by a max pooling layer. The network became deeper, increasing the number of features learned and thus improving the accuracy of the classification. In addition, the classification layer was adapted to classify 1000 classes. Figure 3.12 depicts the VRAI Network chosen for the re-id process.

The re-id phase allow to create an intermediate dataset that can be used to feed the CNN, to better perform the training. To increase the accuracy, the dataset has been balanced, maintaining a constant number of frames for each person, both for training and validation dataset. In particular, the balanced Dataset for 1000 people has the training set with 22 frame/person * 1000 people, hence 22.000 frames. The testing set has 22 frame/person * 1000 people, hence 22000 frames. The data augmentation (Figure 3.11) ensures 1.320.000 frames and it is done by using:

- image flipping, left-to-right and top-to-down;
- image rotation to 90^o, 180^o, 270^o;
- crops 3 × 3 (crop 130 × 130, stride 10 pixel, 3 steps horizontal x 3 steps vertical and resizing at 150 × 150 of the cropped).

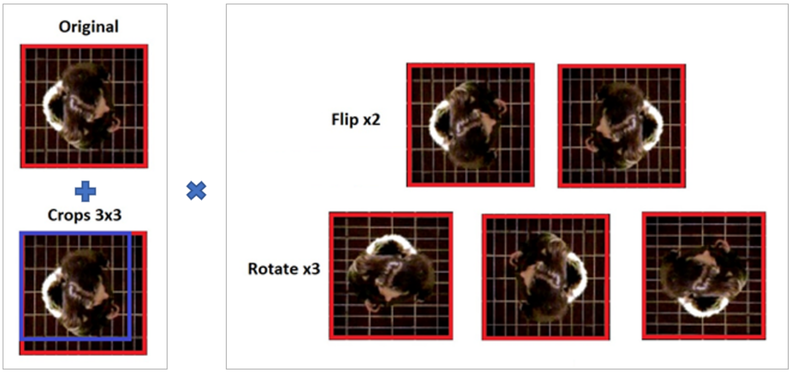


Figure 3.11: Data Augmentation.

3.4.2 TVPR2

The third dataset is TVPR2. This data was collected following the procedure outlined in [5], which described settings that are close to being realistic. This new dataset enabled possibilities in multiple directions, including deep learning, large-scale metric learning, multiple query techniques and search reranking

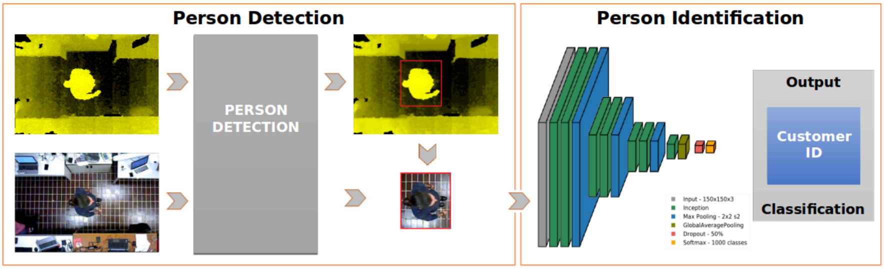
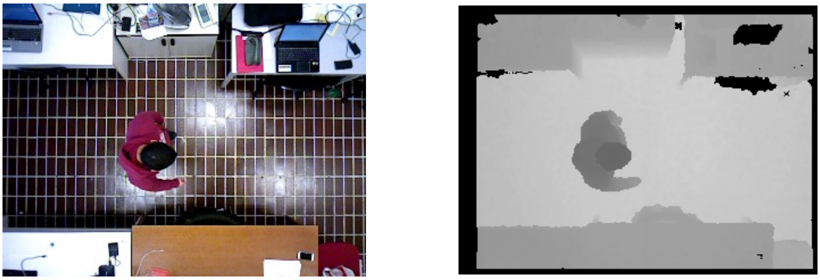


Figure 3.12: Person Re-Identification Workflow consists of 2 steps: Person Detection and Person Identification. The detection of person is carried out by the Depth Channel. For the Identification is designed the *VRAI-Net 3* architecture. The network was adapted to process 150 × 150 pixel images by adding several inception layers followed by a max pooling layer. In addition, the classification layer was adapted to classify 1000 classes based on the TVPR2 dataset.

directions. The dataset contains 235 videos, with RGB and depth channels. Each video recorded the people on the forward path (left to right) for half the time and recorded the same people on the return path (right to left) for the other half of the time, though not necessarily in that order. Resolution is 320 × 240. Recordings have been annotated reporting the people IDs in the order they appear in the video. The number of people present in the videos varied from one to eleven. The total number of unique people in this dataset was 1027.

Figure 3.13 shows an example of a dataset instance.



(a) Color stream.

(b) Depth Stream.

Figure 3.13: TVPR2 dataset. It consists of videos with color stream 3.3a and depth stream 3.13b of people passing under the camera in both direction, with annotations of people IDs.

Table 3.3 briefly summarises the characteristics of the data collected for the VRAI datasets.

Table 3.3: VRAI datasets.

Feature	**TVHeads**	**HaDa**	**TVPR2**
Total number of Images	3630	13856	223950
RGB Images	1815	6928	111975
Depth images	1815	6928	111975
Interactions	-	3464	-
Resolution	320 × 240	80 × 80	320 × 240
Annotation	Semantic Segmentation	4-Class Classification	1027-Class Classification

3.4.3 Performance evaluation and Results

In this subsection, re-id results of *VRAI-Net 3* on the TVPR2 dataset and comparison with those obtained from other state-of-the-art approaches are presented. The results that were obtained are shown in Table 3.4. The results reported in this Table are on a re-id over 1000 contemporary shoppers.

In the classification stage, different classifiers are compared according to the nature of the feature descriptors TVD (depth descriptor) and TVH (color descriptor) introduced in [3] . The overall prediction is performed by averaging the computed posterior probability of each classifier in order to provide the optimal decision rule. Based on TVD and TVH features, five state-of-the-art classifiers, namely k-nearest neighbours (kNN) [106], support vector machine (SVM) [26], decision tree (DT) [107], random forest (RF) [108] and Naïve Bayes (NB) [109] classifiers, are compared to recognize customers.

The network has been compared to another state-of-the-art classification network, the VGG-16 network [110]. To obtain shorter training times, a pre-trained VGG-16 [110] on the ImageNet dataset [86] has been used. Then, network was fine-tuned replacing the final classification layer with a custom layer and finally re-trained the network by using a lower learning rate in the first convolutional layers of the network and a more aggressive learning rate in the last layers.

Data augmentation was used both on the original images and on some of their clippings. Clippings were generated by moving a box of 130 × 130 pixels inside the image with strides of 10 pixels in both directions, making a 3 × 3 grid.

To improve accuracy during the testing phase, it was decided to use a technique called *10-crop validation.* For each image of the validation dataset, the network was tested on the original images of 4 of its crops (top-left, top-right, bottom-left and bottom-right), on the original image flipped left-to-right, and finally, on 4 more of its crops (top-left, top-right, bottom-left and bottom-right). As a result of this classification, the most commonly predicted class for these 10 types of tests is used.

Results reported in Table 3.4, show how *VRAI-Net 3* exceeded, in all metrics, the performance of the other features-based methods. In particular, increase of about 0.2 is obtained, for all the classification metrics compared to the SVM, which is the most common features-based approach. It is also interesting to note that the CMC of rank 1 of our *VRAI-Net 3* was lower than its own accuracy, which is very unique.

In addition to accuracy, precision, recall and f1-score, evaluation of approaches takes into account the CMC. The CMC represents the expectations of finding the correct identity in the first n-predicted identities. This metric is suitable for evaluating performances in recognition problems. Figure 3.14 shows the CMCs of the compared approaches. In particular, the horizontal axis indicates the rank, while the vertical axis indicates the probability of correctly identifying the corresponding rank. From the CMC, the curve of the proposed network *VRAI-Net 3* is always higher than the CMC curves of other state-of-the-art methods.

An additional comparison between the approaches was carried out to evaluate recognition performance according to the number of people identified, as depicted in Figure 3.15. From the graph is clear that the *VRAI-Net 3* is the less affected by accuracy loss when the number of people increase, which is a really important outcome in people re-identification.

Table 3.4: Re-Id Results on TVPR2, i.e. 1027 people in the retail space at the same time.

Approach	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	0.1754	0.2393	0.1578	0.1901
Decision Tree [111]	0.2262	0.2215	0.2104	0.2158
Random Forest [112]	0.3514	0.3552	0.3254	0.3396
K-NN [113]	0.4963	0.4775	0.4792	0.4783
SVM [114]	0.5587	0.5426	0.5458	0.5442
VGG-16 [110]	0.6754	0.7105	0.6592	0.6839
VRAI-Net 3	**0.7448**	**0.7794**	**0.7089**	**0.7425**

Table 3.5 shows the measured simulation runtime at different network sizes for the compared CNNs. The experiments are conducted on a single GPU model NVIDIA Tesla K80. The results reveal that *VRAI-Net 2* does not scale well. In contrast to this, *VRAI-Net 1* finishes the same task faster. The VRAI-Nets performances are aligned with the general purposes of the framework also in term of a correct mix of accuracy and time performances.

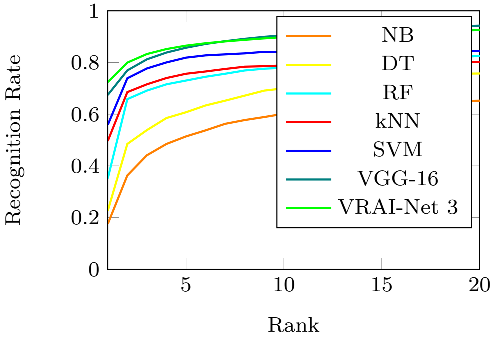


Figure 3.14: CMC on TVPR2 Dataset.

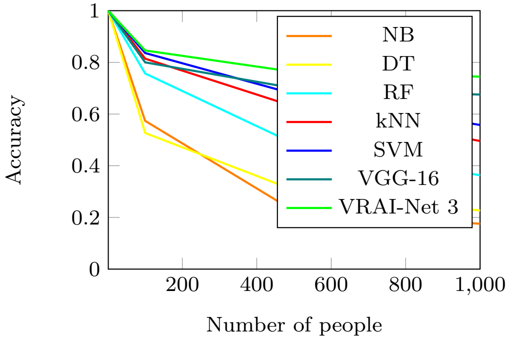


Figure 3.15: Scores of the people in the TVPR2 dataset.

Table 3.5: Computation Time for the testing stage.

	DCNNs	**Weights**	**Epoch Time**
People Counting	SEGNet	7.8M	15s
	RESNet	2.7M	17s
	FRACTAL	4.7M	20s
	UNet 1	7.8M	13s
	UNet 2	470k	12s
	UNet 3	2M	10s
	VRAI-Net 1	4.8 M	9s
Interactions Classification	CNN	2.7M	38s
	CNN2	10.7M	80s
	AlexNet	9.3M	43s
	CaffeNet	21.6M	80s
	NasNet	4.3M	618s
	Xception	20.8M	401s
	VRAI-Net 2	2.3M	1061s
Re-Identification	VGG-16	55M	1750s
	VRAI-Net 3	3M	3000s

Main applications of the re-id are: i) the evaluation of the dwell time inside the store, by the identification of the same person entering and exiting the store; ii) the identification of returning customers both at a store level and category level; iii) the store flow of a single person passing by different categories.

Chapter 4

Discussion: Limitations, challenges and lesson learnt

The systems illustrated and the use cases described so far, have proved that the Deep Learning approach, raised in the Introduction (Chapter 1), is suitable for the development of the challenging applications in intelligent retail environment, in which Deep Learning is the main core. Even if each application has different features and needs, it has been possible to outline a common path in every system. The different Deep Learning approaches experimented, have demonstrated that it is possible to cope with each need. The overriding goal was not only to present interesting retail solutions, but also to introduce challenging computer vision problems in the increasingly important domain chosen, accompanied with benchmark datasets and suitable performance evaluation methods. For a given problem, information can be obtained from multiple sources at different abstraction levels.

4.1 Thesis Contributions

The main contribution of this thesis can be summarized in the following aspects: the definition and development of an efficient and effective framework to tackle 3 important tasks in the shopper behavior understanding. The computer vision applications in which Deep Learning algorithms is the key core in their design, starting from general methods, that can be exploited in more fields, and then passing to methods and techniques addressing the specific problems. The applications are devoted to real-world retail problems. In fact, the issues taken into exam are People Counting, analysis of the interactions between shoppers and products and people re-identification. For each problem, a new deep learning approach is introduced together with a dedicated dataset, collected in a real store and made publicly available for the community. Although many approaches for people counting and re-identification exist in the literature, they mostly take advantage from the frontal view, which is rich in terms of features, but on the other hand are prone to occlusions among people. To make deep learning techniques tailored for the aforementioned challenging applications, considerations such as computational complexity reduction, hardware implementation, software optimization, and strategies for parallelizing solutions must be observed. The overarching goal of the work consisted of selecting the model that best explains the given observations; nevertheless, it does not prioritize in memory and time complexity when matching models to observations. Extensive efforts are devoted to collecting training and testing data and 4 newly challenging datasets are specifically designed for the described task. The design of benchmark involved several issues that range from the objective collection in order not to give any method unfair advantages to the consideration of the specificity of the concrete situation (the methods design that is tuned to a specific problem do not work properly on other problems).

4.2 Limitations

During the studies for this thesis, several limitations emerged that still exist and that are preventing the effectiveness of the applications. Limitations of the proposed approach reside on the nature of processed data, in fact depth maps, more than RGB images are strongly dependent on the sensor that produces them. Every depth sensor produces a different depth map estimating depth in different way often using a proprietary algorithm. Transfer learning between different datasets, acquired with different sensors, is desirable. Moreover different technologies exist in the depth sensor market and while structured light sensors produce accurate depth maps on the other hand they suffer from mutual interference and sunlight interference. Other depth technologies like stereo cameras for example are influenced by the natural light. Also the depth sensor market represents an incognito still far from the mass market adoption due to the their application and costs. These considerations imply that there is no perfect depth sensor and every one comes with its pros and cons to be evaluated and challenged.

The retail environment is complex to model and many different aspect must be taken into account. Interactions analysis performed with a limited number of cameras (typically one per shelf) allow the systems to be installed in a non-intrusive manner, but matching interactions with products based on the 3D location of them assume a perfect respect of the planogram. Planogram compliance on the other hand is still a open challenge and several companies in the retail intelligence market are developing solutions to account for it. Product recognition must be taken into account in the future but new challenges arises: to keep a low number of cameras and install them in the ceiling a great resolution will be needed to cope with such distance from the camera. On the other hand big resolution comes with more expensive computation. Computation resources are a key factor to keep the elaboration on the edge with low cost embedded hardware which is desirable in near real time systems also to deal with the modern regulations about privacy, making more complicated to record and transmit any personal image or video.

4.3 Challenges

The proposed applications, described in Chapter 3 open up a wealth of novel and important opportunities for the retail community. The datasets collected as well as the complex problems taken into exam, make the research challenging. Intensive attention has been drawn to the exploration of tailored learning models and algorithms, and their extension to more application areas. The evolution to the new deep learning algorithms has also demonstrated benefits. However, most existing methods directly borrow the models for multimedia tasks without considering the distinctiveness of multimedia data and multimedia tasks. As a result, these methods hardly fit the requirements of these multimedia tasks. The tailored methods, adopted for the development of the proposed applications, have shown to be capable of extracting complex statistical features and efficiently learning their representations, allowing it to generalize well across a wide variety of computer vision tasks, including image classification and so on. In order to cope with these new multimedia tasks, current models, including their architectures, training and inference methods, must be adapted or even re-designed. A number of fundamental issues had to be solved for emerging multimedia data, multimedia computing and applications. Challenges tackled in this thesis are the use of top-view configuration camera to understand the shoppers behavior. Reduce the number of sensors going from a fully covered area to a sparse camera system using re-identification, increase performances of people counting systems and interactions classification. These systems have to be not only reliable but also cost effective in order to maintain scalability, but efficient enough to run on the edge. Privacy concerns are rising and running analysis in the edge allow to avoid the recording of images or video and just transmit to the cloud synthetic and anonymous information. Winning these challenges paves the road towards high level applications in retail marketing research, answering complex business questions in a fast and efficient way. In the next section more details are provided in a marketing perspective.

4.3.1 Marketing applications of shopper behaviour understanding

As previously mentioned, the technologies discussed in this work have great relevance in the marketing field. In particular, they offer relevant contributions in the field of behavioural science and, more precisely, to consumer behaviour studies by using innovative methodologies and tools. Over the years, various attempts have been made to study in-store consumer behaviours using mainly manual recording techniques [115]. However, the extremely laborious nature of these techniques means that they take a long time to complete, and it is often difficult to obtain a large sample. Moreover, it is almost impossible to obtain a complete picture of a consumer's behaviour during his or her entire shopping journey at any given moment or from a series of moments over time. Another possible technique for measuring in-store behaviour is to interview consumers when they leave the store. However, a study on pedestrian flow in the city centre of Lincoln, Nebraska indicated that such investigative techniques lead to unacceptably high levels of inaccuracy [116]. Because of the obvious limitations of analyses made through manual surveys, over the years, researchers have begun to experiment with passive methods of data collection, which are considered most appropriate for in-store consumer behaviour studies. Implicit behaviour detections using technology have been carried out by many researchers, including Sorensen et al. [115], who used a shopping cart-tracking system, and Dieter Oosterlinck et al. [117], who used Bluetooth technology to detect in-store shopping journeys. The use of the technologies described in this thesis allowed, therefore, to analyse shoppers implicitly and continuously in all the stores in which they were observed, to obtain multiple independent, comparable studies. The main goals of this approach, which is defined in the literature as the "metanalysis approach" [118, 119], consist of:

- assessing Shopping Experience Fundamentals by comparing insights across different categories and in different store formats and

- confirming (or refuting) behavioural science theories using data obtained from actual shopper observations.

For example, by comparing data from multiple categories, it's straightforward to verify the most frequently used category management evidence, stating that, for instance, "the middle third of the shelf performs better than the top and bottom thirds." By using interaction recognition technology between people and products on a shelf, it's straightforward to promptly verify which parts of the shelf are touched more, confirming the theory mentioned in Figure 4.1. Through the same technology, we can also measure interaction dwell time to analyse the relationship between the time spent at the shelf and sales, which is positive up to a ceiling then becomes inverse (Figure 4.2). This mean that the longer amount of time a consumer spends at a shelf, the more purchases that will be made, but only up to a threshold of three products being touched. The use of these technologies, therefore, has a wide range of applications in the

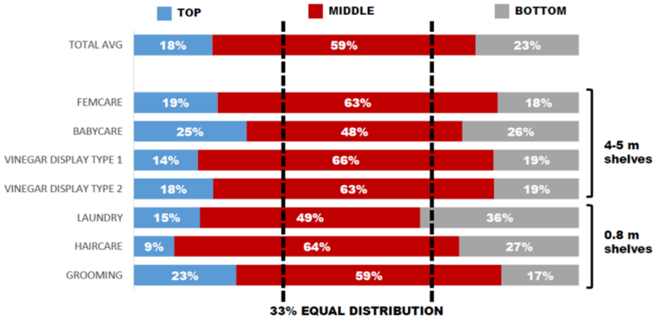


Figure 4.1: Distribution of positive interactions by top, middle and bottom shelf.

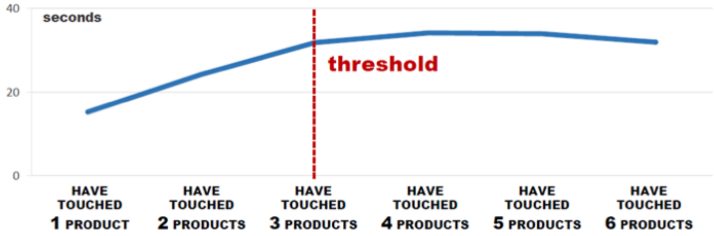


Figure 4.2: Time spent at the shelf by purchasers having had one to six interactions.

marketing field. Further studies should be conducted to deepen these technologies' potential and make useful findings in order to confirm or refute, through implicit observations, consumer behaviour theories.

4.4 Lesson Learnt

Understanding shoppers behavior is challenging in real retail environments and requires a combination of different technologies. On the other hand only a direct observation using depth and RGB sensor can provide useful information about shoppers letting them behave naturally. It is hence important to minimize the intrusiveness of these technology as well as the impact of the installation, to easy integrate them in existing retail environment.

However, the applications described in this thesis deserve some comments. Deep Learning techniques are delivering a promising solution to develop systems and to make innovation at a rapid pace. The combination of ICT technologies offers a framework for building large scale retail applications relying on data gathered from a complex infrastructure of sensors and smart devices. Numerous challenges exist in implementing such a framework, one of them is to meet the data and services requirements on informatics-based applications in terms of energy efficiency, sensing data quality, network resource consumption, and latency. Further, Deep Learning approaches had addressed various challenges such as anomaly detection, multivariate analysis, streaming and visualization of data.

As outlined in Chapter 2, recent literature has addressed the inherent power of Deep Learning for retail applications development. It can provide effective solutions for machine understanding of data (structured/semi structured), optimization problems, specifically, dealing with incomplete or inconsistent information. It is concerned with constructing systems that can improve the experiences in this work. The techniques proposed to meet the challenge of massive data processing, of which semi-supervised learning is a hot topic and should be one of the most important techniques. However, the cost of labelling the data is large because of expert experience or experiments, so only part of the data is labelled. Semi-supervised learning can utilize the unlabelled data. There are different methods to utilize the unlabelled data, of which clustering is a state-of-the-art method. But it does not work when it meets huge data.

Many improvements, from different perspectives, should be considered in the technology, so that the challenging nature of the requirements for the current and future computing environments can be accommodated.

Chapter 5

Conclusions and future works

In this thesis, a novel and powerful methodology and application for shopper behaviour analysis is presented. The system is based on RGB-D video in an intelligent retail environment and is evaluated on real environments, collecting 3 public datasets. Results prove that the proposed methodology is suitable for implicit shopper behaviour analysis with relevant applications in marketing and consumer research field with a particular focus on implicit consumer understanding.

The proposed research starts from the idea of collecting relevant datasets from real scenarios to change the overall methodology from a handcrafted feature based approach to a fully deep learning method with three concurrent CNNs processing the same frame to: i) segment and count people count with high accuracy (more than 99%) even in crowded environments); ii) measure and classify interactions between shoppers and shelves classifying positive, negative, neutral and refill actions with a good accuracy also compared with cashier sell-out; iii) perform a re-identification over contemporary shoppers (up to 1000 people in the same area at the same time) with a good accuracy to detect massive behavioural data on the best performing categories (more than 80% with 100 or 250 contemporary shoppers in the area).

For every task a public dataset is collected and shared together with the framework source codes to ensure comparisons with the proposed method and future improvements and collaborations over this challenging problems.

Research presented in this thesis has been conducted in fully cooperation and collaboration with Grottini Lab, a company operating in the retail intelligence market. The proposed systems are already operating in real retail environment in different countries around the world demonstrating the feasibility of the proposed approaches.

Future works should improve and better integrate the three CNNs with more complex architectures able to improve performances. Incremental learning methods will be investigated to improve the on-line performances of the re-identification algorithm. Further investigation on CNNs generalisations are needed to prove the effectiveness of the approach in very different retail categories (from grocery to fashion) and in cross-country human behaviours and attitudes. Hand segmentation for interaction classification can be integrated in the pipeline improving the overall accuracy of the system. Product recognition should be put in the loop. These improvement will have to take into account the considerations made in this thesis about low intrusiveness of installation to be integrated in already existing store and keeping elaboration on the edge to be compliant to the continuously increasing privacy concerns.

Bibliography

[1] 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 *Video Analytics for Audience Measurement*, C. Distante, S. Battiato, and A. Cavallaro, Eds. Cham: Springer International Publishing, 2014, pp. 146–157.

[2] M. Paolanti, D. Liciotti, R. Pietrini, A. Mancini, and E. Frontoni, “Modelling and forecasting customer navigation in intelligent retail environments,” *Journal of Intelligent & Robotic Systems*, vol. 91, no. 2, pp. 165–180, 2018.

[3] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, “People detection and tracking from an rgb-d camera in top-view configuration: Review of challenges and applications,” in *International Conference on Image Analysis and Processing.* Springer, 2017, pp. 207–218.

[4] D. Liciotti, M. Paolanti, R. Pietrini, E. Frontoni, and P. Zingaretti, “Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment,” in *Pattern Recognition (ICPR), 2018 24rd International Conference on.* IEEE, 2018.

[5] 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 *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, K. Nasrollahi, C. Distante, G. Hua, A. Cavallaro, T. B. Moeslund, S. Battiato, and Q. Ji, Eds. Cham: Springer International Publishing, 2017, pp. 1–11.

[6] M. J. Arnold and K. E. Reynolds, “Hedonic shopping motivations,” *Journal of retailing*, vol. 79, no. 2, pp. 77–95, 2003.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.

[8] M. Quintana, J. M. Menéndez, F. Alvarez, and J. Lopez, “Improving retail efficiency through sensing technologies: A survey,” *Pattern Recognition Letters*, vol. 81, pp. 3–10, 2016.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[10] S. R. Ahmed, “Applications of data mining in retail business,” in *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, vol. 2. IEEE, 2004, pp. 455–459.

[11] A. Berson and S. J. Smith, *Building data mining applications for CRM.* McGraw-Hill, Inc., 2002.

[12] C. Giraud-Carrier and O. Povel, “Characterising data mining software,” *Intelligent Data Analysis*, vol. 7, no. 3, pp. 181–192, 2003.

[13] S. Mitra, S. K. Pal, and P. Mitra, “Data mining in soft computing framework: a survey,” *IEEE transactions on neural networks*, vol. 13, no. 1, pp. 3–14, 2002.

[14] C. Wang and Y. Xi, “Convolutional neural network for image classification,” *Johns Hopkins University Baltimore, MD*, vol. 21218, 2015.

[15] D. P. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability.* John Wiley & Sons, Inc., 2001.

[16] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, “A review on generative adversarial networks: Algorithms, theory, and applications,” *arXiv preprint arXiv:2001.06937*, 2020.

[17] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2013.

[18] D. M. Chu and A. W. Smeulders, “Thirteen hard cases in visual tracking,” in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance.* IEEE, 2010, pp. 103–110.

[19] E. Vildjiounaite, S.-M. Mäkelä, S. Järvinen, T. Keränen, and V. Kyllönen, “Predicting consumers’ locations in dynamic environments via 3d sensor-based tracking,” in *2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies.* IEEE, 2014, pp. 100–105.

[20] 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*, vol. 53, pp. 16–23, 2015.

[21] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, Oct 2013.

[22] 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*, vol. 81, pp. 30 – 40, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786551600057X

[23] B. Dan, Y. Kim, Suryanto, J. Jung, and S. Ko, "Robust people counting system based on sensor fusion," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 1013–1021, August 2012.

[24] J. Han, E. J. Pauwels, P. M. de Zeeuw, and P. H. N. de With, "Employing a rgb-d sensor for real-time tracking of humans across multiple re-entries in a smart environment," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 255–263, May 2012.

[25] R. Ravnik, F. Solina, and V. Zabkar, "Modelling in-store consumer behaviour using machine learning and digital signage audience measurement data," in *International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage.* Springer, 2014, pp. 123–133.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[27] H.-B. Li, W. Wang, H.-W. Ding, and J. Dong, "Mining paths and transactions data to improve allocating commodity shelves in supermarket," in *Proceedings of 2012 IEEE International Conference on Service Operations and Logistics, and Informatics.* IEEE, 2012, pp. 102–106.

[28] 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.* Springer, 2014, pp. 146–157.

[29] R. Pierdicca, D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, and P. Zingaretti, "Low cost embedded system for increasing retail environment intelligence," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW).* IEEE, 2015, pp. 1–6.

[30] J. Melià-Seguí and R. Pous, "Human-object interaction reasoning using rfid-enabled smart shelf," in *2014 International Conference on the Internet of Things (IOT).* IEEE, 2014, pp. 37–42.

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

[32] M. George and C. Floerkemeier, "Recognizing products: A per-exemplar multi-label image classification approach," in *European Conference on Computer Vision.* Springer, 2014, pp. 440–455.

[33] G. Varol and R. S. Kuzu, "Toward retail product recognition on grocery shelves," in *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, vol. 9443. International Society for Optics and Photonics, 2015, p. 944309.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[35] P. Viola, M. Jones *et al.*, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)*, vol. 1, no. 511-518, p. 3, 2001.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[37] E. Cruz, S. Orts-Escolano, F. Gomez-Donoso, C. Rizo, J. C. Rangel, H. Mora, and M. Cazorla, "An augmented reality application for improving shopping experience in large retail stores," *Virtual Reality*, vol. 23, no. 3, pp. 281–291, 2019.

[38] V. J. Nogueira, H. Oliveira, J. A. Silva, T. Vieira, and K. Oliveira, "Retailnet: A deep learning approach for people counting and hot spots detection in retail stores," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images.* IEEE, 2019, pp. 1–8.

[39] Y. Song, Y. Xue, C. Li, X. Zhao, S. Liu, X. Zhuo, K. Zhang, B. Yan, X. Ning, Y. Wang *et al.*, "Online cost efficient customer recognition system for retail analytics," in *2017 IEEE Winter Applications of Computer Vision Workshops (WACVW).* IEEE, 2017, pp. 9–16.

[40] C. E. Kim, M. M. D. Oghaz, J. Fajtl, V. Argyriou, and P. Remagnino, "A comparison of embedded deep learning methods for person detection," *arXiv preprint arXiv:1812.03451*, 2018.

[41] X. Liu, Y. Jiang, P. Jain, and K.-H. Kim, "Tar: Enabling fine-grained targeted advertising in retail stores," in *Proceedings of the 16th Annual*

International Conference on Mobile Systems, Applications, and Services, 2018, pp. 323–336.

[42] N. T. Karim, S. Jain, J. Moonrinta, M. N. Dailey, and M. Ekpanyapong, "Customer and target individual face analysis for retail analytics," in *2018 International Workshop on Advanced Image Technology (IWAIT)*. IEEE, 2018, pp. 1–4.

[43] A. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in *2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 2018, pp. 1–6.

[44] F. Allegrino, P. Gabellini, L. Di Bello, M. Contigiani, and V. Placidi, "The vending shopper science lab: Deep learning for consumer research," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 307–317.

[45] L. Karlinsky, J. Shtok, Y. Tzur, and A. Tzadok, "Fine-grained recognition of thousands of object categories with single-example training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4113–4122.

[46] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. Yuille, "Scalenet: Guiding object proposal generation in supermarkets and beyond," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1791–1800.

[47] A. Tonioni and L. Di Stefano, "Domain invariant hierarchical embedding for grocery products recognition," *Computer Vision and Image Understanding*, vol. 182, pp. 81–92, 2019.

[48] A. Franco, D. Maltoni, and S. Papi, "Grocery product detection and recognition," *Expert Systems with Applications*, vol. 81, pp. 163–176, 2017.

[49] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[50] A. Tonioni, E. Serra, and L. Di Stefano, "A deep learning pipeline for product recognition on store shelves," in *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*. IEEE, 2018, pp. 25–31.

[51] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[52] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[53] T. Chong, I. Bustan, and M. Wee, “Deep learning approach to planogram compliance in retail stores,” *CS229 Stanford University*, 2016.

[54] M. Cotter, S. Advani, J. Sampson, K. Irick, and V. Narayanan, “A hardware accelerated multilevel visual classifier for embedded visual-assist systems,” in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2014, pp. 96–100.

[55] S. Advani, B. Smith, Y. Tanabe, K. Irick, M. Cotter, J. Sampson, and V. Narayanan, “Visual co-occurrence network: using context for large-scale object recognition in retail,” in *2015 13th IEEE Symposium on Embedded Systems for Real-time Multimedia (ESTIMedia)*. IEEE, 2015, pp. 1–10.

[56] L. Kong, X. Fan, and J. Lussier, “Item removal detection for retail environments with neural networks,” *CS231n Stanford University*, 2016.

[57] F. D. S. Ribeiro, F. Caliva, M. Swainson, K. Gudmundsson, G. Leontidis, and S. Kollias, “An adaptable deep learning system for optical character verification in retail food packaging,” in *2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. IEEE, 2018, pp. 1–8.

[58] A. L. Loureiro, V. L. Miguéis, and L. F. da Silva, “Exploring the use of deep neural networks for sales forecasting in fashion retail,” *Decision Support Systems*, vol. 114, pp. 81–93, 2018.

[59] F. Femling, A. Olsson, and F. Alonso-Fernandez, “Fruit and vegetable identification using machine learning for retail applications,” in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2018, pp. 9–15.

[60] G. Agnihotram, N. Vepakomma, S. Trivedi, S. Laha, N. Isaacs, S. Khatravath, P. Naik, and R. Kumar, “Combination of advanced robotics and computer vision for shelf analytics in a retail store,” in *2017 International Conference on Information Technology (ICIT)*. IEEE, 2017, pp. 119–124.

[61] M. Paolanti, M. Sturari, A. Mancini, P. Zingaretti, and E. Frontoni, "Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning," in *Mobile Robots (ECMR), 2017 European Conference on.* IEEE, 2017, pp. 1–6.

[62] B. Hu, Z. Nuoya, Z. Qiang, W. Xinggang, and L. Wenyu, "Diffnet: A learning to compare deep network for product recognition," *IEEE Access*, vol. 8, pp. 19 336–19 344, 2020.

[63] W. Ye, Y. Xu, and Z. Zhong, "Robust people counting in crowded environment," in *2007 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2007, pp. 1133–1137.

[64] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007.

[65] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 688–703.

[66] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[67] M. S. L. B. Dong Seon Cheng, Marco Cristani and V. Murino, "Custom pictorial structures for re-identification," in *Proceedings of the British Machine Vision Conference.* BMVA Press, 2011, pp. 68.1–68.11, http://dx.doi.org/10.5244/C.25.68.

[68] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *First International Workshop on Re-Identification*, October 2012.

[69] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 297–312.

[70] M. Paolanti, C. Kaiser, R. Schallner, E. Frontoni, and P. Zingaretti, "Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks," in *Image Analysis and Processing - ICIAP 2017*, S. Battiato, G. Gallo, R. Schettini, and F. Stanco, Eds. Cham: Springer International Publishing, 2017, pp. 402–413.

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

[72] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, “Efficient object localization using convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

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

[74] 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*, 2016, pp. 770–778.

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

[76] 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*, 2017, pp. 203–211.

[77] P. Jaccard, “Étude comparative de la distribution florale dans une portion des alpes et des jura,” *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[78] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[79] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, “A multi-stream bi-directional recurrent neural network for fine-grained action detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[80] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, “Customers’ activity recognition in intelligent retail environments,” in *International Conference on Image Analysis and Processing*. Springer, 2013, pp. 509–516.

[81] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,”

in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[82] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[83] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[84] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[85] 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*, 1990, pp. 396–404.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[87] 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.* ACM, 2014, pp. 675–678.

[88] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *arXiv preprint arXiv:1707.07012*, vol. 2, no. 6, 2017.

[89] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint*, pp. 1610–02 357, 2017.

[90] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[91] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[92] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[93] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1971–1978.

[94] A. Urooj and A. Borji, "Analysis of hand segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4710–4719.

[95] B. Lavi, M. F. Serj, and I. Ullah, "Survey on deep learning techniques for person re-identification task," *CoRR*, vol. abs/1807.05284, 2018.

[96] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An Enhanced Deep Feature Representation for Person Re-identification," *ArXiv e-prints*, Apr. 2016.

[97] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238 – 250, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316304447

[98] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification," *ArXiv e-prints*, Apr. 2016.

[99] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification," *ArXiv e-prints*, Oct. 2017.

[100] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 513–524, March 2017.

[101] D. Yi, Z. Lei, and S. Z. Li, "Deep Metric Learning for Practical Person Re-Identification," *ArXiv e-prints*, Jul. 2014.

[102] A. Franco and L. Oliveira, "Convolutional covariance features: Conception, integration and performance in person re-identification,"

Pattern Recognition, vol. 61, pp. 593 – 609, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316301625

[103] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993 – 3003, 2015, discriminative Feature Learning from Big Data for Visual Recognition. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320315001296

[104] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-Scalable Deep Hashing With Regularized Similarity Learning for Image Retrieval and Person Re-Identification," *IEEE Transactions on Image Processing*, vol. 24, pp. 4766–4779, Dec. 2015.

[105] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 475–491.

[106] T. H. Bø, B. Dysvik, and I. Jonassen, "Lsimpute: accurate estimation of missing values in microarray data with least squares methods," *Nucleic acids research*, vol. 32, no. 3, pp. e34–e34, 2004.

[107] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[108] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[109] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[110] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[111] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on IEEE*, 2008, pp. 1–6.

[112] Z. Li, Y.and Wu and R. Radke, "Multi-shot re-identification with random-projection-based random forests," in *Applications of Computer*

Vision (WACV), 2015 IEEE Winter Conference on. IEEE, 2015, pp. 373–380.

[113] B. S.D., “Nearest neighbor classification from multiple feature subsets,” in *Intelligent data analysis*, 1999, pp. 191–209.

[114] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q. Mary, “Person re-identification by support vector ranking,” in *BMVC*, vol. 2, 2010, p. 6.

[115] H. Sorensen, S. Bogomolova, K. Anderson, G. Trinh, A. Sharp, R. Kennedy, B. Page, and M. Wright, “Fundamental patterns of in-store shopper behavior,” *Journal of Retailing and Consumer Services*, vol. 37, pp. 182 – 194, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969698916303186

[116] H. Phillips and R. Bradshaw, “Camera tracking: a new tool for market research and retail management,” *Management Research News*, vol. 14, no. 4/5, pp. 20–22, 1991. [Online]. Available: https://doi.org/10.1108/eb028133

[117] D. Oosterlinck, D. F. Benoit, P. Baecke, and N. V. de Weghe, “Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits,” *Applied Geography*, vol. 78, pp. 55 – 65, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0143622816307330

[118] E. Roedel, “Fisher, r. a.: Statistical methods for research workers, 14. aufl., oliver & boyd, edinburgh, london 1970. xiii, 362 s., 12 abb., 74 tab., 40 s,” *Biometrische Zeitschrift*, vol. 13, no. 6, pp. 429–430, 1971. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.19710130623

[119] W. G. Kochran, “The combination of estimates from different experiments,” *Biometrics*, vol. 10, 03 1954.