Machine Vision and Applications (2020) 31:66
https://doi.org/10.1007/s00138-020-01118-w

ORIGINAL PAPER

Deep understanding of shopper behaviours and interactions using RGB-D vision

Marina Paolanti[1] · Rocco Pietrini[1] · Adriano Mancini[1] · Emanuele Frontoni[1] · Primo Zingaretti[1]

Received: 14 April 2019 / Revised: 8 August 2020 / Accepted: 24 August 2020 / Published online: 13 September 2020
© The Author(s) 2020

Abstract
In retail environments, understanding how shoppers move about in a 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 analytic and re-identification were developed with the aim of learning shopper skills 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 human behaviour. In this paper, a novel VRAI deep learning application that uses three convolutional neural networks 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 has been introduced. The framework is evaluated on the following three new datasets that are publicly available: TVHeads for people counting, HaDa for shopper–shelf interactions and TVPR2 for people re-identification. The experimental results show that the proposed methods significantly outperform all competitive state-of-the-art methods (accuracy of 99.5% on people counting, 92.6% on interaction classification and 74.5% on re-id), bringing to different and significative insights for implicit and extensive shopper behaviour analysis for marketing applications.

Keywords Intelligent retail environment · RGB-D camera · Deep learning · Human behaviour analysis

1 Introduction

In retail environments, understanding consumer behaviour is of great importance and one of the keys to success for retailers [1]. Many efforts have been devoted in particular towards monitoring how shoppers move about in 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 [2]. Popular sensors that are used for this task are RGB-D cameras because of their affordability, reliability and availability. 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 high-resolution depth and visual information has led to new challenges and opportunities for activity

✉ Marina Paolanti
m.paolanti@univpm.it

Rocco Pietrini
r.pietrini@pm.univpm.it

Adriano Mancini
a.mancini@univpm.it

Emanuele Frontoni
e.frontoni@univpm.it

Primo Zingaretti
p.zingaretti@univpm.it

[1] Department of Information Engineering (DII), Università Politecnica delle Marche, Via Brecce Bianche, Ancona, Italy

recognition and people tracking for many retail applications based on human–environment interactions.

In several research manuscripts, the top-view configuration was adopted to counter these challenges because it facilitates tasks and makes it more simple to extract different trajectory features. This setup choice also increases the robustness, because it reduces occlusions among individuals, it has the advantage of preserving privacy, because faces are not recorded by the camera, and it is more easy to set up in a retail environment. Furthermore, 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 (with up to 99% accuracy) application type, especially in very crowded scenarios (more than three 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 determine statistics and deduce insights [4–6]. 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 who go into the store and interact with the products, those who buy a product and those who interact with a product without buying it). People's re-identification (re-id) in different categories is also crucial to understanding the shopping trip 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 [7].

However, after moving into the era of multimedia big data, machine-learning approaches have 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 [8].

In this work, a novel VRAI[1] deep learning framework is introduced with the goal of improving our existing applications [4–6], and it is suggested that this evolution of machine intelligence 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. Figure 1 shows a camera installation in one of the stores where our experiments were performed.

This store has an area of about 1500 m^2, and 24 RGB-D (Asus Xtion Pro Live) cameras were installed in it in a top-view configuration without overlapping one another.

For maximum coverage using this relatively small number of cameras, the store was covered with two RGB-D cameras 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 in front of shelves, counting the number of people and re-identifying them in every top-seller category.

This test installation, together with that of the other four stores in Italy, China, Indonesia and the USA, became the basis for the dataset and results presented in this paper based on a two-year experience that measured 10.4 million shoppers and about 3 million interactions.

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.

To this end, newly challenging datasets were specifically designed for the tasks described in this study. In fact, each described application involved the construction 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.

Based on these evaluation configurations and datasets, a novel VRAI deep learning framework that uses three 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 is introduced. The system is able to process data at 10 frames per second, ensuring high performances even in cases of very brief shopper–shelf interactions.

The proposed methods were evaluated using three new publicly available datasets: TVHeads, HaDa and TVPR2, which are described more thoroughly in the next sections. 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.

The paper is organised as follows. Section 2 provides a description of the approaches that were adopted using

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

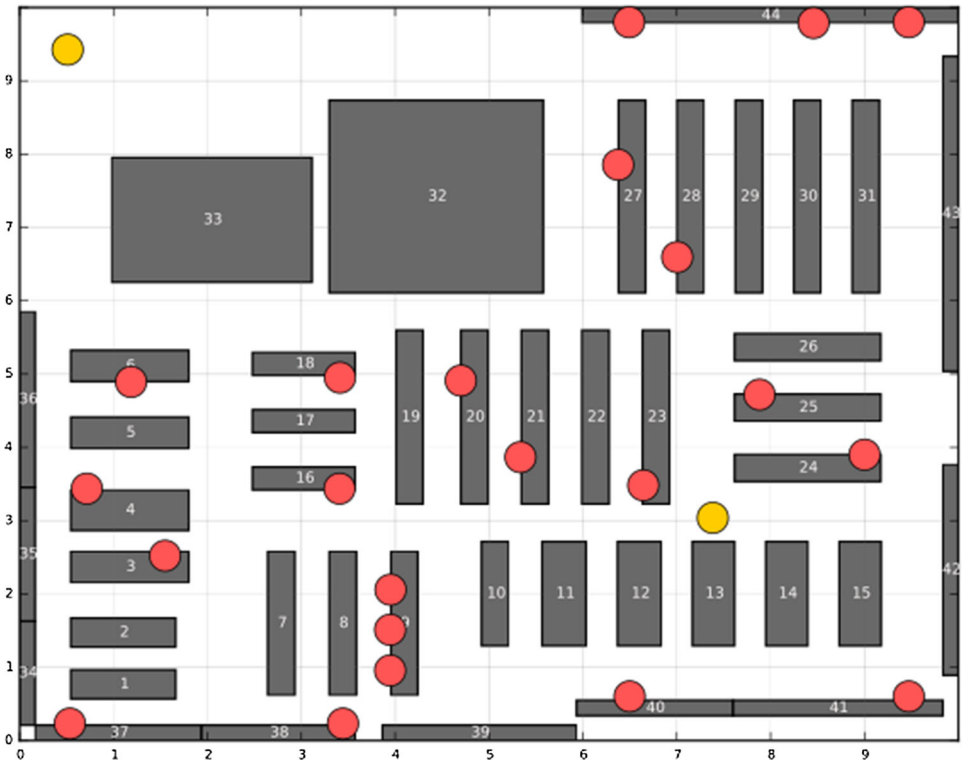


Fig. 1 Camera installations in the target store where our experiments are performed. This store has an area of about 1500 m^2 and was covered with a total of 24 RGB-D cameras that were installed in a top-view configuration. In particular, two RGB-D cameras were used for counting and identifying shoppers at the store's entrances (marked in yellow), and the other 22 cameras, in order to measure shoppers' attractions, attentions and interactions, were installed in front of shelves, counting the people and re-identifying them in every top-seller category (marked in red)

RGB-D sensors installed in a top-view configuration. Section 3 describes our approach to evolving our systems towards VRAI deep learning and offers details on "VRAI datasets", three new, challenging datasets that are publicly available. In Sect. 4, an extensive comparative evaluation of our approach with respect to the state-of-the-art "VRAI datasets" is offered, as well as a detailed analysis of each component of our approach. Finally, in Sect. 5, conclusions and discussion about future directions for this field of research are drawn.

2 Related work

In this section, the relevant literature concerning the human behaviour analysis in crowded environment is reviewed. Then, it focuses on the applications, discussing different existing approaches.

2.1 People detection and tracking in crowded environment

Detecting and tracking people in video sequences has recently attracted a great deal of attention in research, with wide application particularly in security surveillance and human–computer interaction fields [2]. Considerable research has been conducted on robust methods for tracking isolated and small numbers of humans, for which only transient occlusion exists. Still, however, the challenging task remains studying and tracking people in crowded situations that exhibit persistent occlusions or changes of appearance, complex and dynamic backgrounds, since conventional surveillance technologies often have difficulties understanding images [9].

Several early works have studied the deployment of conventional video cameras that lack depth information [10]. Among these methods, the most popular are those that extracted spatially global features [11] and those that used statistical learning with local features and boosting, such as edgelet [12] and histograms of oriented gradient [13]. Even though some works reported that these methods can lead to satisfying detection and tracking results, their performance does deteriorate in more challenging applications, which is due to the limitations in the use of conventional cameras in complex background situations.

To cope with crowded environments, several approaches have proposed the use of depth data produced by stereo rigs [14]. In a stationary camera setting, moving pixels can usually be extracted by changing the detection techniques, as outlined in [15]. Blobs are defined as groups of moving pixels according to their connectivity. On the other hand, although these methods have advantages over those that use conventional cameras, in realistic scenarios, a blob-based analysis may face several difficulties. For instance, a single blob may contain multiple humans, a single object may be fragmented into several blobs when a low colour contrast occurs, and

the last blobs may contain pixels corresponding to shadows or reflections caused by the moving objects. Several approaches have been proposed to help solve these problems. For example, [16] studied vertical blob projection in order to segment a large blob among multiple humans, and [17] analysed foreground boundaries in order to detect head candidates. However, in highly crowded environments, these methods may not be successful.

More recently, the quality of depth maps has greatly improved through the use of depth cameras such as Kinect [18], Xtion and TOF (time-of-flight) cameras, 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 [19]. Usually, depth cameras are placed either vertically overhead [3,20] or horizontally at the same level as humans [21]. Recent works have shown that the use of RGB-D cameras and depth maps can provide valuable additional information that significantly improves tracking and detection results [2,3].

Understanding a scene, which is one of the main problems in computer vision and semantic segmentation, is considered a high-level task that should lean into its understanding. In the past, these problems were addressed by various traditional computer vision and machine-learning techniques. More recently, deep learning architectures and CNNs have been proposed and are becoming more popular due to the improved accuracy and efficiency compared to former ones [5,22–25]. Semantic segmentation refers to the understanding of an image at the pixel level. In order to do this, each pixel of the image has to be assigned to an object class. In the literature, the widely known standards that have made significant contributions in the deep networks field are AlexNet [26], VGG-16 [27], GoogleNet [28] and ResNet [29]. Today, these are exploited as building blocks for many segmentation architectures.

The current milestone in deep learning techniques for semantic segmentation appears to be the fully convolutional network (FCN), as described by Long et al. [30]. Contemporary classification networks are adapted into FCNs, and they transfer their learned representations by fine-tuning them according to the segmentation task. A significant improvement has been achieved in segmentation accuracy through this method compared to traditional methods on standard datasets, such as PASCAL VOC, while also preserving inference efficiencies [31]. Although the FCN model is known for its power and flexibility, it still lacks various features. For instance, its inherent spatial invariance does not take into account useful global context information, it exhibits no instance-awareness by default, its efficiency has not yet reached real-time execution at high resolutions, and it is not completely suitable for unstructured data. Novel, state-of-the-art solutions such as SegNet, DeepLab, ENet and U-Net have been proposed in the literature in an attempt to overcome these challenges [31].

2.2 People counting

People counting can be classified into two categories: counting the number of people in a certain area and counting the number of people passing through a passage [32]. This work deals with the second category, i.e. counting people passing through a door or passage. In this regard, there are mainly vision-based studies, using radio frequency (RF) signals, such as Wi-Fi, ZigBee and UWB [33]. The main issues related to people counting in a given area have focused on the characteristics and the detection of the person, and they have tried to define indexes, which tend to change with the number of people and measuring the number of people based on these indexes. Instead, when the goal is counting people passing under a door or a gate, the binary direction along with the detection should also taken into exam [34].

Recent works have demonstrate the validity of CNN for density map estimation in single image for crowd counting. The first researchers that adopted CNN-based methods to crowd density estimation are Wang et al. [35] and Fu et al. [36]. At the same time, in [37], the authors propose a cross-scene crowd counting. The key idea is to map images into crowd counts and adapt this mapping to new target scenes for cross-scene counting. However, the drawback of this method is the need for perspective maps both on training scenes and test scenes.

Del Pizzo et al. [38] have described a vision-based approach for counting the number of persons which cross a virtual line. In their work, they have analysed the video stream acquired by a camera mounted in a zenithal position with respect to the counting line, allowing to determine the number of persons that cross the virtual line and providing the crossing direction for each person.

In [39], an image representation method is proposed. It combines semantic attributes and spatial cues to increase the discriminative power of feature representation. Shang et al. have presented an end-to-end network composed of CNN model and LSTM decoder to predict the number of people [40]. A deep spatial regression model for counting the number of people present in a still image with arbitrary perspective and arbitrary resolution based on CNN and LSTM is described in [41].

Recently, different approaches focus on combining additional for people counting such as detection, attention, localisation and synthetic data [42]. In [43], the authors presented a large synthetic crowd counting dataset, as well as a spatial fully convolutional network to improve real-world performance with synthetic data. All these approaches have reached huge success in crowd counting. However, these single image crowd counting methods may lead to inconsistent

head counts for neighbouring frames in video crowd counting. Liciotti et al. have introduced a novel modified U-Net architecture for head segmentation, first modifying the U-Net to U-Net 3 and then making it even more robust and efficient [5].

2.3 Human–object interactions

This topic is essentially concerned with engineering feature extraction methods that can promptly detect and represent motion in the input sequence of frames. There have been many approaches introduced from early stages of space-time pyramids [44] till recent years with the ones as attempts to substitute deep neural network feature extractors. In particular, human–object interaction is a considerable problem in computer vision. However, the types of studied interactions are mostly around sports activities [45], cooking [46] or everyday activities [47].

In retail environment, the shoppers interaction is completely different. Firstly, all the objects (e.g. products) could be considered a single category of the object without hurting the recognition. Secondly, only the movement of them in the scene helps distinguish some activities (picking from shelf vs putting back). In [48], the authors have proposed a method to recognise actions involving objects. The authors use graph neural networks to fuse human pose and object pose data for action recognition from surveillance cameras. However, their approach is heavily reliant on the quality of input pose information. In [49], the authors have introduced a framework for integrating human pose and object motion to both temporally detect and classify the activities in a fine-grained manner. They have combined partial human pose and interaction with the objects in a multi-stream neural network architecture to guide the spatiotemporal attention mechanism for more efficient activity recognition. They have also proposed to use the generative adversarial network (GAN) to generate exact joint locations from noisy probability heat maps. Furthermore, they have integrated the second stream of object motion to our network as a prior knowledge that we quantitatively show improves the recognition results. The approach has been applied on MERL dataset, a dataset composed by six activities, namely: "reach the shelf", "retract from the shelf", "hand in the shelf", "inspect the product", "inspect the shelf" and the background (or no action) class. It is recorded using a roof-mounted camera to simulate the real shopping store environment, albeit not specific. On the contrary, our proposed approach is evaluated on dataset collected in a real retail environment.

2.4 Person re-identification

Common re-id approaches are generally based on frontal image datasets, but sensors installed in top-view configuration have been revealed as especially effective in crowded environments [5].

Regarding person re-id problems, on the other hand, many recent works have addressed this issue. These works have focused mainly on either developing new descriptors for a person's appearance or on the learning techniques for a person re-id [50]. The problem of appearance models for person recognition, re-acquisition and tracking together was first introduced in [51]. The authors propose the cumulative match curve (CMC) as the performance evaluation metric and introduce the viewpoint invariant pedestrian recognition (VIPeR) dataset for re-id. Regarding discriminating features, hue–saturation–value (HSV) and red–green–blue (RGB) histogram colours were used initially due to their robustness to variations in resolution and perspective [52]. Later on, anthropometric measures were proposed based on calculating the physical conformation values of people, such as their heights, which were estimated using RGB cameras [53]. Another example is the use of anthropometric measurements in combination with clothing descriptors with both extracted through RGB-D cameras [54]. An estimation of the body's pose to guide feature extraction was proposed in [55]. With a similar performance as the former example, another state-of-the-art approach proposed an appearance model that does not rely on body parts but is based on a descriptor called the Mean Riemannian Covariance Grid [56].

Liciotti et al. [6], [57] proposed a method to extract anthropometric features through image processing techniques, then training machine-learning algorithms for re-id tasks. Their tests were carried out on a dataset of 100 people acquired using a top-view RGB-D camera.

Haque et al. [58] developed an attention-based model that deduces human body shape and motion dynamics by using depth information. Their approach was a combination of convolutional and recurrent neural networks leveraging unique 4-D spatiotemporal signatures to identify small discriminative regions indicative of human beings. Their tests were assessed on a DPI-T dataset, which consisted of 12 persons appearing in 25 videos while wearing different sets of clothing and holding different objects.

In [59], the authors started with a two-flow convolutional neural network (CNN) (one for RGB and one for depth) and a final fusion layer. They improved on this approach with a multi-modal attention network [60], adding an attention module to extract local and discriminative features that were fused with globally extracted features. In another work, Lejbolle et al. [61] presented a SLATT network with two types of attention modules (one spatial and one layer-wise). The authors collected also the OPR dataset from a university canteen, which was composed of 64 persons captured twice (entering and leaving a room). However, these datasets are not publicly available. The approach used in the current work, top-view re-id, matches the one proposed by the authors in [6].

2.5 Comparisons and contributions

In this work, a novel VRAI deep learning approach into existing applications [4–6] is introduced to evolve the machine-learning and features-based methods into deep learning methods, which are a more powerful and efficient way of dealing with the massive amounts of data generated from modern approaches.

The main contributions of this paper with respect to the state-of-the-art approach are: i) solutions, for real retail environments with a great variability in data acquired, derived from a large experience over 10.4 million shoppers observed in two years in different types of stores and in different countries; ii) 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. Three datasets are also provided that are publicly available to the scientific community for testing and comparing different approaches.

Table 1 is a comparative overview with the key differences between the different previous approaches and the VRAI proposed method. The triple deep network guaranties a single identification for every user entering the area (even in case of multi-cameras), and an association between the user identification and its interactions directly at a frame level, ensuring also multi-user identification and a correct association of men–shelf interactions to every user in the scene. Finally, the common frame pre-processing flow and the triple deep network inferences on the same input frame allow parallelisation on multi-core or multi-CPU architectures, ensuring low computational time for the whole framework processing; the system is able to process data at 10 frames per second, ensuring high performances even in cases of very fast shopper–shelf interactions. The networks use the same data stream and a common pre-processing phase that produce a proper input for each network.

3 From a geometric and features-based approach to VRAI deep learning

In this section, this evolution process as well as the datasets used for the evaluation is described. The framework is depicted in Fig. 2 and comprises three main systems: people counting [5], interaction classification [4] and re-id [6]. Specially designed new VRAI-Nets are used: *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 towards deep learning. 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 everything happens in front of the shelf, many fake interactions could occur because of unintentional interactions or something new in a scene; thus, an object that appears present may not actually be present.

Further details pertaining to this are described in the following subsections.

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 Sect. 3.4.

3.1 VRAI-Net 1 for people counting

Semantic segmentation has proven to be a high-level task when dealing with 2D images, 2D videos and even 3D data [24]. It paves the way towards the complete understanding of scenes and is being tackled using deep learning architectures, primarily deep convolutional neural networks (DCNNs), because they perform more accurately and sometimes even more efficiently compared to machine-learning and features-based approaches [23]. 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, it has been designed a novel *VRAI-Net 1*.

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 two 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 [63], the number of filters of the first convolution block was halved in the current study. A simpler network is obtained, going from 7.8 million parameters to 2 million.

The *VRAI-Net 1* architecture is shown in Fig. 3.

VRAI-NET 1 is even more robust and efficient than the work proposed in [5]. In the current work, the expansive path of the network was modified after being replaced by a refinement procedure. This procedure is composed of four layers, and each layer combines two types of the features map. It is basically formed by two 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 two branches are merged through an equivalent layer thickness (ELT) layer, which determines the element-wise sum of the outputs. The output of each

Table 1 Comparative overview table with the key differences between the different previous approaches and the VRAI proposed method (PC stands for people counting and HOI stands for human–object interactions)

	PC	HOI	Re-id	Position	RGB	Depth
Wang et al. [35]	X			Frontal	X	
Fu et al. [36]	X			Frontal	X	
Zhang et al. [37]	X			Frontal	X	
Del Pizzo et al. [38]	X			Top-View	X	X
Sheng et al. [39]	X			Frontal	X	
Shang et al. [40]	X			Frontal	X	
Yao et al. [41]	X			Frontal	X	
Fang et al. [42]	X			Frontal	X	
Wang et al. [43]	X			Frontal	X	
Liciotti et al. [5]	X			Top-View		X
Laptev et al. [44]		X		Frontal	X	
Yao et al. [45]		X		Frontal	X	
Rohrbach et al. [46]		X		Frontal	X	
Soomro et al. [47]		X		Frontal	X	
Kim et al. [48]		X		Frontal	X	
Moghaddam et al. [49]		X		Top-View	X	
Liciotti et. al [62], [4]		X		Top-View	X	X
Lisanti et al. [50]			X	Frontal	X	
Gray et al. [51], [52]			X	Frontal	X	
Madden et al. [53]			X	Frontal	X	
Pala et al. [54]			X	Frontal	X	
Cheng et al. [55]			X	Frontal	X	
Bak et al. [56]			X	Frontal	X	
Haque et al. [58]			X	Frontal	X	X
Lejbolle et al. [59], [60], [61]			X	Top-View	X	X
Liciotti et al. [6], [57]			X	Top-View	X	X
VRAI-Nets	X	X	X	X	Top-view	X

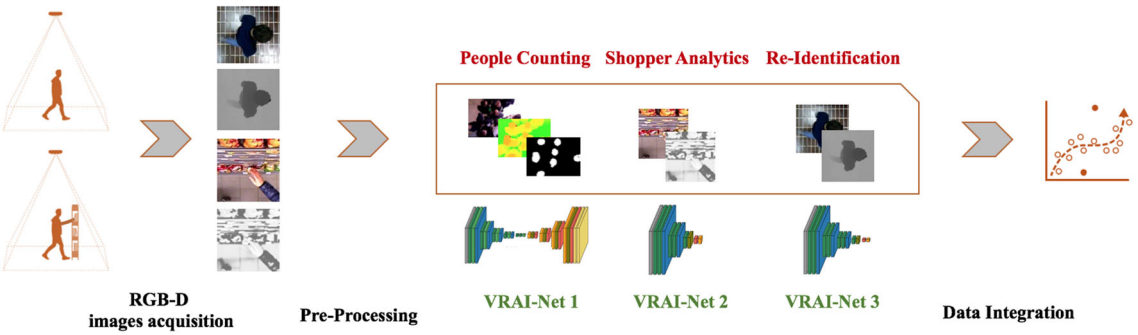


Fig. 2 VRAI framework for the deep understanding of shoppers' behaviour

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 [64] has been implemented, performing standard Bernoulli trials on the training phase and 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 towards increasing the robustness of our 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 four compared to the expansive part. Then, a 1×1 convolutional layer with a single channel was added both between the two sides and at the end of the expansive part. The number of every fea-

ture maps of the contraction part's layers was increased four times compared to the feature maps of the relative expansive part's layers. This was done because a good trade-off 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.

3.2 VRAI-Net 2 for interactions classification

In [4], 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 colour frame and analysed. People detection and tracking is performed in the depth stream, while the final step of the interaction analysis uses the colour frame. A depth-based hand detection method is used for this analysis: the system searches for every object that has an intersection with the "virtual wall" in the depth channel, and a crop of the RGB channels is performed around the intersection point with a resolution of 80x80 px. The final step of the interaction analysis, based on the proposed deep network, uses only the colour frame. The analysis thus involves 2 images per interaction: the first entering the "virtual wall" and the second exiting. Interactions can be occasionally performed also by shoulder or other parts of the body; these kind of interactions will be classified as a neutral interaction. The depth frame is also used 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). 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, and thus, the interaction is positive. 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). 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. In this paper, 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 it allows 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 the key idea is to rely on the aforementioned methods and classify the interaction colour images ROIs independently with a deep learning approach and 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.

Regarding this context, *VRAI-Net 2* was designed to have efficient architecture for classifying shopper interactions. This architecture can be adapted to classify either three (negative, neutral and positive) or four (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 [28] and also uses the improvements described in [65] and [66]. It has been attempted to scale up the network, but at the same time, the aim was to reduce the number of parameters and the computational power required. In a typical CNN layer, it has been chosen 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.

For this task, the main idea of the typical architecture of a convolutional network has been followed in which going deep into the net also means making a subsample. A max pooling layer of 2×2 (stride 2) has been used after every two inception modules. At the same time, the map of features learned in each module has to increase. In this way, the output has been halved but doubled the number of feature channels.

The main architecture of the *VRAI-Net 2* is composed of two 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 our dataset; starting from 80×80 pixels images, it has been extracted feature maps with growing volumes step by step, up to dimensions (width and height) that were neither too small nor too large, until the classification layer was reached. In this way, a number of learned parameters that were not too high are maintained.

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, it has been decided to use a global average pooling (GAP) layer after the last module, as in [67]. 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.

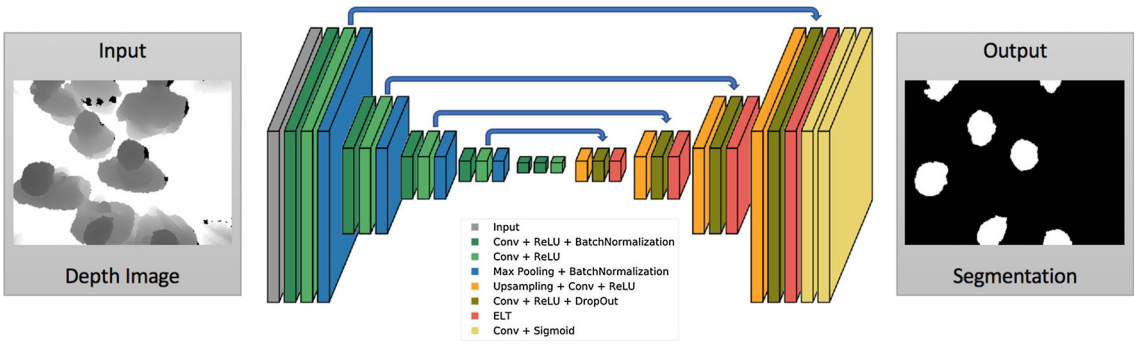


Fig. 3 VRAI-Net 1 architecture. It is composed of two main parts: a contracting path (left side) and an expansive path (right side). Its main contribution is its use of a refinement process in the expansive path; each step is a combination of two branches: one from the upsampling and the other from the 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 towards improving the robustness of the network in a shorter amount of time

To speed up the learning and increase the stability of the neural network, batch normalisation after each layer has been added [68]. 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 %) before the classification layer has been added. The *VRAI-Net 2* architecture is shown and described in Fig. 4.

VRAI-Net 1 and 2 share only the pre-processing layers, and to limit computational overload, VRAI-Net 2 is only activated when a person is detected and its hand interacts with the shelf.

3.3 VRAI-Net 3 for Top-View Person Re-Identification

As stated in the previous sections, the RGB-D cameras installed in the stores were devoted not only to counting and classifying the interactions but also to re-identifying the customers.

In a previous work [69], a Top-View Person Re-id (TVPR) dataset was built that contained videos of 100 persons recorded from an 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 colour 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 [6] for re-identifying customers in the store. In particular, with this methodology, important statistics of the shoppers have been extracted, 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 two 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 it has been found that the people in our dataset had average dimensions between 80×80 and 125×125 pixels (Fig. 5).

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-Net 2. The network has been 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 6 depicts the VRAI Network chosen for the re-id process.

The re-id phase allows to create an intermediate dataset that can be used to feed our DCNN, to better perform the training. To increase the accuracy, the dataset is 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, i.e. 22.000 frames. The testing set has 22 frame/person * 1000 people, i.e. 22000 frames. The data augmentation (Fig. 5) ensures 1.320.000 frames, and it is done by using:

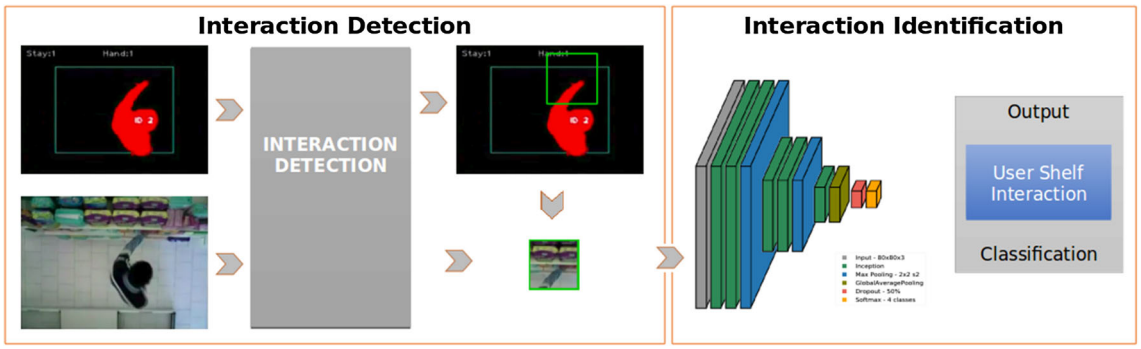


Fig. 4 VRAI-Net 2 architecture. It is composed of two inception modules followed by a max pooling layer of 2×2 (stride 2) and one inception module. The last block of the network should be used to map the features learned in the desired number of classes. A GAP layer is used after the last module. These values are fed directly into a softmax layer. This can remove the need for fully connected layers in a CNN-based classifier

- 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).

The proposed number is intended as a maximum number of contemporary people in the re-id gallery (closed world setting), and it is a credible number for a week-based re-id process in a store. Moreover, the system is able to be retrained at every new customer and, in a real installation done in Germany and reported in the last part of the paper, a specialised processing unit is able to train the network in a proper time frame, given also the very slow dynamic of a store visit. (Avg dwell time is 18 minutes in the test store.) Even if the deep network is trained again at every customer, the total amount of different people in the gallery is constant (1000) to ensure a high and stable re-id accuracy over time.

3.4 VRAI datasets

In this work, the first study on understanding shoppers' behaviours using an RGB-D camera installed in a top-view configuration is provided. As discussed in Sect. 1, 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 preserving privacy, since faces are not recorded by the camera.

Three new datasets are constructed from the images and videos acquired from the RGB-D cameras that were installed in a top-view configuration in different areas of the target store. The "VRAI datasets" were composed of the following three datasets:

- *TVHeads*[2] Dataset;
- HaDa[3] Dataset;
- TVPR2[4] Dataset.

The *TVHeads* dataset contained 1815 depth images (16 bit) with dimensions 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 three people per square metre and physical contact between them.

Following the pre-processing phase, a suitable scaling method was applied to the images, which allowed us to switch to 8 bits per pixel instead of the original 16. In this way, it is possible to obtain a more highlighted profile of the heads, improving the contrast and brightness of the image. The ground truth, for the head detection, was manually labelled by human operators.

Figure 7 shows an example of a dataset instance that includes the two images described above (8-bit depth image and the corresponding ground truth).

The HaDa dataset was composed of 13856 manually labelled frames. These frames were the same type as those used in the aforementioned features-based approach, thus, for each interaction, resulting in a total of four images (first RGB, first DEPTH, last RGB and last DEPTH). This dataset was acquired in a real retail environment over a period of three months using seven different cameras located in four different shelf categories (Chips, Women's Care, Baby Care and Spirits) above a total of ten shelves.

Frames were labelled in the following four frame interaction-classifying classes:

[2] http://vrai.dii.univpm.it/tvheads-dataset.

[3] http://vrai.dii.univpm.it/content/hada-dataset.

[4] http://vrai.dii.univpm.it/content/tvpr2-dataset.

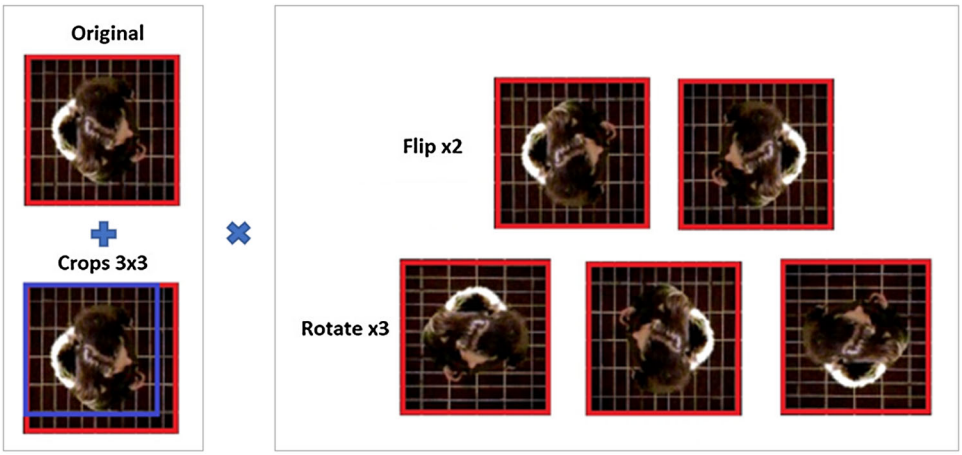


Fig. 5 Data augmentation

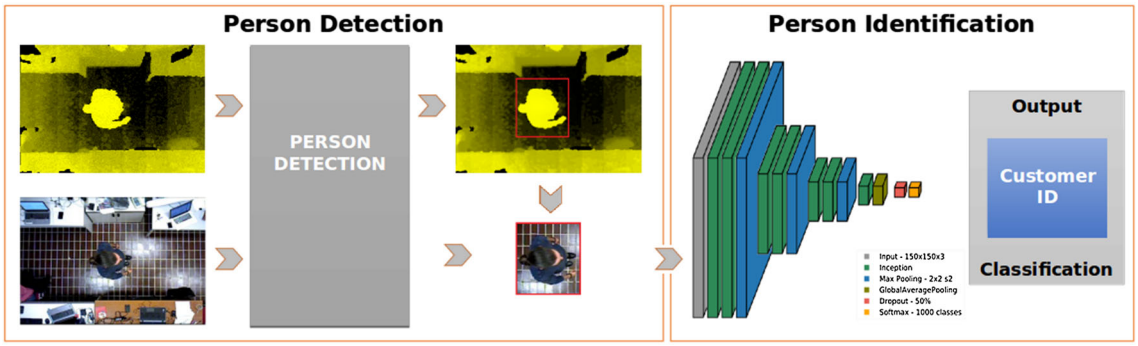
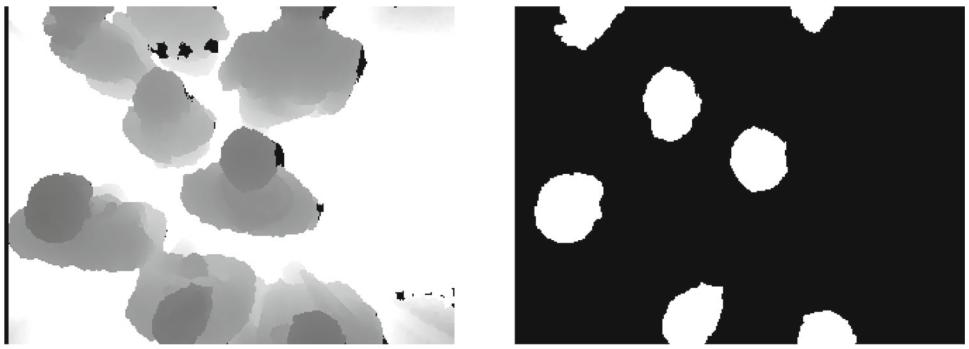


Fig. 6 Person Re-Identification Workflow consists of two steps: person detection and person identification. The detection of person is carried out by the depth channel. For the identification is designed the *VRAI-Net 3* architecture. The network was adapted to process 150 × 150 pixel images by adding several inception layers followed by a max pooling layer. In addition, the classification layer was adapted to classify 1000 classes based on our TVPR2 dataset



(a) 8-bit depth image. **(b)** Ground truth.

Fig. 7 TVHeads dataset. It consists of an 8-bit scaled depth image (**a**) and the corresponding ground truth (**b**)

– *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; and
– *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 8 depicts four samples of HaDa dataset classes.

The third dataset is TVPR2. This data was collected following the procedure outlined in [6], 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

(a) Positive **(b)** Negative **(c)** Neutral **(d)** Refill

Fig. 8 HaDa dataset

search re-ranking directions. The dataset was composed of 235 videos, containing RGB channels 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. The number of people present in the videos varied from one to eleven. The total number of people in this dataset was 1027.

Table 2 briefly summarises the characteristics of the data collected for the VRAI datasets.

4 Results and discussion

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 two years. This paper focused specifically on an analysis of a supermarket. Several days of video data were recorded from 24 cameras (2 used solely as counters and 22 used for counting, interaction classification and re-id) and processed by the system. Computation resources are a key factor to keep the elaboration on the edge with low-cost embedded hardware which is desirable in real applications in the retail scenario (i.e. high number of stores, high number of categories, etc.). In the following subsections, the results of our VRAI deep learning framework are evaluated and compared with the state-of-the-art framework.

4.1 People counting

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 [70], ResNet [29], FractalNet [71], U-Net, U-Net 2 [63,72] and U-Net 3 [5] to attempt to solve the problem of head image segmentations.

Each DCNN is trained using two types of depth images to highlight head silhouettes: 16-bit (original depth images) and 8-bit (scaled images). In this way, it is possible to improve each image's contrast and brightness. In the training phase, the dataset is split into training and validation sets with a ratio of 10%. The learning process is conducted for 200 epochs using a learning rate equal to 0.001 and an Adam optimisation algorithm. Semantic segmentation performances are shown in Table 3, which also reports the Jaccard [73] and Dice [74] indices for training and validation, respectively, as well the results in terms of accuracy, precision, recall and F1-score. Those metrics mainly concern the quality of the segmentation and not the counting of people. However, there are the Dice and Jaccard metrics, which are based on the area of predicted segmentation compared to the ground truth. Since their results are very good, it means that the segmentation is very accurate. We can count people from the mask of the predicted segmentation by simply using an image processing algorithm, which detects and counts the contours of the segmented areas. As it is possible to infer, the current study's *VRAI-Net 1* network outperformed the state-of-the-art network 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 our 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, the best performance used mainly images scaled to 8 bits.

The obtained deep learning results are compared with image processing algorithm: multi-level segmentation [75] and water filling [76]. Table 4 shows the results of algorithms in terms of precision, recall and F1-score. The algorithms reach lower values of performances. In particular, when the heads are along the edge of image, their accuracies are decreased. Instead, the multi-level segmentation algorithm looks more accurate than water filling algorithm.

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; and iii) the store flow modelling also for high traffic areas (e.g. promo areas). To

Table 2 VRAI datasets

Feature	TVHeads	HaDa	TVPR2
Total number of images	3630	13856	223,950
RGB Images	1815	6928	111,975
Depth images	1815	6928	111,975
Interactions	–	3464	–
Resolution	320 × 240	80 × 80	320 × 240
Annotation	Semantic segmentation	4-Class classification	1027-Class classification

Table 3 Jaccard and Dice indices comparison and segmentation results obtained for different DCNN architectures

CNN	Bit	Jaccard	Dice	Accuracy	Precision	Recall	F1-score
Fractal [71]	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 [70]	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 [29]	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 [63]	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 [72]	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 [5]	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

Table 4 Image processing algorithms performances

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

better understand the applications of the proposed method and the high impact on the shopper marketing area, other aggregated results of the proposed framework are reported in Sect. 4.4.

4.2 Interaction classification

To evaluate the HaDa dataset, the interaction frames are independently classified (the first and last ones) of each interaction and combined 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. [77], AlexNet [26] and CaffeNet [78]. Then, the CNN structure has been modified, deepening it by duplicating the main block, which was composed of two convolution layers and a max pooling layer, CNN^2.

Table 5 outlines the classification results for the interactions frame according to the classes defined in Sect. 3.2. From the same test set, each type of interaction has been extracted and compared it 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 has been 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. (These fake interactions are discussed in Sect. 3.2.) 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 they had earlier been misinterpreted by the features-based approaches. After excluding the fake interactions and the interactions labelled "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% has been obtained. 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 DCNN in our case, achieved 92% for the entire test set, and thus the same accuracy for the interaction-type classification, outperforming the previous features-based results. To prove 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 analysed 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 sell-out 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, other aggregated results of the proposed framework are reported in Sect. 4.4.

4.3 Re-identification

In this subsection, the re-id results of *VRAI-Net 3* on the TVPR2 dataset are presented and compared with those obtained from other state-of-the-art approaches. The results that were obtained are shown in Table 6. 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 (colour descriptor) . 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) [81], support vector machine (SVM) [82], decision tree (DT) [83], random forest (RF) [84] and Naïve Bayes (NB) [85] classifiers, are compared to recognise customers.

Our network has been compared to another state-of-the-art classification network, the VGG-16 network [27]. To obtain shorter training times, a pre-trained VGG-16 [27] on the ImageNet dataset [26] is used. Then, a network fine-tuning has been performed; the final classification layer has been replaced with our own custom layer and then 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.

For the training phase, it has been decided to also use data augmentation techniques. In particular, it has been used:

- 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).

These techniques were used both on the original images and on some of their clippings. The clippings were generated by moving a box of 130×130 pixels inside the image with steps 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 four of its crops (top-left, top-right, bottom-left and bottom-right), on the original image flipped left to right, and finally, on four 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 was used.

Moreover, the results obtained by training a DCNN with a triplet loss, pretrained on ImageNet dataset, are reported. Given an input image (called anchor), this function tends to bring it closer to images of its same class (hard positive) while simultaneously moving it away from images of the other classes (hard negative) [86]. GoogleNet [87] has been chosen as backbone, pretrained on ImageNet dataset [88]; the triplet loss function was instead based on the work of Hermans et al. [89].

From the results reported in Table 6, it can be observed how *VRAI-Net 3* exceeded, in all metrics, the performance of the other features-based methods. In particular, an increase of about 0.2 was 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 one of our *VRAI-Net 3* was lower than its own accuracy, which is very unique.

There are also several real checks in-store through human observations. Data have been collected in two target stores in Germany, and it has been observed for 2 hours people entering and leaving the stores. The accuracy of real in-store tests gave an accuracy of 73% comparable with the VRAI-Net 3 results on the laboratory dataset.

In addition to accuracy, precision, recall and F1-score, our approaches have been evaluated using 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 9 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, it is possible to infer that the curve of our proposed network *VRAI-Net 3* was always higher than the CMC curves of other state-of-the-art methods.

Table 5 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 [26]	0.6115	0.7993	0.8164	0.7711	0.7928
CaffeNet	0.7608	0.8731	0.8768	0.8720	0.8743
NASNet [79]	0.3316	0.9089	0.9124	0.9078	0.9300
Xception [80]	0.3362	0.9002	0.9066	0.8959	0.9011
VRAI-Net 2	0.2251	0.9260	0.9347	0.9254	0.9300

Table 6 Re-Id Results on TVPR2, i.e. 1027 contemporary people in the retail space

Approach	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.1754	0.2393	0.1578	0.1901
Decision Tree [90]	0.2262	0.2215	0.2104	0.2158
Random Forest [91]	0.3514	0.3552	0.3254	0.3396
K-NN [92]	0.4963	0.4775	0.4792	0.4783
SVM [93]	0.5587	0.5426	0.5458	0.5442
VGG-16 [27]	0.6754	0.7105	0.6592	0.6839
Triple Loss (TL) DCNN	0.6212	0.6380	0.6075	0.6227
VRAI-Net 3	0.7448	0.7794	0.7089	0.7425

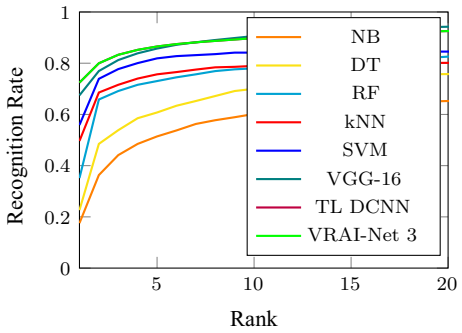


Fig. 9 CMC on TVPR2 dataset

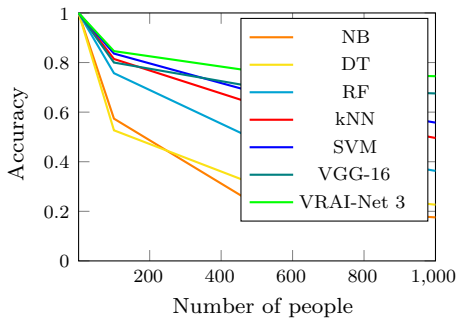


Fig. 10 Scores of the people in the TVPR2 dataset

An additional comparison between the approaches was carried out to evaluate recognition performance according to the number of people identified, as depicted in Fig. 10.

Table 7 shows the measured simulation runtime at different network sizes for the compared DCNNs. The experiments are conducted on GPU 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 terms of a correct mix of accuracy and time performances. These systems are not only reliable but also cost-effective in order to maintain scalability, but efficient enough to run on the edge. The designed network is fast and light in terms of parameters and learns with good performance.

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; and iii) the store flow of a single person passing by different categories. Next section will better clarify the impact of the proposed methods on the shopper marketing.

4.4 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

Table 7 Parameters and computational time comparison

	DCNNs	Number of parameters	Training-time for Epoch
People counting	SEGNet	7.8M	15s
	RESNet	2.7M	17s
	FRACTAL	4.7M	20s
	U-Net 1	7.8M	13s
	U-Net 2	470k	12s
	U-Net 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

[94]. 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 [95]. 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. [94], who used a shopping cart-tracking system, and Dieter Oosterlinck et al. [96], who used Bluetooth technology to detect in-store shopping journeys. The use of the technologies described in this paper allowed us, 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 "meta-analysis" [97,98], 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 is possible to verify that 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 is possible to promptly verify which parts of the shelf are touched more, confirming the theory mentioned in Fig. 11.

Through the same technology, it is also possible to measure interaction dwell time to analyse the relationship between the time spent at the shelf and sales, which is positive up to a ceiling and then becomes inverse (Fig. 12). This mean that the longer the amount of time a consumer spends at a shelf, the more the 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 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.

5 Conclusion

In this paper, 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.

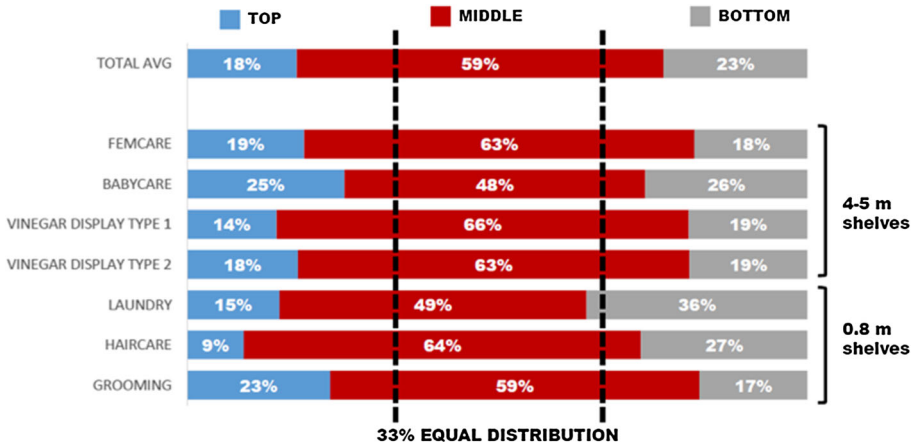


Fig. 11 Distribution of positive interactions by top, middle and bottom shelf

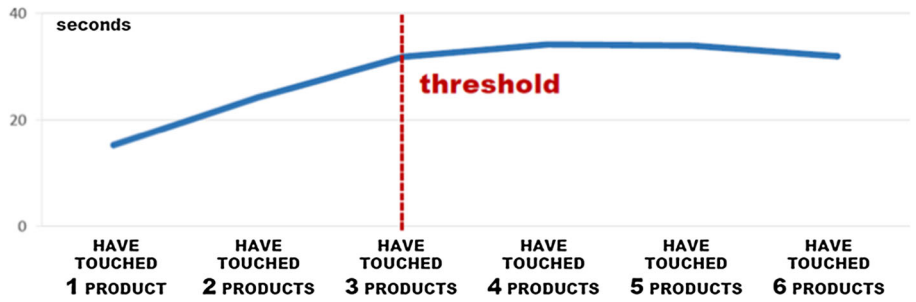


Fig. 12 Time spent at the shelf by purchasers having had one to six interactions

The proposed research starts from the idea of collecting relevant datasets from real scenarios to change the overall methodology from a feature-based and geometric 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; and 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 purpose, 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. The paper describes one of the more extensive tests based on real data from real retail scenarios in the literature. The system is also designed to deal with the modern regulations about privacy, avoiding to record and transmit any personal image or video, computing all frames on edge, and transmitting to the cloud only synthetic and anonymised data.

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 online performances of the re-identification algorithm. Further investigation on CNNs generalisations is 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.

Acknowledgements This work was funded by Grottini Lab (www.grottinilab.com). The authors would like to thank Andrea Felicetti, Massimo Martini, Lorenzo Nardi, Claudia Norscini and Raffaele Vaira for their support.

Funding Open access funding provided by Università Politecnica delle Marche within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Paolanti, M., Liciotti, D., Pietrini, R., Mancini, A., Frontoni, E.: Modelling and forecasting customer navigation in intelligent retail environments. J. Intell. Robot. Syst. **91**(2), 165–180 (2018)
2. Liu, J., Liu, Y., Zhang, G., Zhu, P., Chen, Y.Q.: Detecting and tracking people in real time with rgb-d camera. Pattern Recogni. Lett. **53**, 16–23 (2015)
3. Liciotti, D., Paolanti, M., Frontoni, E., Zingaretti, P.: 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, pp. 207–218. Springer (2017)
4. Liciotti, D., Contigiani, M., Frontoni, E., Mancini, A., Zingaretti, P., Placidi, V.: Shopper analytics: a customer activity recognition system using a distributed rgb-d camera network. In: Distante, C., Battiato, S., Cavallaro, A. (eds.) Video Analytics for Audience Measurement, pp. 146–157. Springer, Cham (2014)
5. Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., Zingaretti, P.: Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment. In: 2018 24rd International Conference on Pattern Recognition (ICPR). IEEE (2018)
6. Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., Zingaretti, P.: Person re-identification dataset with rgb-d camera in a top-view configuration. In: Nasrollahi, K., Distante, C., Hua, G., Cavallaro, A., Moeslund, T.B., Battiato, S., Ji, Q. (eds.) Video Analytics. Face and Facial Expression Recognition and Audience Measurement, pp. 1–11. Springer, Cham (2017)
7. Arnold, M.J., Reynolds, K.E.: Hedonic shopping motivations. J. Retail. **79**(2), 77–95 (2003)
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
9. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. IEEE Trans. Pattern Anal. Mach. Intell. **30**(7), 1198–1211 (2008)
10. Bogdan Rusu, R., Sundaresan, A., Morisset, B., Hauser, K., Agrawal, M., Latombe, J.C., Beetz, M.: Leaving flatland: efficient real-time three-dimensional perception and motion planning. J. Field Robot. **26**(10), 841–862 (2009). https://doi.org/10.1002/rob.20313
11. Felzenszwalb, P.F.: Learning models for object recognition. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I–I (2001). https://doi.org/10.1109/CVPR.2001.990647
12. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. Int. J. Comput. Vis. **75**(2), 247–266 (2007). https://doi.org/10.1007/s11263-006-0027-7
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893 vol. 1 (2005). https://doi.org/10.1109/CVPR.2005.177
14. Ess, A., Leibe, B., Schindler, K., van Gool, L.: Robust multiperson tracking from a mobile platform. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1831–1846 (2009). https://doi.org/10.1109/TPAMI.2009.109
15. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 780–785 (1997)
16. Haritaoglu, I., Harwood, D., Davis, L.S.: W/sup 4/: real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 809–830 (2000). https://doi.org/10.1109/34.868683
17. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1208–1221 (2004). https://doi.org/10.1109/TPAMI.2004.73
18. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans. Cybern. **43**(5), 1318–1334 (2013). https://doi.org/10.1109/TCYB.2013.2265378
19. Sturari, M., Liciotti, D., Pierdicca, R., Frontoni, E., Mancini, A., Contigiani, M., Zingaretti, P.: Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. Pattern Recognit. Lett. **81**, 30–40 (2016). https://doi.org/10.1016/j.patrec.2016.02.010
20. Dan, B., Kim, Y., Suryanto, Jung, J., Ko, S., : Robust people counting system based on sensor fusion. IEEE Trans. Consum. Electron. **58**(3), 1013–1021 (2012). https://doi.org/10.1109/TCE.2012.6311350
21. Han, J., Pauwels, E.J., de Zeeuw, P.M., de With, P.H.N.: Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. IEEE Trans. Consum. Electron. **58**(2), 255–263 (2012). https://doi.org/10.1109/TCE.2012.6227420
22. Hu, L., Hong, C., Zeng, Z., Wang, X.: Two-stream person re-identification with multi-task deep neural networks. Machine Vision and Applications pp. 1–8 (2018)
23. Paolanti, M., Kaiser, C., Schallner, R., Frontoni, E., Zingaretti, P.: Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. In: Battiato, S., Gallo, G., Schettini, R., Stanco, F. (eds.) Image Analysis and Processing—ICIAP 2017, pp. 402–413. Springer, Cham (2017)
24. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision—ECCV 2014, pp. 297–312. Springer, Cham (2014)
25. Paolanti, M., Sturari, M., Mancini, A., Zingaretti, P., Frontoni, E.: Mobile robot for retail surveying and inventory using visual and textual analysis of monocular pictures based on deep learning. In: 2017 European Conference on Mobile Robots (ECMR), pp. 1–6. IEEE (2017)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, pp. 1097–1105. Curran Associates Inc., USA (2012). http://dl.acm.org/citation.cfm?id=2999134.2999257
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014). arXiv:1409.1556
28. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
30. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017). https://doi.org/10.1109/TPAMI.2016.2572683
31. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. Int. J. Multimed. Inf. Retr. **7**(2), 87–93 (2018). https://doi.org/10.1007/s13735-017-0141-z
32. Choi, J.W., Quan, X., Cho, S.H.: Bi-directional passing people counting system based on ir-uwb radar sensors. IEEE Internet Things J. **5**(2), 512–522 (2017)
33. Mrazovac, B., Bjelica, M.Z., Kukolj, D., Todorovic, B.M., Samardzija, D.: A human detection method for residential smart energy systems based on zigbee rssi changes. IEEE Trans. Consum. Electron. **58**(3), 819–824 (2012)

34. García, J., Gardel, A., Bravo, I., Lázaro, J.L., Martínez, M., Rodríguez, D.: Directional people counter based on head tracking. IEEE Trans. Ind. Electron. **60**(9), 3991–4000 (2012)
35. Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X.: Deep people counting in extremely dense crowds. In: Proceedings of the 23rd ACM international conference on Multimedia, pp. 1299–1302 (2015)
36. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C.: Fast crowd density estimation with convolutional neural networks. Eng. Appl. Artif. Intell. **43**, 81–88 (2015)
37. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841 (2015)
38. Del Pizzo, L., Foggia, P., Greco, A., Percannella, G., Vento, M.: Counting people by RGB or depth overhead cameras. Pattern Recognit. Lett. **81**, 41–50 (2016)
39. Sheng, B., Shen, C., Lin, G., Li, J., Yang, W., Sun, C.: Crowd counting via weighted vlad on a dense attribute feature map. IEEE Trans. Circuits Syst. Video Technol. **28**(8), 1788–1797 (2016)
40. Shang, C., Ai, H., Bai, B.: End-to-end crowd counting via joint learning local and global count. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1215–1219. IEEE (2016)
41. Yao, H., Han, K., Wan, W., Hou, L.: Deep spatial regression model for image crowd counting. arXiv preprint arXiv:1710.09757 (2017)
42. Fang, Y., Gao, S., Li, J., Luo, W., He, L., Hu, B.: Multi-level feature fusion based locality-constrained spatial transformer network for video crowd counting. Neurocomputing (2020)
43. Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8198–8207 (2019)
44. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
45. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17–24. IEEE (2010)
46. Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., Schiele, B.: Recognizing fine-grained and composite activities using hand-centric features and script data. Int. J. Comput. Vis. **119**(3), 346–373 (2016)
47. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
48. Kim, S., Yun, K., Park, J., Choi, J.Y.: Skeleton-based action recognition of people handling objects. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 61–70. IEEE (2019)
49. Moghaddam, M.M.K., Abbasnejad, E., Shi, J.: Follow the attention: Combining partial pose and object motion for fine-grained action detection. arXiv preprint arXiv:1905.04430 (2019)
50. Lisanti, G., Masi, I., Bagdanov, A.D., Bimbo, A.D.: Person re-identification by iterative re-weighted sparse ranking. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1629–1642 (2015). https://doi.org/10.1109/TPAMI.2014.2369055
51. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro (2007)
52. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) Computer Vision—ECCV 2008, pp. 262–275. Springer, Berlin (2008)
53. Madden, C., Piccardi, M.: Height measurement as a session-based biometric for people matching across disjoint camera views. In: In Image and Vision Computing New Zealand, p. 29 (2005)
54. Pala, F., Satta, R., Fumera, G., Roli, F.: Multimodal person reidentification using rgb-d cameras. IEEE Trans. Circuits Syst. Video Technol. **26**(4), 788–799 (2016). https://doi.org/10.1109/TCSVT.2015.2424056
55. Dong Seon Cheng Marco Cristani, M.S.L.B., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference, pp. 68.1–68.11. BMVA Press (2011). https://doi.org/10.5244/C.25.68
56. Bąk, S., Corvee, E., Brémond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 179–184 (2011). https://doi.org/10.1109/AVSS.2011.6027316
57. Paolanti, M., Romeo, L., Liciotti, D., Pietrini, R., Cenci, A., Frontoni, E., Zingaretti, P.: Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. Sensors **18**(10), 3471 (2018)
58. Haque, A., Alahi, A., Fei-Fei, L.: Recurrent attention models for depth-based person identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1229–1238 (2016)
59. Lejbolle, A.R., Nasrollahi, K., Krogh, B., Moeslund, T.B.: Multimodal neural network for overhead person re-identification. In: 2017 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. IEEE (2017)
60. Lejbolle, A.R., Krogh, B., Nasrollahi, K., Moeslund, T.B.: Attention in multimodal neural networks for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 179–187 (2018)
61. Lejbølle, A.R., Nasrollahi, K., Krogh, B., Moeslund, T.B.: Person re-identification using spatial and layer-wise attention. IEEE Transactions on Information Forensics and Security (2019)
62. Liciotti, D., Frontoni, E., Mancini, A., Zingaretti, P.: Pervasive system for consumer behaviour analysis in retail environments. In: Video Analytics. Face and Facial Expression Recognition and Audience Measurement, pp. 12–23. Springer (2016)
63. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: arXiv preprint arXiv:1505.04597 (2015)
64. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
65. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
66. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
67. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
68. Carneiro, Z.L.G.: On the importance of normalisation layers in deep learning with piecewise linear activation units. Methods for Understanding and Improving Deep Learning Classification Models p. 58 (2017)
69. Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., Zingaretti, P.: Person re-identification dataset with rgb-d camera in a top-view configuration. In: Video Analytics. Face and Facial Expression Recognition and Audience Measurement, pp. 1–11. Springer (2016)

70. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In: CoRR (2015)
71. Larsson, G., Maire, M., Shakhnarovich, G.: Fractalnet: Ultra-deep neural networks without residuals. In: arXiv preprint arXiv:1605.07648 (2016)
72. Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V.: Learning and incorporating shape models for semantic segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 203–211 (2017)
73. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull. Soc. Vaudoise Sci. Nat. **37**, 547–579 (1901)
74. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
75. Frontoni, E., Paolanti, M., Pietrini, R.: People counting in crowded environment and re-identification. In: RGB-D Image Analysis and Processing, pp. 397–425. Springer (2019)
76. Zhang, X., Yan, J., Feng, S., Lei, Z., Yi, D., Li, S.Z.: Water filling: Unsupervised people counting via vertical Kinect sensor. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 215–220 (2012). https://doi.org/10.1109/AVSS.2012.82
77. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems, pp. 396–404 (1990)
78. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678. ACM (2014)
79. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. arXiv preprint arXiv:1707.07012 **2**(6) (2017)
80. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint pp. 1610–02357 (2017)
81. Bø, T.H., Dysvik, B., Jonassen, I.: Lsimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. **32**(3), e34–e34 (2004)
82. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
83. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
84. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
85. Rish, I.: An empirical study of the naive bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, pp. 41–46. IBM New York (2001)
86. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. arXiv preprint arXiv:1912.07863 (2019)
87. Zhong, Z., Jin, L., Xie, Z.: High performance offline handwritten chinese character recognition using googlenet and directional feature maps. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 846–850. IEEE (2015)
88. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
89. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
90. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: Second ACM/IEEE International Conference on Distributed Smart Cameras, 2008. ICDSC 2008. IEEE, pp. 1–6 (2008)
91. Li Y.and Wu, Z., Radke, R.: Multi-shot re-identification with random-projection-based random forests. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 373–380 (2015)
92. Bay, S.D.: Nearest neighbor classification from multiple feature subsets. In: Intelligent Data Analysis, pp. 191–209 (1999)
93. Prosser, B., Zheng, W., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC, vol. 2, p. 6 (2010)
94. Sorensen, H., Bogomolova, S., Anderson, K., Trinh, G., Sharp, A., Kennedy, R., Page, B., Wright, M.: Fundamental patterns of in-store shopper behavior. J. Retail. Consum. Serv. **37**, 182–194 (2017). https://doi.org/10.1016/j.jretconser.2017.02.003
95. Phillips, H., Bradshaw, R.: Camera tracking: a new tool for market research and retail management. Manag. Res. News **14**(4/5), 20–22 (1991). https://doi.org/10.1108/eb028133
96. Oosterlinck, D., Benoit, D.F., Baecke, P., de Weghe, N.V.: Bluetooth tracking of humans in an indoor environment: an application to shopping mall visits. Appl. Geogr. **78**, 55–65 (2017). https://doi.org/10.1016/j.apgeog.2016.11.005
97. Roedel, E.: 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 **13**(6), 429–430 (1970). https://doi.org/10.1002/bimj.19710130623. https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.19710130623
98. Cochran, W.G.: The combination of estimates from different experiments. Biometrics **10**(1), 101–129 (1954)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Marina Paolanti is a post-doc research fellow and contract professor at the Department of Information Engineering (DII) of Universitá Politecnica delle Marche. Her research focuses on deep learning, machine learning, image processing, and computer vision. During her PhD, she has worked at GfK Verein (Nuremberg, Germany) for visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. Her research focuses on deep learning and computer vision techniques applied to retail, social media intelligence and cultural heritage. She is a member of IEEE, CVPL and AI*IA.

Rocco Pietrini is a R&D Engineer at Grottini Lab, a company working worldwide in human behavior analysis in Retail environment. In 2020, he earned a PhD in Information Engineering from Universitá Politecnica delle Marche working on deep learning and image processing for human behavior analysis. He spent part of his PhD at the University of Central Florida (Orlando, USA) working on semantic segmentation for action detection.

Adriano Mancini received the PhD degree in intelligent artificial systems from the Department of Information Engineering (DII), Universitá Politecnica delle Marche, in 2010. His PhD thesis was on "A new methodological framework for land use/land cover mapping and change detection". He currently holds an assistant professor position with DII. His research focuses on mobile robotics also for assisted living, machine learning, image processing, and geographic information systems.

Emanuele Frontoni is currently a Professor of computer science with the Universitá Politecnica delle Marche, Italy. He received the PhD degree in intelligent artificial systems from the Department of Information Engineering (DII), Universitá Politecnica delle Marche, in 2006. His PhD thesis was on "vision-based robotics". His research focuses on artificial intelligence and computer vision techniques applied to robotics, internet of things, e-health, and ambient assisted living. He is a member of the ASME MESA TC, CVPL, and AI*IA.

Primo Zingaretti is currently Professor in the Department of Information Engineering at UNIVPM, the Polytechnic University of Marche in Ancona, Italy. He graduated in Electronic Engineering at the University of Ancona (former name of UNIVPM) in 1984. He has authored over 150 scientific research papers in English. He is a founding member (1988) of AI*IA (Italian Association for Artificial Intelligence), valued Member of the IEEE (Institute of Electrical and Electronics Engineers) since 1990 and then a Senior Member of the IEEE since 2007, a Member of the ASME (American Society of Mechanical Engineers) since 2011, and a member of CVPL (Italian Association for the research in computer vision, pattern recognition, and machine learning), founded in 1983 and affiliated to the IAPR (International Association for Pattern Recognition).