

Chapter 18

People Counting in Crowded Environment and Re-identification



Emanuele Frontoni, Marina Paolanti and Rocco Pietrini

Abstract Nowadays, detecting people and understanding their behaviour automatically is one of the key aspects of modern intelligent video systems. This interest arises from societal needs. Security and Video Analytics, Intelligent Retail Environment and Activities of Daily Living are just a few of the possible applications. The problem remains largely open due to several serious challenges such as occlusion, change of appearance, complex and dynamic background. Nevertheless, in recent years, privacy concerns are arising making these system designs more challenging, also to cope with different worldwide country regulations. Popular sensors for this task are RGB-D cameras because of their availability, reliability and affordability. Studies have demonstrated the great value (both in accuracy and efficiency) of depth camera in coping with severe occlusions among humans and complex background. In particular, RGB-D cameras show their great potential if used in a top-view configuration achieving high performances even in a crowded environment (considering at least 3 people per square meter in the area of the camera) minimizing occlusions and also being the most privacy-compliant approach. The first step in people detection and tracking is the segmentation to retrieve people silhouette, for this reason different methods will be covered in this chapter, ranging from classical handcraft feature based approaches to deep learning techniques. These techniques also solve the nontrivial problem of blob collision, occurring when two or more people are close enough to form a unique blob from the camera point of view. Multilevel segmentation and water filling algorithms will be presented to the reader in this chapter as handcraft feature based, in addition a deep learning approach is also introduced from the literature. In the methods presented in this chapter, the elaboration occurs live (there is no image recording) and occurs on the edge, following an IoT paradigm. Live analysis

E. Frontoni (✉) · M. Paolanti · R. Pietrini
Department of Information Engineering, Università Politecnica delle Marche,
Ancona 60121, Italy
e-mail: e.frontoni@univpm.it

M. Paolanti
e-mail: m.paolanti@univpm.it

R. Pietrini
e-mail: r.pietrini@pm.univpm.it

also strengthens the aforementioned concept of privacy compliance. The last part of this chapter is dedicated to person re-identification (re-id), which is the process to determine if different instances or images of the same person, recorded in different moments, belong to the same subject. Person re-id has many important applications in video surveillance, because it saves human efforts on exhaustively searching for a person from large amounts of video sequences. Identification cameras are widely employed in most of the public places like malls, office buildings, airports, stations and museums. These cameras generally provide enhanced coverage and overlay large geospatial areas because they have non-overlapping fields-of-views. Huge amounts of video data, monitored in real time by law enforcement officers are used after the event for forensic purposes, are provided by these networks. An automated analysis of these data improves significantly the quality of monitoring, in addition to processing the data faster. Handcrafted anthropomorphic features coupled with a machine learning approach will be exploited in this chapter, then a deep learning approach in comparison is presented. Different metrics are then adopted to evaluate the above algorithms and to compare them.

18.1 Introduction

Nowadays, detecting and tracking people is an utmost important and challenging task for various applications. It can be viably used in many interactive and intelligent systems such as visual surveillance and human–computer interaction. Recently, a considerable research has been made towards this topic and robust methods have been developed to track isolated or a small number of humans in case of the existence of transient occlusion [15, 72]. Nonetheless, tracking in crowded situations with a high number of people present in the image is experiencing many issues such as the exhibition of persistent occlusion, change of appearance, dynamic and complex background [41]. Indeed, these issues are the cause of the severe problems in the case of crowded environment since the conventional surveillance technologies cease to understand the image [68].

The methods dealing with people counting problems can be divided into twofold: detection-based methods and mapping-based methods. The first ones refer to running a detector, counting or clustering the output. The different features can include body, head, skin, hair, etc. For effective detection algorithms, they can have a high output accuracy for not highly crowded environments, but are not scalable for large crowds [68]. The mapping-based methods are referring to feature extraction and mapping them to a value. They use edge points, background, texture, optical flow, etc., as the features. Compared to detection-based methods, these methods can be scalable to large crowds. To address people detecting and tracking problems, sensors viably adopted are RGB-D cameras. Compared to conventional cameras, their performance results in increased reliability, availability and affordability. The efficiency and accuracy of depth cameras have been proven to be elevated in cases with severe occlusions among humans and complex background [53]. The combination of high-resolution

depth and visual information opens up new opportunities in many applications in the field of activity recognition and people tracking. Tracking and detection results can be significantly improved by the use of reliable depth maps [32].

In existing works, depth cameras are often placed either vertically overhead (top-view configuration) [8], or horizontally at the same level as humans (front-view configuration) [21]. The preferred choice is the RGB-D camera in a top-view configuration since it offers greater suitability compared with a front-view configuration. It moreover reduces the problem of occlusions and has the ability of being privacy preserving since a person's face is not recorded by the camera [32]. In [33], it has been shown that top-view people counting applications are the most accurate (with accuracy up to 99%) even in highly crowded scenarios, defined as situations with more than three people per square meter. However, this configuration is showing also an important limitation: inability to retrieve features connected to the front view, since the front-view configuration has been highly employed by many state-of-the-art approaches [60, 61].

In the literature exist several datasets using RGB-D technology for the study of person re-id mainly in the front-view configuration such as VIPeR [19], the iLIDS multi-camera tracking scenario [58], ETHZ [11], CAVIAR4REID [4, 6]. They cover many aspects of the existing problems such as shape deformation, occlusions, illumination changes, very-low-resolution images and image blurring. The top-view dataset is introduced in [31], called TVPR (Top- View Person Re-identification) dataset.

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

This chapter is organized as follows: Sect. 18.2 is giving the state of the art on the algorithms and approaches for person detection and tracking, Sect. 18.3 is giving results and use cases and in final Sect. 18.4 the conclusions are given.

18.2 Algorithms and Approaches

Many vision techniques and algorithms for person detection and tracking have been proposed during the past years. In this section, we give the state of the art on algorithms and techniques applied for tracking and detecting humans from top-view RGB-D data, covering both early and recent literature. In particular, the approaches related to segmentation using background subtraction, water filling, statistical algorithms, and finally machine learning and deep learning techniques are considered.

18.2.1 Approaches Using Background Subtraction

In most of the approaches, it is necessary to remove the background to obtain better accuracy in the next stages. In re-id, the methods that use the same reference back-

ground frame are not useful since data are grabbed from different cameras and with different backgrounds [49]. The most Naïve approach for background elimination method is the manual silhouette segmentation as proposed in [1]. Many works in the literature [48, 64, 65] use Gaussian mixture models (GMMs) for the background classification, introduced by [52]. The limitation is the sensitivity to fast illumination variations. Another approach is proposing the notion of a structure element [23], referring to a probabilistic element of an entire image class and widely employed for background elimination [12, 50]. The main advantage is that it can be used in the case of still images of a dataset with different backgrounds. However, the limitation is that it is time consuming. In [25], the authors use depth information from the Kinect camera, by exploiting the distance between the head and the floor.

In [67], a system is proposed for passenger counting in buses based on stereovision. The counting system involves different steps dedicated to the detection, segmentation, tracking and counting. The height maps have been segmented for highlighting the passengers' heads at different levels, resulting in kernels-binary images containing information related to the heads [34]. It makes use of the idea of computed tomography (CT) and the depth images are segmented to different layers along the transverse plane. The depth images are segmented into K layers as the CT slides with a fixed value of depth spacing between two adjacent layers. Then, based on the classic contour finding algorithm, the region of each slide can be found. Afterwards, an SVM classifier is trained to classify the activities. Another work is proposing a method of low-level segmentation and tracking, namely, the system detects the interactions with products on the shelves but also the movement of the people inside the store [43].

The authors [17] employ Microsoft Kinect depth sensor in an on-ceiling configuration and propose an automatic indoor fall detection method based on the analysis of depth frames. A segmentation algorithm is used to recognize the elements acquired. It extracts the elements and implements a solution to classify all the blobs in the scene. The human subjects are recognized by anthropometric relationships and features. In [10], an approach is presented for low-level body part segmentation based on RGB-D data gathered from the RGB-D sensor. The object classes are certain human body parts. In order to generate data for training the classifier, the authors make use of synthetic representation of the human body in a virtual environment in combination with Kinect skeleton tracking data.

The approach with multiple depth cameras is presented in [55], with the goal to develop real-time indoor surveillance system for tracking. The system tries to overcome the well-known problems such as severe occlusion, similar appearance, illumination changes, etc. It is based on the background subtraction of the stitched top-view images. Different phases have been employed in the detection scheme such as the graph-based segmentation, the head hemiellipsoid model, and the geodesic distance map, resulting in improved robustness and efficiency when compared to other state-of-the-art techniques.

The algorithm proposed in [24] shows an improvement of the classical segmentation techniques. It begins with nearest neighbour interpolation in order to fill the holes in the depth map. After that, the median filter with a 5×5 window on the depth

array is executed to smoothen the data. In order to extract the person, the algorithm extracts the floor and removes its corresponding pixels from the depth map. After that, to confirm the presence of the tracked subject as well as to provide head's location, a Support Vector Machine (SVM) based person finder is used. The algorithm shows promising results in achieving high sensitivity and specificity of fall detection in poor lighting conditions. Method for people counting in public transportation has also been presented in [36] using a segmentation approach. Data from Kinect sensor contains an image database of 1–5 persons, with and without body poses of holding a handrail. However, in this case, the image is processed in blocks with the goal to find potential local maxima, in order to find head candidates in the next step. Finally, non-head objects are filtered out.

The approach in [5] employs a novel active sensor based on a Time of Flight (TOF) technology applied in real time people tracking system, successful also in severe low-lighting conditions. A simple background subtraction procedure based on a pixel-wise parametric statistical model is performed. The system has proven to be reliable by the experiments conducted under changing lighting conditions and involving multiple people in close interaction.

A method using top-view camera system is presented in [45], used for human detection. The authors introduce a new feature descriptor to train a head-shoulder detector by the usage of discriminative class scheme, with excellent runtime performance. A final tracking step reliably propagates detection providing stable tracking results.

The papers [38, 39] have addressed the problem of the tracking of 3D human gestures by particle filtering and by using the Xtion PRO-LIVE camera. A hybrid 2D–3D method is proposed, consisting in separating human body in two parts; head and shoulders tracked in the 2D space, and arms tracked in the 3D space.

18.2.2 Water Filling

Water Filling is a computer vision algorithm adopted by many researchers while considering RGB-D cameras in top-view configuration.

In [71], the authors have built a system which uses vertical Kinect sensor for people counting, and the depth information is used to remove the effect of the appearance variation. They propose a novel unsupervised water filling method that can find suitable local minimum regions (people heads) with the property of robustness, locality and scale-invariance, even in crowded scenarios. The limitation is seen in the fact that water filling cannot handle the situation, where some moving object is closer to the sensor than head.

People counting system using water filling technique is also presented in [7]. In this case, the sensor is placed perpendicular to the ceiling. Water filling algorithm is used for the determination of the people's heads. Robot Operating System (ROS) is used to capture depth information from the Kinect sensor, and the images are

processed by using open-source library OpenCV. Some minor contributions have been obtained such as elimination of chattering, tracking the person lost in the image for a short time.

Water filling algorithm finds local minimum regions in depth images simulating the rain and the flooding of ground, with some raindrop, with a uniform distribution. Moving the raindrops towards the local minimum points puddles are formed since the water flows to some local minimum regions. Contour lines can be computed considering the distance from the local minimum as a function of the total raindrops. A depth image can be seen as a function $f(x, y)$ that can be non-derivable or even discontinuous, due to the noise of depth sensor. Finding people means to find local minimum regions in f . The problem can be defined as finding the region A and N satisfying the following equation [71]:

$$E_A(f(x, y)) + \eta \leq E_{N \setminus A}(f(x, y)) \quad (18.1)$$

where $A \in N$, A is the local region and N is its neighbourhood, both of arbitrary shape, $E(\cdot)$ is an operation to relate the depth information in the region to a real value reflecting the total depth information in the region. η is a threshold to ensure that depth in A should be lower than $N \setminus A$ within a tolerance. Zhang et al. [71] define an additional measure function $g(x, y)$ as

Definition 18.1 $g(x, y)$ is a measure function of $f(x, y) \iff \exists \epsilon > 0, \forall (x_1, y_1), (x_2, y_2)$, s.t. $\|(x_1 - x_2)^2 + (y_1 - y_2)^2\| < \epsilon$, if $f(x_1, y_1) \leq f(x_2, y_2)$

$$f(x_1, y_1) + g(x_1, y_1) \leq f(x_2, y_2) + g(x_2, y_2) \quad (18.2)$$

$$\begin{aligned} g(x_1, y_1) &\geq g(x_2, y_2) \\ g(x_1, y_1) &\geq 0, g(x_2, y_2) \geq 0 \end{aligned}$$

The form of $g(x, y)$ can be trivial, for example, a zero function. The use of $g(x, y)$ helps to infer the $f(x, y)$. Definition of $g(x, y)$ allows us to solve efficiently the equation and be robust to noise.

A general solution of $g(x, y)$ is not necessary for these contexts, but a proper nontrivial form is acceptable. The form of the function $f(x, y)$ can be seen as a land with humps and hollows. In the hump, raindrops flow to the neighbourhood hollow due to the force of gravity. Step by step the hollow region will be filled with raindrops. Function $g(x, y)$ represents the quantity of raindrop in the point (x, y) . When the rain stops, regions that collected rain drops can be classified as hollow. Algorithm 18.1 proposed by Liciotti et al. [30] is an improved version of the original method proposed in [71]. In particular, the drops are chosen according to the segmentation of the foreground image (line 4). This procedure improves the execution time of the algorithm. The main characteristics are depicted in Fig. 18.1.

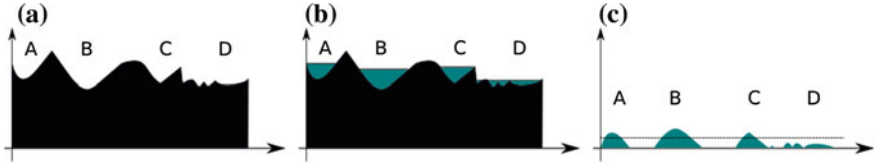


Fig. 18.1 The main characteristics of water filling algorithm. A, B, C correspond to three people, respectively, and D is a noise region (figure a). Region A has smaller scale compared with B and C, and the absolute height of A is larger than noise region D. After the water filling process (figure b), the measure function $g(x, y)$ which reflects the property of $f(x, y)$ is obtained (figure c). Finally, the people are detected by a threshold operation on measure function $g(x, y)$

Algorithm 18.1 Water Filling

```

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

```

The total number of raindrops is $K = tMN$, where t is usually set to be 100. Every iteration (line 5), (x, y) is randomly generated through a discrete uniform distribution (line 6). If there is a point (x^*, y^*) in the neighbourhood of (x, y) that satisfies Eq. 18.2 then the raindrop in (x, y) flows towards (x^*, y^*) and the loop is restarted until a local minimum is reached. When a local minimum is reached, the measure function $g(\cdot)$ is increased (line 13). After all the K raindrops find their stable places, measure function $g(x, y)$ is calculated and by applying a threshold T , it is possible to extract the heads of people that are under the camera (line 18).

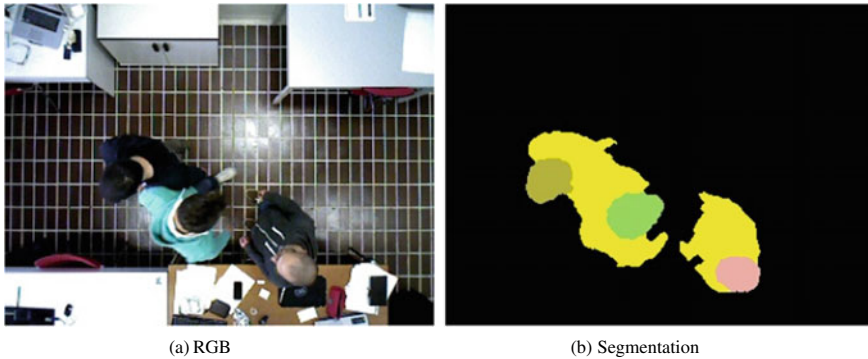


Fig. 18.2 Multilevel segmentation algorithm. Head recognition **(b)**: different colours of the blob highlight the head of the people detected in the scene **(a)**. Images from [29]

18.2.3 Multilevel Segmentation

The Multilevel Segmentation algorithm overcomes the limitations of common binary segmentation methods in case of collisions among people. In a normal segmentation, when two people collide they become a single blob (person). Multilevel segmentation ensures that, when a collision occurs, even if two people are in a single blob, their heads are detected and taken into account as discriminant elements [13]. In Fig. 18.2, it is possible to see the head of each person obtained by the multilevel segmentation algorithm. In case of collisions both people's heads are detected in the yellow blob (Figs. 18.2a and b). Multilevel segmentation is explained in detail in the pseudocode Algorithm 18.2. The MULTILEVELSEGM function takes the foreground image ($f(x, y)$) as input. FINDPOINTMAX function calculates the highest point in the image (max) and its location coordinates ($point_{max}$). In line 3, the level counter assumes the *threshold* value, a fixed value corresponding to average height of the human head ([13] adopted the value 10 cm, which is quite reasonable). When the segmentation level becomes negative (above the floor) the iteration stops. In line 5, there is a segmentation function that yields in output a binary image with blobs representative of moving objects that are above the segmentation level ($max - level$). This binary image is the input of FINDCONTOURS, a function that returns a vector of points for each blob. Then, the FILTERCONTOURS function deletes noise (condition on size and shape can be applied). The highest point/depth value (FINDPOINTMAX function) of each blob identified by means of the FILTERMASK function is then inserted in the vector *points*. Finally, MULTILEVELSEGM function returns a vector with all maximum local points. The length of this vector is exactly the number of people in the current frame.

Algorithm 18.2 Multi level segmentation algorithm

```

1: function MULTILEVELSEGMENT( $f(x, y)$ )
2:   ( $max, point_{max}$ ) = FINDPOINTMAX( $f(x, y)$ )
3:    $level = threshold$ 
4:   while ( $max - level$ ) > 0 do
5:      $f_{level}(x, y) = f(x, y) > (max - level)$ 
6:      $contours = FINDCONTOURS(f_{level}(x, y))$ 
7:     FILTERCONTOURS( $contours$ )
8:     for each contour  $i \in contours$  do
9:        $f_{mask}(x, y) = FILTERMASK(f_{level}(x, y), i)$ 
10:       $v_{max}, p_{max} = FINDPOINTMAX(f_{mask}(x, y))$ 
11:      if  $p_{max} \notin points$  then
12:         $points.PUSHBACK(p_{max})$ 
13:      end if
14:    end for
15:     $level = level + threshold$ 
16:  end while
17:  return  $points$ 
18: end function

```

18.2.4 Semantic Segmentation with Deep Learning

One of the main problems in computer vision is the semantic segmentation of images, video and 3D data. Semantic segmentation is one of the high-level tasks that leads to complete scene understanding. Scene understanding started with the goal of building machines that can see like humans to infer general principles and current situations from imagery, but it has become much broader than that. Applications such as image search engines, autonomous driving, computational photography, vision for graphics, human-machine interaction, were unanticipated and other applications keep arising as scene understanding technology develops [16]. As a core problem of high-level CV, while it has enjoyed some great success in the past 50 years, a lot more is required to reach a complete understanding of visual scenes. In the past, such a problem has been addressed using different traditional CV and machine learning techniques. Despite the popularity of those kinds of methods, the deep learning marked a significant change so that many CV problems are being tackled using deep architectures, usually Convolutional Neural Networks (CNNs), which are surpassing other approaches by a large margin in terms of accuracy and sometimes even efficiency. This section presents a particular case study describing five approaches from the literature based on CNN architectures and implementation methods for semantic segmentation. In this case, the goal is to segment people heads from a top-view configuration, so different CNN architectures have been tested for this. People heads are always visible from a top-view configuration and thus after successful detection,

counting became trivial, because we already have a binary mask to apply common computer vision algorithm to extract contours. In Sect. 18.3, a comparison between the different architectures is reported from [33].

18.2.4.1 U-NET

U-Net architecture proposed in [47] is shown in Fig. 18.3. It is composed of two main parts:

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

The first path follows the typical architecture of a CNN. It consists of the repeated application of two 3×3 convolutions (unpadded convolutions), each followed by a Rectified Linear Unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. At each downsampling step, the number of feature channels is doubled. Every step in the expansive path consists of an upsampling of the feature map followed by a 2×2 convolution (“up-convolution”) that halves the number of feature channels, a concatenation with the corresponding cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU. At the final layer, a 1×1 convolution is used to map each 32-component feature vector to the desired number of classes. Similarly, the authors of [46] revisited the classic U-Net by removing two levels of max pooling and changing the ReLU activation function with a LeakyReLU (Fig. 18.4). Another U-Net architecture is proposed in [33]. In particular, a batch normalization is added after the first ReLU activation function and after each max pooling and upsampling functions (Fig. 18.5).

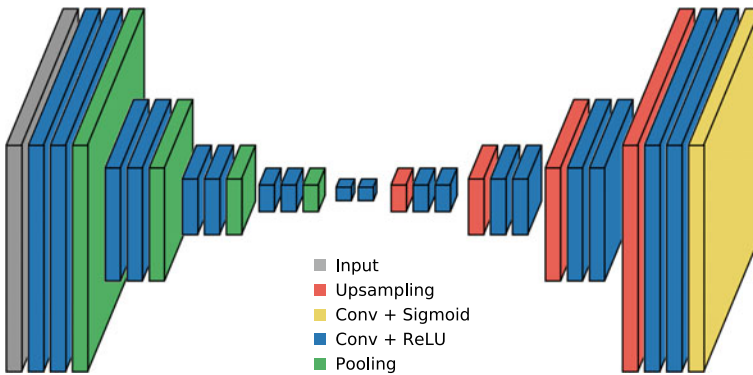


Fig. 18.3 U-Net architecture

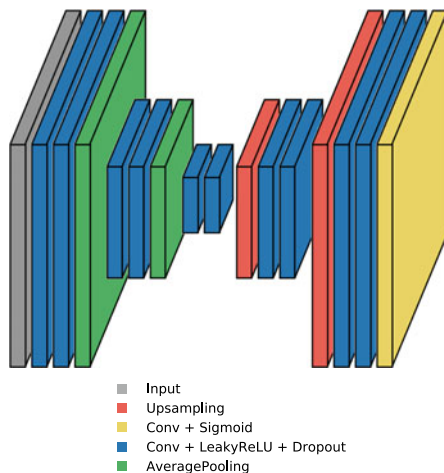


Fig. 18.4 U-Net2 architecture

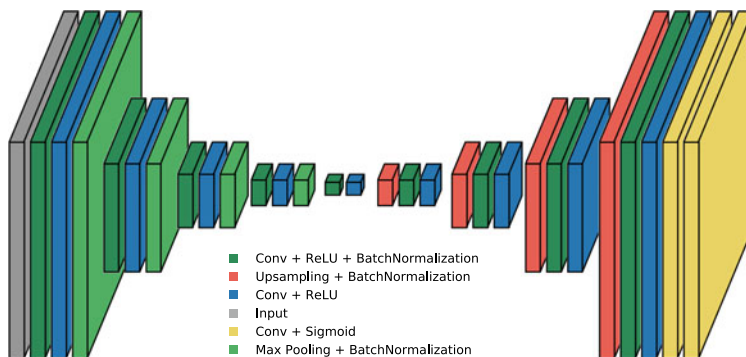


Fig. 18.5 U-Net3 architecture [33]

18.2.4.2 SegNet

SegNet, presented by Vijay et al. in [2], is depicted in Fig. 18.6. The architecture consists of a sequence of non-linear processing layers (encoders) and a corresponding set of decoders followed by a pixel-wise classifier. Typically, each encoder consists of one or more convolutional layers with batch normalization and a ReLU non-linearity, followed by non-overlapping max pooling and subsampling. The sparse encoding, due to the pooling process, is upsampled in the decoder using the max pooling indices in the encoding sequence. The max pooling indices are used in the decoders to perform upsampling of low-resolution feature maps. This has the important advantage of retaining high-frequency details in the segmented images and also reducing the total number of trainable parameters in the decoders. The entire architecture can be trained end-to-end using stochastic gradient descent. The

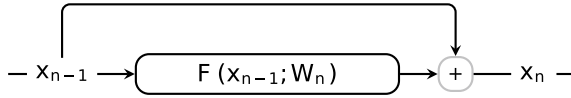


Fig. 18.7 Residual unit

18.2.4.4 FractalNet

Fractal network is introduced by Larsson et al. in [26]. Let C denote the index of a truncated fractal $f_C(\cdot)$ (i.e., a few stacked layers) and the base case of a truncated fractal is a single convolution:

$$f_1(z) = \text{conv}(z)$$

According to the expansion rule,

$$z' = \text{conv}(z)$$

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

can be defined recursively for the successive fractals, where $\oplus$ is a join operation and $\text{conv}(\cdot)$ is a convolution operator. Two blobs are merged by the join operation $\oplus$. As these two blobs contain features from different visual levels, joining them can enhance the discrimination capability of our network. Generally, this operation can be summation, maximization and concatenation. In order to enlarge the receptive field and enclose more contextual information, downsampling and upsampling operations are added in the above expansion rule. In particular, a max pooling with a stride of 2 and a deconvolution also with a stride of 2 are added. After the downsampling operation, the receptive field of a fractal becomes broader. When combining different

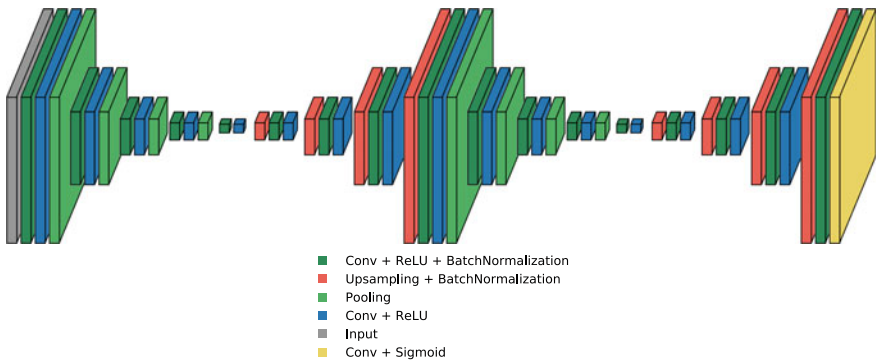


Fig. 18.8 FractalNet architecture

receptive fields through the join operation, the network can harness multi-scale visual cues and promote itself in discriminating. The Fractal Net architecture is depicted in Fig. 18.8.

18.2.5 *Re-identification*

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

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

An automated analysis of these data improves significantly the quality of monitoring, in addition to processing the data faster [56]. The behaviour characterization of people in a scene and their long-term activity can be possible using video analysis, which is required for high-level surveillance tasks in order to alert the security personnel. Over the past years, in the field of object recognition a significant amount of research has been performed by comparing video sequences. Colour-based features of video sequences are usually described with the use of a set of keyframes that characterizes well a video sequence. The HSV colour histogram and the RGB colour histogram are robust against the perspective and the variability of resolution [20]. The clothing colour histograms taken over the head, trousers and shirt regions, together with the approximated height of the person, have been used as discriminative features. Research works on person re-id can be divided into two categories: feature-based and learning-based [37]. The use of anthropometric measures for re-id was proposed for the first time in [35]. In this case, height was estimated from RGB cameras as a cue for associating tracks of individuals coming from non-overlapping views. In [18], the authors proposed the use of local motion features to re-identify people across camera views. They obtained correspondence between body parts of different persons through space-time segmentation. On these body parts, colour and edge histograms are extracted. In this approach, person re-id is performed by matching the body parts based on the features and correspondence. Shape and appearance context, which computes the co-occurrence of shape words and visual words for person re-id is proposed in [59]. Human body is partitioned into L parts with the shape context and a learned shape dictionary. Then, these parts are further segmented into

M subregions by a spatial kernel. The histogram of visual words is extracted on each subregion. Consequently, for the person re-id the $L \times M$ histograms are used as visual features. In [12], the appearance of a pedestrian is represented by combining three kinds of features (sampled according to the symmetry and asymmetry axes obtained from silhouette segmentation): the weighted colour histograms, the maximally stable colour regions and recurrent highly structured patches. Another method to face the problem of person re-id is learning discriminant models on low-level visual features. Adaboost is used to select an optimal ensemble of localized features for pedestrian recognition in [20]. The partial least squares method is used to perform person re-id in [51]. Instead, Prosser et al. [44] have used a ranking SVM to learn the ranking model. In the past years, it is well-known to use metric learning for person re-id. A probabilistic relative distance comparison model has been proposed in [73]. It maximizes the probability that the distance between a pair of true match is smaller than the distance between an incorrect match pair. In [40], the authors investigate whether the re-id accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. They also propose a dissimilarity-based framework for building and fusing the multimodal descriptors of pedestrian images for re-id tasks, as an alternative to the widely used score-level fusion.

Recently, CNNs are being widely employed to solve the problem of person re-id. Deep Learning models in the person re-id problem are still suffering from the lack of training data samples. The reason for this is that most of the datasets provide only two images per individual [27]. Several CNN models have been proposed in the literature to improve the performance of person re-id. Specifically, two models have been employed in re-id area: a classification model and a Siamese model based on either pair or triplet comparisons.

The model based on classification requires determining the individual identity. In [63], a novel feature extraction model called Feature Fusion Net (FFN) is proposed for pedestrian image representation. The presented model makes use of both CNN feature and handcrafted features. The authors utilize both colour histogram features and texture features. The extracted features are followed by a buffer layer and a fully connected layer which are acting as the fusion layer. The effectiveness was demonstrated on the three challenging datasets. In [62], a hybrid deep architecture for person re-id is presented, composed of Fisher vectors and multiple supervised layers. The network has been trained employing the linear discriminative analysis (LDA) as an objective function, with the goal of maximizing margin between classes. The authors in [66] propose a method based on learning deep feature representations from multiple domains by using CNNs with the aim to discover effective neurons for each training dataset. The authors propose Domain Guided Dropout algorithm in order to improve the feature learning process by discarding useless neurons. They evaluate on various datasets, with the limitation that some neurons are effective only for a specific data set and useless for another one. The authors in [28] designed a multi-scale context-aware network. The network is learning powerful features over the body and body parts. It can capture knowledge of the local context by stacking convolutions of multiple scales in each layer. They also propose to learn and

locate deformable pedestrian parts through networks of spatial transformers with new spatial restrictions, instead of using predefined rigid parts.

Since the person re-id research area lacks training instances, Siamese network models have been widely and viably employed. Siamese neural network is a type of neural network architectures which contains two or more identical sub-networks. A Siamese network is employed as pairwise (in the case of two sub-networks), or triplet (the case of three sub-networks). Some examples of pairwise research can be found in [14, 57, 69]. The authors in [57] combined four CNNs, each of them embedding images from different scale or different body part. Each of sub-CNN is trained with adaptive list-wise loss function. In addition, they adopted sharpness parameter and an adaptive margin parameter to automatically focus more on the hard negative samples in the training process. In [69], a Siamese neural network has been proposed to learn pairwise similarity. The method can learn at the same time the colour feature, texture feature and metric in a unified framework. The network is a symmetrical structure containing two sub-networks, which are connected by Cosine function. Binomial deviance is also used to deal with the big variations of person images [14]. The authors propose a novel type of features based on covariance descriptors—the convolutional covariance features. There are three steps, first a hybrid network is trained for person recognition, next another hybrid network is employed to discriminate the gender, and finally the output of the two networks are passed through the coarse-to-fine transfer learning method to a pairwise Siamese network in order to accomplish the final person re-id. In [9], the authors presented a scalable distance driven feature learning framework based on the deep neural network in order to produce feature representation from a raw person images. A CNN network is trained by a set of triplets to produce features that can satisfy the relative distance constraints. In [70], a supervised learning framework is proposed to generate compact and bit-scalable hashing codes from raw images. Training images were organized into a batch of triplet samples, two images with the same label and one with a different label. The deep convolutional neural network is utilized to train the model in an end-to-end fashion, with the simultaneous optimization of the discriminative image features and hash functions. In [54], a three-stage training is proposed: a deep convolutional neural network is first trained on an independent dataset labelled with attributes, then it is fine-tuned on another dataset that is only labelled with person IDs using a particular triplet loss they define, and finally, the updated network predicts attribute labels for the target dataset.

18.2.5.1 Person Re-identification in Top-View Configuration

The re-id in top-view configuration has been studied in [31, 42]. In order to face this task, the authors have built a dataset namely TVPR¹ (Top-View Person Re-identification). An Asus Xtion PRO LIVE RGB-D camera has been used because it allows to acquire colour and depth information in an affordable and fast way. The

¹<http://vrai.dii.univpm.it/re-id-dataset>.

camera is installed on the ceiling above the area to be analysed. Data of 100 people are collected, acquired across intervals of days and at different times. Each person walked with an average gait within the recording area in one direction, stopping for a few seconds just below the camera, then it turned around and repeated the same route in the opposite direction, always stopping under the camera for a while. This choice is due to its greater suitability compared with a front-view configuration, usually adopted for gesture recognition or even for video gaming. The top-view configuration reduces the problem of occlusions [30] and has the advantage of being privacy preserving, because the face is not recorded by the camera. The process of extraction of a high number of significant features derived from both depth and colour information is presented. The first step is the processing of the data acquired from the RGB-D camera. The camera captures depth and colour images, both with dimensions of 640×480 pixels, at a rate up to approximately 30 fps and illuminates the scene/objects with structured light based on infrared patterns. Seven out of the nine features selected are the anthropometric features extracted from the depth image:

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

The remaining two *colour-based features* are acquired by the colour image, such as in [3], with $n = 10$ bin quantization, for both H channel and S channel.

- colour histogram for the head/hair, H_h ;
- colour histogram for the outwear, H_o .

It has also been defined in three descriptors: TVH , TVD and $TVDH$.

- TVH is the colour descriptor:

$$TVH = \{H_h^p, H_o^p\} \quad (18.4)$$

- TVD is the depth descriptor:

$$TVD = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p\} \quad (18.5)$$

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

$$TVDH = \{d_1^p, d_2^p, d_3^p, d_4^p, d_5^p, d_6^p, d_7^p, H_h^p, H_o^p\} \quad (18.6)$$

Figure 18.9 depicts the set of features considered: anthropometric and the colour-based ones.

The 100 people dataset was acquired in 23 registration sessions. Each of the 23 folders contains the video of one registration session. The recording time [s] for

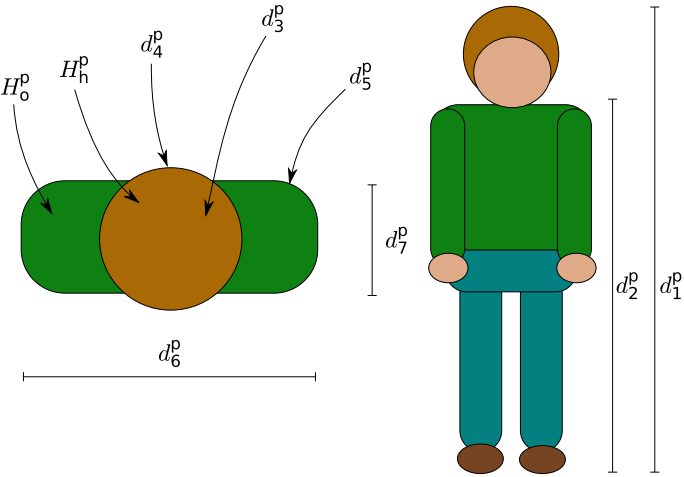


Fig. 18.9 Anthropometric and colour-based features

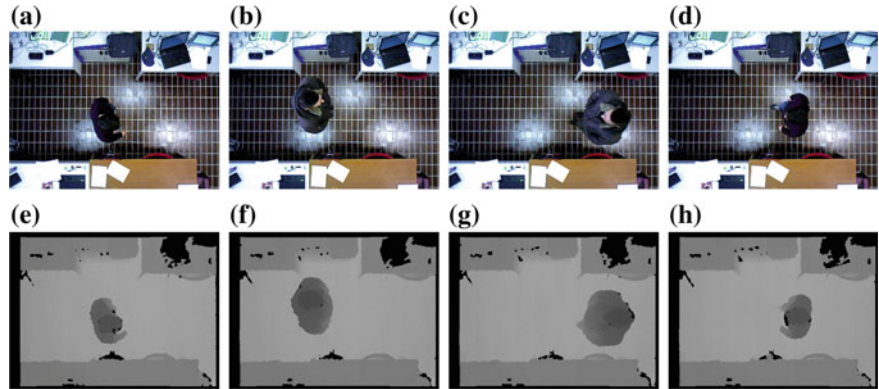


Fig. 18.10 Snapshots of a registration session of the recorded data, in an indoor scenario, with artificial light. People had to pass under the camera installed on the ceiling. The sequences **a–e**, **b–f** correspond to the sequences **d–h**, **c–g** respectively for the training and testing sets of the classes 8–9 for the registration session g003 [31]

the session and the number of persons of that session are reported in Table 18.1. Acquisitions have been performed in 8 days and the total recording time is about 2000s. Registrations are made in an indoor scenario, where people pass under the camera installed on the ceiling. Another big issue is environmental illumination. In each recording session, the illumination condition is not constant, because it varies in function of the different hours of the day and it also depends on natural illumination due to weather conditions. The video acquisitions, in this scenario, are depicted in Fig. 18.10, which are examples of person registration respectively with sunlight and artificial light. Each person during a registration session walked with an average gait

Table 18.1 Time [s] of registration for each session and the number of people of that session [31]

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

within the recording area in one direction, then it turned back and repeated the same route in the opposite direction. This methodology is used for a better split of TVPR in training set (the first passage of the person under the camera) and testing set (when the person passed again under the camera). The recruited people are aged between 19–36 years: 43 females and 57 male; 86 with dark hair, 12 with light hair and 2 are hairless. Furthermore, of these people 55 have short hair, 43 have long hair. The subjects were recorded in their everyday clothing like T-shirts/sweatshirts/shirts, loose-fitting trousers, coats, scarves and hats. In particular, 18 subjects wore coats and 7 subjects wore scarves. All videos have fixed dimensions and a frame rate of about 30 fps. Videos are saved in native .oni files, but can be converted to any other format. Colour stream is available in a non-compressed format. Figure 18.11 reports the histogram of each extracted anthropometric feature. Due to the dissimilarity of the analysed subjects a Gaussian curve is obtained from the data. The Cumulative Matching Characteristic (CMC) curve represents the expectation of finding the correct match in the top n matches. It is equivalent to the Receiver Operating Characteristic (ROC) curve in detection problems. This performance metric evaluates recognition problems, by some assumptions about the distribution of appearances in a camera network. It is considered the primary measure of identification performance among biometric researchers. As well-established in recognition and in re-id tasks, for each testing item we ranked the training gallery elements using standard distance metrics. Three distance measures have been examined as the matching distance metrics: the L1 City block, the Euclidean Distance and the Cosine Distance. To evaluate the TVPR, performance results are reported in terms of recognition rate, using the CMC curves, illustrated in Fig. 18.12. In particular, the horizontal axis is the rank of the matching score, and the vertical axis is the probability of correct identification. Considering the dataset, a comparison among TVH and TVD in terms of CMC curves are depicted, to

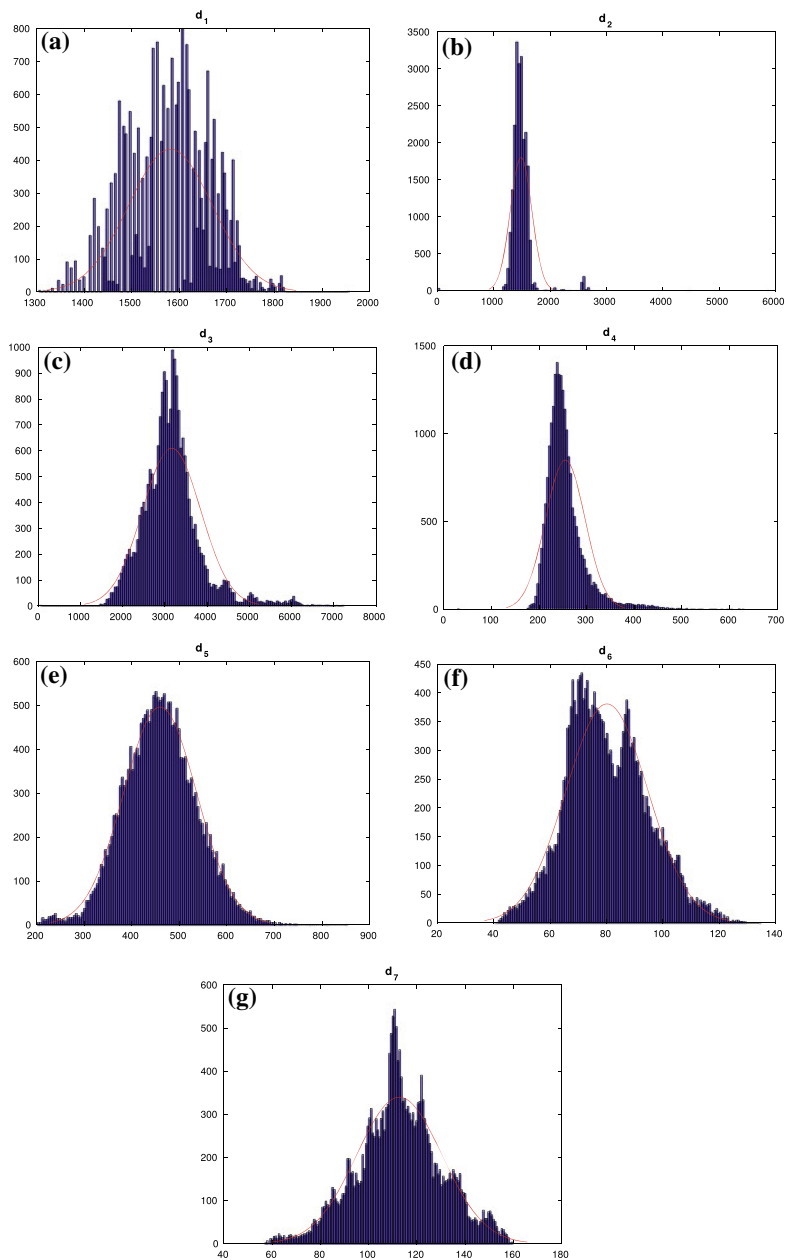


Fig. 18.11 Statistics histogram for each feature (18.11a d_1 distance between floor and head; 18.11b d_2 distance between floor and shoulders; 18.11c d_3 area of head surface; 18.11d d_4 Head circumference, 18.11e d_5 shoulders circumference, 18.11f d_6 shoulders breadth; 18.11g d_7 thoracic antero-posterior depth). The resultant Gaussian curve (in red) is due to the dissimilarity of the analysed subjects [31]

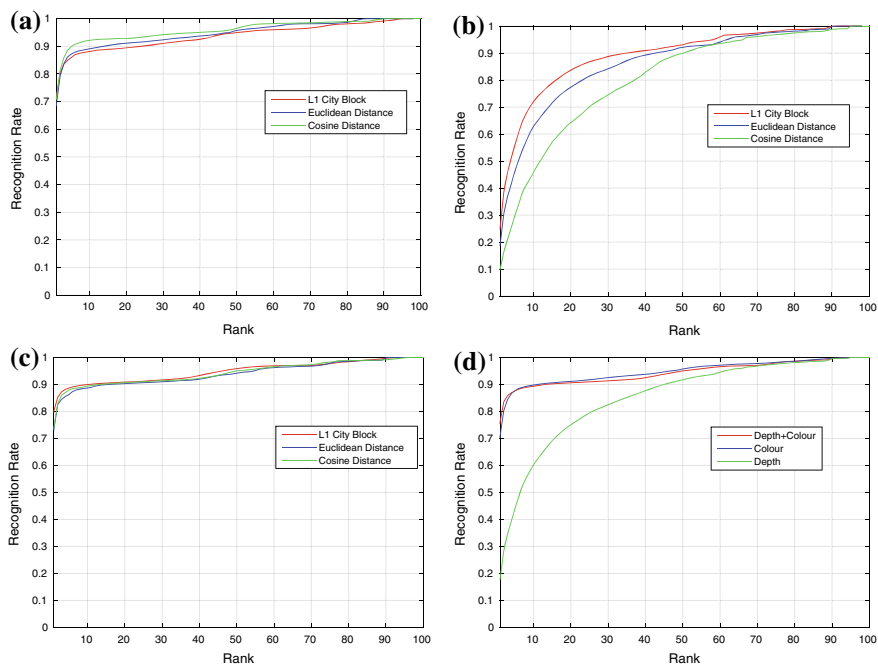


Fig. 18.12 The CMC curves obtained on TVPR Dataset [42]

compare the ranks returned by using these different descriptors. Figure 18.12a provides the CMC obtained for TVH. Figure 18.12b represents the CMC obtained for TVD. These results are compared with the average obtained by TVH and TVD. The average CMC is displayed in Fig. 18.12c. It is observed that the best performance is achieved by the combination of descriptors. In Fig. 18.12d, it can be seen that the combination of descriptors improves the results obtained by each of the descriptor separately. This result is due to the depth contribution that can be more informative. In fact, the depth outperforms the colour, giving the best performance for rank values higher than 15 (Fig. 18.12b). Its better performance suggests the importance and potential of this descriptor.

18.3 Results and Use Cases

In this section, the results of experiments performed for testing the performance of multilevel segmentation and water filling, DCNNs for people counting and re-id feature based approach are reported.

Table 18.2 Image processing algorithms performances

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

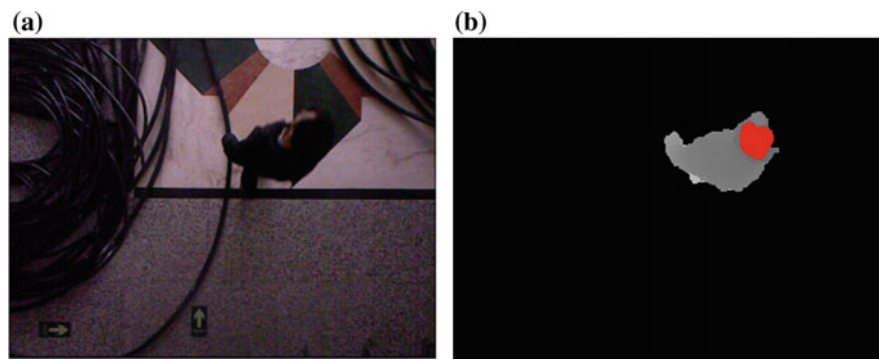


Fig. 18.13 CBSR dataset

In particular, the multilevel segmentation and water filling algorithms are assessed evaluating a restricted part of CBSR Dataset.² This dataset includes a total of 3884 images with 6094 heads. It contains depth images after background subtraction and in the ground truth the heads are manually painted as red colour.

Table 18.2 shows the results of algorithms in term of Precision, Recall and F1-score. The algorithms reach high values of performances. However, when the heads are along the edge of image their accuracies are decreased. Instead, the multilevel segmentation algorithm looks more accurate than water filling algorithm (Fig. 18.13).

Regarding the evaluation of deep learning approaches *TVHeads*³ dataset has been used. It contains depth images of people from top-view configuration. In particular, the purpose of this dataset is to localise the heads of people who are present below the camera. It contains a total of 1815 depth images (16 bit) with a dimension of 320×240 pixels. Furthermore, after an image preprocessing phase, the depth images are also converted, with an appropriate scaling, in order to obtain images (8 bit), where the heads silhouette is highlighted by improving image contrast and brightness (Fig. 18.14).

Each CNN implementation is trained with two types of depth images:

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

²<https://goo.gl/MwtWKR>.
³<http://vrai.dii.univpm.it/tvheads-dataset>.



Fig. 18.14 TVHeads dataset. It consists of 8-bit scaled depth images (a), 16-bit original images (b) and the corresponding ground truth (c) [33]

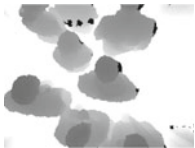
Table 18.3 Jaccard and Dice indices of different CNN architectures. From [33]

Net	Bit	Jaccard	Jaccard	Dice	Dice
		Train	Validation	Train	Validation
Fractal [26]	8	0.960464	0.948000	0.979833	0.973306
	16	0.961636	0.947762	0.980443	0.973180
U-Net [47]	8	0.896804	0.869399	0.945595	0.930138
	16	0.894410	0.869487	0.944262	0.930188
U-Net2 [46]	8	0.923823	0.939086	0.960403	0.968586
	16	0.923537	0.938208	0.960249	0.968119
U-Net3[33]	8	0.962520	0.931355	0.980902	0.964458
	16	0.961540	0.929924	0.980393	0.963690
SegNet [2]	8	0.884182	0.823731	0.938531	0.903347
	16	0.884162	0.827745	0.938520	0.905756
ResNet [22]	8	0.932160	0.856337	0.964889	0.922609
	16	0.933436	0.848240	0.965572	0.917889

Table 18.4 Semantic segmentation results of different ConvNet architectures. From [33]

Net	Bit	Accuracy	Precision	Recall	F1-score
Fractal [26]	8	0.994414	0.991400	0.993120	0.992235
	16	0.994437	0.992667	0.993297	0.992970
U-Net [47]	8	0.992662	0.946475	0.950483	0.948408
	16	0.992569	0.945083	0.948957	0.946938
U-Net2 [46]	8	0.993156	0.970013	0.969206	0.969568
	16	0.993165	0.967884	0.970557	0.969123
U-Net3[33]	8	0.994572	0.990451	0.990387	0.990419
	16	0.994559	0.989382	0.989411	0.989396
SegNet [2]	8	0.992683	0.946304	0.953136	0.949625
	16	0.992699	0.946237	0.953342	0.949658
ResNet [22]	8	0.993789	0.968399	0.968374	0.968359
	16	0.993819	0.968765	0.969256	0.968992

Table 18.5 Qualitative result of prediction. From [33]

	8-bit	16-bit	Label
			
FractalNet [26]			
U-Net [47]			
U-Net2 [46]			
U-Net3 [33]			
SegNet [2]			
ResNet [22]			

In this case, training, test and validation are chosen, respectively, to learn model parameters. Once this phase is completed, the best model is also evaluated over the never before seen test set. In the following experiments, 70%, 10% and 20% of the datasets are chosen, respectively, for training, test and validation. Furthermore, different combinations of hyperparameters are tested, a learning rate equal to 0.001 and an Adam optimization algorithm have been used. Semantic segmentation performances are divided into two different tables. Table 18.3 shows Jaccard and Dice indices for training and for validation, respectively, while Table 18.4 reported the results in terms of accuracy, precision, recall and F1-score. Both tables refer to a learning process conducted during 200 epochs. In Table 18.3, the best CNN architecture is the U-Net3 8-bit version. Indeed, Jaccard index reaches a value equal to 0.962520. The second best is Fractal Net 16-bit version also obtaining higher values as regards validation performances. Table 18.4 reports the best CNN architecture, in terms of accuracy, is U-Net3 8-bit version, while Fractal Net 16-bit version exceeds slightly in precision, recall and F1-score metrics. Qualitative results in Table 18.5.

18.4 Conclusions

The RGB-D cameras installed in top-view configuration have several advantages for tracking and detecting people especially in a heavy crowded environment with values of accuracy that reaches 99%. The aim of this chapter is to demonstrate the potential of these sensors installed in top-view configuration for two tasks: people counting and re-id. These kind of solutions are successfully applied in several applications because of their great suitability compared with a front-view configuration, usually adopted for gesture recognition or even for video gaming. The combination of depth information coupled with visual images provides their success in retail domain.

Future works would include the integration of these systems with an audio framework and the use of other types of RGB-D sensors, such as time of flight (TOF) ones. The system can additionally be integrated as a source of high semantic-level information in a networked ambient intelligence scenario, to provide cues for different problems, such as detecting abnormal speed and dimension outliers, that can alert of a possible uncontrolled circumstance. Further investigation will be devoted to improve the top-view configuration approach for people counting and re-id by extracting other comprehensive features and setting up it for the real-time processing of video images, in particular, in the retail scenario.

References

1. Annesley J, Orwell J, Renno J (2005) Evaluation of MPEG7 color descriptors for visual surveillance retrieval. In: 2005 IEEE international workshop on visual surveillance and performance

- evaluation of tracking and surveillance, pp 105–112. <https://doi.org/10.1109/VSPETS.2005.1570904>
2. Badrinarayanan V, Kendall A, Cipolla R (2015) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint [arXiv:1511.00561](https://arxiv.org/abs/1511.00561)
 3. Baltieri D, Vezzani R, Cucchiara R (2013) Learning articulated body models for people re-identification. In: Proceedings of the 21st ACM international conference on multimedia, pp 557–560. ACM
 4. Barbosa BI, Cristani M, Del Bue A, Bazzani L, Murino V (2012) Re-identification with RGB-D sensors. In: First international workshop on re-identification
 5. Bevilacqua A, Stefano LD, Azzari P (2006) People tracking using a time-of-flight depth sensor. In: 2006 IEEE international conference on video and signal based surveillance, pp 89–89. <https://doi.org/10.1109/AVSS.2006.92>
 6. Cheng DS, Cristani M, Stoppa M, Bazzani L, Murino V (2011) Custom pictorial structures for re-identification. In: Proceedings of the british machine vision conference, pp 68.1–68.11. BMVA Press. <http://dx.doi.org/10.5244/C.25.68>
 7. Coşkun A, Kara A, Parlaktuna M, Ozkan M, Parlaktuna O (2015) People counting system by using Kinect sensor. In: 2015 International symposium on innovations in intelligent systems and applications (INISTA), pp 1–7. <https://doi.org/10.1109/INISTA.2015.7276740>
 8. Dan B, Kim YS, Jung J, Ko S (2012) Robust people counting system based on sensor fusion. *IEEE Trans Consum Electron* 58(3):1013–1021. <https://doi.org/10.1109/TCE.2012.6311350>
 9. Ding S, Lin L, Wang G, Chao H (2015) Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit* 48(10):2993–3003. <https://doi.org/10.1016/j.patcog.2015.04.005>. <http://www.sciencedirect.com/science/article/pii/S0031320315001296>
 10. Ditttrich F, Woern H, Sharma V, Yayilgan S (2014) Pixelwise object class segmentation based on synthetic data using an optimized training strategy. In: 2014 First international conference on networks & soft computing (ICNSC2014), pp 388–394. IEEE
 11. Ess A, Leibe B, Gool LV (2007) Depth and appearance for mobile scene analysis. In: 2007 IEEE 11th international conference on computer vision, pp 1–8. <https://doi.org/10.1109/ICCV.2007.4409092>
 12. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 2360–2367. <https://doi.org/10.1109/CVPR.2010.5539926>
 13. Ferracuti N, Norscini C, Frontoni E, Gabellini P, Paolanti M, Placidi V (2019) A business application of RTLS technology in intelligent retail environment: Defining the shopper's preferred path and its segmentation. *J Retail Consum Serv* 47:184–194
 14. Franco A, Oliveira L (2017) Convolutional covariance features: conception, integration and performance in person re-identification. *Pattern Recognit* 61:593–609. <https://doi.org/10.1016/j.patcog.2016.07.013>. <http://www.sciencedirect.com/science/article/pii/S0031320316301625>
 15. Frontoni E, Zingaretti P (2005) A vision based algorithm for active robot localization. In: 2005 International symposium on computational intelligence in robotics and automation, pp 347–352. IEEE
 16. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J (2017) A review on deep learning techniques applied to semantic segmentation. [arXiv:1704.06857](https://arxiv.org/abs/1704.06857)
 17. Gasparrini S, Cipitelli E, Spinsante S, Gambi E (2014) A depth-based fall detection system using a Kinect® sensor. *Sensors* 14(2):2756–2775. <https://doi.org/10.3390/s140202756>. <http://www.mdpi.com/1424-8220/14/2/2756>
 18. Gheissari N, Sebastian TB, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: IEEE conference on computer vision and pattern recognition, pp 1528–1535. IEEE
 19. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: IEEE international workshop on performance evaluation for tracking and surveillance, Rio de Janeiro

20. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European conference on computer vision, pp 262–275. Springer
21. Han J, Pauwels EJ, de Zeeuw PM, de With PHN (2012) 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. <https://doi.org/10.1109/TCE.2012.6227420>
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
23. Jojic N, Perina A, Cristani M, Murino V, Frey B (2009) Stel component analysis: modeling spatial correlations in image class structure. In: 2009 IEEE conference on computer vision and pattern recognition, pp 2044–2051. <https://doi.org/10.1109/CVPR.2009.5206581>
24. Kepski M, Kwolek B (2014) Detecting human falls with 3-axis accelerometer and depth sensor. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 770–773. <https://doi.org/10.1109/EMBC.2014.6943704>
25. Kouno D, Shimada K, Endo T (2012) Person identification using top-view image with depth information. In: 2012 13th ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing, pp 140–145. <https://doi.org/10.1109/SNPD.2012.47>
26. Larsson G, Maire M, Shakhnarovich G (2016) Fractalnet: ultra-deep neural networks without residuals. [arXiv:1605.07648](https://arxiv.org/abs/1605.07648)
27. Lavi B, Serj MF, Ullah I (2018) Survey on deep learning techniques for person re-identification task. *CoRR*. [arXiv:1807.05284](https://arxiv.org/abs/1807.05284)
28. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification
29. Liciotti D, Frontoni E, Mancini A, Zingaretti P (2016) Pervasive system for consumer behaviour analysis in retail environments. Video analytics. Face and facial expression recognition and audience measurement. Springer, Berlin, pp 12–23
30. Liciotti D, Massi G, Frontoni E, Mancini A, Zingaretti P (2015) Human activity analysis for in-home fall risk assessment. In: 2015 IEEE international conference on communication workshop (ICCW), pp 284–289. IEEE
31. Liciotti D, Paolanti M, Frontoni E, Mancini A, Zingaretti P (2017) Person re-identification dataset with RGB-D camera in a top-view configuration. In: Nasrollahi K, Distant C, Hua G, Cavallaro A, Moeslund TB, Battiato S, Ji Q (eds) Video analytics. Face and facial expression recognition and audience measurement. Springer International Publishing, Cham, pp 1–11
32. Liciotti D, Paolanti M, Frontoni E, Zingaretti P (2017) People detection and tracking from an RGB-D camera in top-view configuration: review of challenges and applications. In: Battiato S, Farinella GM, Leo M, Gallo G (eds) New trends in image analysis and processing - ICIAP 2017, pp 207–218. Springer International Publishing, Cham
33. Liciotti D, Paolanti M, Pietrini R, Frontoni E, Zingaretti P (2018) Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment. In: 2018 24th international conference on pattern recognition (ICPR), pp 1384–1389. IEEE
34. Lin S, Liu A, Hsu T, Fu L (2015) Representative body points on top-view depth sequences for daily activity recognition. In: 2015 IEEE international conference on systems, man, and cybernetics, pp 2968–2973. <https://doi.org/10.1109/SMC.2015.516>
35. Madden C, Piccardi M (2005) Height measurement as a session-based biometric for people matching across disjoint camera views. In: Image and vision computing conference. Wickliffe Ltd
36. Malawski F (2014) Top-view people counting in public transportation using Kinect. *Chall Mod Technol* 5(4):17–20
37. Messelodi S, Modena CM (2015) Boosting Fisher vector based scoring functions for person re-identification. *Image Vis Comput* 44:44–58
38. Migniot C, Ababsa F (2013) 3D human tracking in a top view using depth information recorded by the xtion pro-live camera. In: Bebis G, Boyle R, Parvin B, Koracin D, Li B, Porikli F, Zordan V, Klosowski J, Coquillart S, Luo X, Chen M, Gotz D (eds) Advances in visual computing. Springer, Berlin, pp 603–612

39. Migniot C, Ababsa F (2016) Hybrid 3D–2D human tracking in a top view. *J R-Time Image Process* 11(4):769–784. <https://doi.org/10.1007/s11554-014-0429-7>
40. Pala F, Satta R, Fumera G, Roli F (2016) Multimodal person reidentification using RGB-D cameras. *IEEE Trans Circuits Syst Video Technol* 26(4):788–799
41. Paolanti M, Liciotti D, Pietrini R, Mancini A, Frontoni E (2018) Modelling and forecasting customer navigation in intelligent retail environments. *J Intell Robot Syst* 91(2):165–180
42. Paolanti M, Romeo L, Liciotti D, Pietrini R, Cenci A, Frontoni E, Zingaretti P (2018) 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
43. Paolanti M, Romeo L, Martini M, Mancini A, Frontoni E, Zingaretti P (2019) Robotic retail surveying by deep learning visual and textual data. *Robot Auton Syst* 118:179–188
44. Prosser BJ, Zheng WS, Gong S, Xiang T, Mary, Q (2010) Person re-identification by support vector ranking. In: *Bmvc*, vol 2, p 6
45. Rauter M (2013) Reliable human detection and tracking in top-view depth images. In: 2013 IEEE conference on computer vision and pattern recognition workshops, pp 529–534. <https://doi.org/10.1109/CVPRW.2013.84>
46. Ravishankar H, Venkataramani R, Thiruvankadam S, Sudhakar P, Vaidya V (2017) Learning and incorporating shape models for semantic segmentation. In: International conference on medical image computing and computer-assisted intervention, pp 203–211. Springer
47. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation
48. Roy A, Sural S, Mukherjee J (2012) A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification. *Pattern Recognit Lett* 33(14):1891–1901. <https://doi.org/10.1016/j.patrec.2012.02.003>. <http://www.sciencedirect.com/science/article/pii/S0167865512000359>. Novel pattern recognition-based methods for re-identification in biometric context
49. Saghaei MA, Hussain A, Zaman HB, Saad MHM (2014) Review of person re-identification techniques. *IET Comput Vis* 8(6):455–474. <https://doi.org/10.1049/iet-cvi.2013.0180>
50. Satta R, Fumera G, Roli F (2011) Exploiting dissimilarity representations for person re-identification. In: Pelillo M, Hancock ER (eds) *Similarity-based pattern recognition*. Springer, Berlin, pp 275–289
51. Schwartz WR, Davis LS (2009) Learning discriminative appearance-based models using partial least squares. In: 2009 XXII Brazilian symposium on computer graphics and image processing (SIBGRAPI), pp 322–329. IEEE
52. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. In: *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, vol 2, pp 246–252. <https://doi.org/10.1109/CVPR.1999.784637>
53. Sturari M, Liciotti D, Pierdicca R, Frontoni E, Mancini A, Contigiani M, Zingaretti P (2016) Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognit Lett* 81:30–40. <https://doi.org/10.1016/j.patrec.2016.02.010>. <http://www.sciencedirect.com/science/article/pii/S016786551600057X>
54. Su C, Zhang S, Xing J, Gao W, Tian Q (2016) Deep attributes driven multi-camera person re-identification. In: Leibe B, Matas J, Sebe N, Welling M (eds) *Computer vision - ECCV 2016*. Springer International Publishing, Cham, pp 475–491
55. Tseng T, Liu A, Hsiao P, Huang C, Fu L (2014) Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In: 2014 IEEE/RSJ international conference on intelligent robots and systems, pp. 4077–4082. <https://doi.org/10.1109/IROS.2014.6943136>
56. Tu PH, Doretto G, Krahnstoeber NO, Perera AA, Wheeler FW, Liu X, Rittscher J, Sebastian TB, Yu T, Harding, KG (2007) An intelligent video framework for homeland protection. In: *Unattended ground, sea, and air sensor technologies and applications IX*, vol 6562, p 65620C. International Society for Optics and Photonics
57. Wang J, Wang Z, Gao C, Sang N, Huang R (2017) DeepList: learning deep features with adaptive listwise constraint for person reidentification. *IEEE Trans Circuits Syst Video Technol* 27(3):513–524. <https://doi.org/10.1109/TCSVT.2016.2586851>

58. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) *Computer vision - ECCV 2014*. Springer International Publishing, Cham, pp 688–703
59. Wang X, Doretto G, Sebastian T, Rittscher J, Tu P (2007) Shape and appearance context modeling
60. Wang Z, Hu R, Liang C, Yu Y, Jiang J, Ye M, Chen J, Leng Q (2016) Zero-shot person re-identification via cross-view consistency. *IEEE Trans Multimed* 18(2):260–272. <https://doi.org/10.1109/TMM.2015.2505083>
61. Wu A, Zheng W, Lai J (2017) Robust depth-based person re-identification. *IEEE Trans Image Process* 26(6):2588–2603. <https://doi.org/10.1109/TIP.2017.2675201>
62. Wu L, Shen C, van den Hengel A (2017) Deep linear discriminant analysis on Fisher networks: a hybrid architecture for person re-identification. *Pattern Recognit* 65:238–250. <https://doi.org/10.1016/j.patcog.2016.12.022>. <http://www.sciencedirect.com/science/article/pii/S0031320316304447>
63. Wu S, Chen YC, Li X, Wu AC, You JJ, Zheng WS (2016) An enhanced deep feature representation for person re-identification
64. Xiang JP (2012) Active learning for person re-identification. In: 2012 International conference on machine learning and cybernetics, vol 1, pp 336–340. <https://doi.org/10.1109/ICMLC.2012.6358936>
65. Xiang ZJ, Chen Q, Liu Y (2014) Person re-identification by fuzzy space color histogram. *Multimed Tools Appl* 73(1):91–107. <https://doi.org/10.1007/s11042-012-1286-7>
66. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification
67. Yahiaoui T, Meurie C, Khoudour L, Cabestaing F (2008) A people counting system based on dense and close stereovision. In: Elmoataz A, Lezoray O, Nouboud F, Mamass D (eds) *Image and signal processing*. Springer, Berlin, pp 59–66
68. Ye W, Xu Y, Zhong Z (2007) Robust people counting in crowded environment. In: 2007 IEEE international conference on robotics and biomimetics (ROBIO), pp 1133–1137. <https://doi.org/10.1109/ROBIO.2007.4522323>
69. Yi D, Lei Z, Li SZ (2014) Deep metric learning for practical person re-identification
70. Zhang R, Lin L, Zhang R, Zuo W, Zhang L (2015) Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Trans Image Process* 24:4766–4779. <https://doi.org/10.1109/TIP.2015.2467315>
71. Zhang X, Yan J, Feng S, Lei Z, Yi D, Li SZ (2012) Water filling: unsupervised people counting via vertical Kinect sensor. In: 2012 IEEE 9th international conference on advanced video and signal-based surveillance, pp 215–220. <https://doi.org/10.1109/AVSS.2012.82>
72. Zhao T, Nevatia R, Wu B (2008) Segmentation and tracking of multiple humans in crowded environments. *IEEE Trans Pattern Anal Mach Intell* 30(7):1198–1211. <https://doi.org/10.1109/TPAMI.2007.70770>
73. Zheng WS, Gong S, Xiang T (2011) Person re-identification by probabilistic relative distance comparison