WatchNet: Efficient and Depth-based Network for People Detection in Video Surveillance Systems

M. Villamizar[*], A. Martínez-González[*†], O. Canévet[*]and J-M. Odobez[*†]

Abstract

We propose a deep-learning approach for people detection on depth imagery. The approach is designed to be deployed as an autonomous appliance for identifying people attacks and intrusion in video surveillance scenarios. To this end, we propose a fully-convolutional and sequential network, named WatchNet, that localizes people in depth images by predicting human body landmarks such as head and shoulders. We use a large synthetic dataset to train the network with abundant data and generate automatic annotations. Adaptation to real data is performed via fine tuning with real depth images.

The proposed method is validated in a novel and challenging database with about 29k top view images collected from several sequences including different people assaults. A comparative evaluation is given between our approach and other standard methods, showing remarkable detection results and efficiency. The network runs in 10 and 28 FPS using CPU and GPU, respectively.

1. Introduction

In recent years we have seen a large deployment of computer vision systems for people detection and counting in video surveillance and analysis applications [2, 4, 15, 16]. These systems can be of primary necessity for security in public and private places, *e.g*. banks, airports, and corporate buildings. Specifically for restricted areas where the access of people is monitored to prevent attacks and intruders.

In this paper, we study the problem of detecting intruders in airlocks from monocular depth cameras. More precisely, we focus on the detection of multiple people in restricted areas where one person is exclusively allowed at a time. This is a difficult problem since the video surveillance system must be able to detect people attacks and trickeries (such as tailgating and piggybacking to fool the system). See Figure 1 for an example.

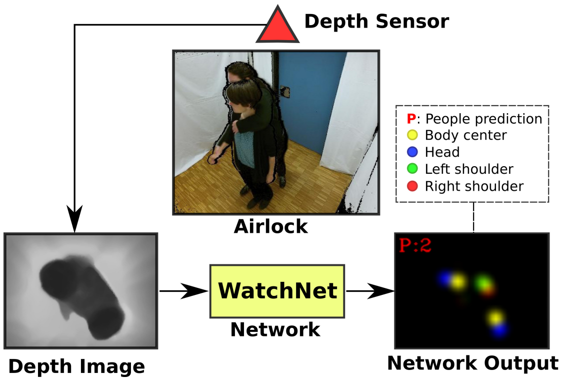

Figure 1. We propose an efficient depth-based network (called WatchNet) for people detection in video surveillance applications. The proposed method is able to identify people intrusion by detecting human landmarks (*i.e*. head and shoulders) with high accuracy.

The performance of people detection systems that rely on monocular cameras may degrade given the occlusions caused by scene elements or other people in the scene. Blind regions makes the task even more challenging since they can be potentially used to hide unallowed objects or people from the camera. To solve this problem, the camera is commonly placed in a zenithal position in such a way that it is much harder to deceive the detection system when people are exactly below the camera [5, 14, 18].

Another aspect related to surveillance systems is the privacy and data protection regulations imposed in many countries. Systems based on color cameras have to apply algorithms and controls to maintain people's privacy. This leads to the use of other technologies such as depth cameras that are a great source for people detection but when used alone can avoid this legal inconvenience [1, 6, 8].

Thus, in this work, we introduce a deep-learning approach for people detection from top view depth cameras (Figure 1). The approach is based on an efficient network, called WatchNet, that predicts the location of head and shoulders as well as the body center in order to estimate the number of people in the airlock. This network is trained with artificial and real samples to boost the performance on detection. This approach results in a robust and discriminative surveillance system able to detect people even under attack or intrusion situations.

[*] Idiap Research Institute, Switzerland. {michael.villamizar, angel.martinez, olivier.canevet, odobez}@idiap.ch

[†] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Related Work: To control the access of people to public or private buildings, video surveillance systems are usually based on counting the number of people in the scene. Techniques for this can be divided into two main categories: feature-based and counting-by-detection methods.

Feature-based counting methods formulate the task as a regression problem, avoiding people detection, where image features are exploited to predict the number of people in the scene. This approach is particularly convenient for crowded scenarios such as public events or demonstrations since the detection of people is very challenging due to the high degree of occlusion [2, 12]. The second category relies on visual detectors to localize each person in the image. To cope with occlusions, the detectors are mainly focused on localizing the head and shoulders of people [1, 6, 9, 15, 18]. This has shown good results especially for overhead and depth cameras.

Approaches can also be divided according to whether they are unsupervised or supervised. In both cases, a background segmentation technique is often exploited to facilitate the extraction of features and ease the detection process. People detection approaches based on unsupervised techniques have shown good efficiency in real world applications [1, 5, 8, 13, 16, 18], but their performance is compromised when people appear in static postures. In addition, the performance is heaviliy subject to the quality of background subtraction.

Supervised approaches to detect people have shown remarkable results. They normally require higher computational costs for both training and testing as well as a representative dataset with annotations for supervised learning. These approaches make use of machine learning algorithms, such as SVM or Boosting, to compute discriminative classifiers [14, 17, 19].

Recently, the use of deep networks has also shown impressive results for people detection using color cameras [3]. However, these methods have a high computational cost that makes them unfeasible for deployment in many real-world applications.

Contributions: In this paper we propose a counting-by-detection method for identifying attacks and intrusion in building entrances using a top view sensor. Specifically, we present a fully-convolutional and sequential network that detects people by predicting the location of head, shoulders, and the body centers (see Figure 1).

The proposed network, called WatchNet, is inspired by the Convolutional Pose Machines (CPM) for people pose estimation in color images [3]. However, WatchNet is a lightweight and efficient version of CPM thanks to the use of depth data rather than color, allowing to reduce the number of convolutional layers and parameters, and includes other network characteristics like skip connections useful for multi-resolution analysis. As a result, WatchNet is perfectly suited to video surveillance systems where real-time performance is a crucial requirement. Our network runs in 10 and 28 FPS using CPU and GPU cards, respectively.

To train the network with a large amount of data and reduce the human annotation effort, we make use of a synthetic dataset with its corresponding annotations generated automatically, observe Figure 2. This dataset has about 80k artificial depth images.

In contrast to earlier methods, WatchNet does not use any technique for background subtraction and temporal consistency. Yet, we consider that these techniques may be helpful to enhance our detection results, especially for those cases where people are not fully visible in the scene.

2. Method

In this section we describe the contributions of the paper: the proposed network (Section 2.2) and the synthetic database for training this network (Section 2.1).

2.1. Synthetic Dataset

The supervised learning of deep network models requires to have a large and diverse enough dataset to boost the network performance and prevent overfitting. Yet, the data is sometimes scarce for scenarios with task-based specifications. In addition, generating the images' annotations for supervised learning presents another inconvenient. This process is usually done manually and requires large amounts of human effort. An attractive alternative is to work with synthetic data. The benefits of this approach are twofold: 1) synthetic data can be generated according to a given scenario for a specific problem, and 2) high quality annotations are generated at no cost.

Thus, to overcome the need for annotated training data, we present a systematic way to generate artificial depth images displaying people inside an airlock and the corresponding annotations (ground truth). We introduce a Synthetic Data Generator (SDG) built on Blender[1] to render people performing multiple behaviors inside a virtual airlock by motion simulation (Figure 2). The airlock was designed following the specifications mentioned in the Unicity database [7]. Specifically, the airlock has an area of 2×2 meters and the camera is placed at the center of the airlock at two different heights: 2.1 and 2.5 meters.

A challenge in generating synthetic data is to introduce enough variability. We achieve this point by considering different body shapes and as many body pose configurations as possible. First, we use 24 3D human characters created with the modeling software Makehuman[2]. The different characters show variations in physical features, such as height and weight, and have been dressed with different clothing outfits to increase shape variation.

[1] http://www.blender.org

[2] http://www.makehuman.org/

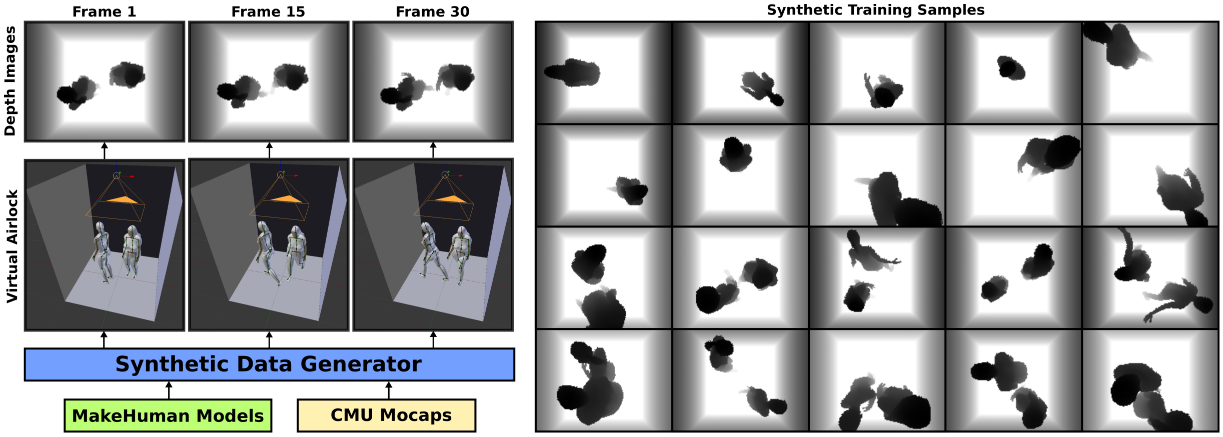

Figure 2. Synthetic Data Generator (SDG). Left: SDG creates a virtual airlock with one or two people performing different actions to reproduce similar depth images to the Unicity database [7]. Right: Some image examples generated by SDG for training WatchNet.

We add variability in body pose configurations by relying on the publicly available motion capture database from CMU labs[3]. We selected motion captures sequences of people performing diverse actions, such as walking or jumping.

To synthesize depth images along with the required annotations, our SDG works as follows. At each iteration, we randomly select up to two 3D characters along with the corresponding number of mocap sequences, randomly selected. The 3D characters are randomly placed inside the airlock, in such a way that there is no collision between them. Subsequently, SDG samples one every 15 frames from the mocap sequence, performs motion retargeting and generates the corresponding synthetic depth image along with annotations. This is illustrated in Figure 2 (left). As a result, the synthetic database has more than 80k images containing up to two people, observe Figure 2 (right).

2.2. WatchNet

In this section we describe our design choices for the architecture of WatchNet. It can be seen as a lightweight version of the Convolutional Pose Machines (CPM) network designed for human pose estimation [3]. Similar to CPM, the WatchNet network can be thought of as comprising a feature extraction sub-network and a series of prediction stages that progressively refine the localization of human body landmarks in the image. Figure 3 shows a general view of the proposed network architecture.

Feature Extraction sub-network: This sub-network computes discriminative features for body landmark prediction that will be shared among the prediction stages. Since we use depth images as input, the complexity of this stage can be reduced compared to [3], and we can therefore deploy a smaller and more efficient feature extractor sub-network, contrary to the original CPM framework that relied on a very deep network to compute features (VGG-19).

We propose to use a sub-network composed of 7 convolutional layers with filters' size of 3×3, three max-pooling operations, and one up-sampling operation (see Figure 3). Furthermore, all our convolutional layers use 64 filters to reduce the number of parameters in the network and speed up the forward pass.

A major design choice we follow and which differs from [3], is the use of skip connections [11] to combine features from different resolutions. Specifically, while the layer $C4$ with filters' size of 1×1 computes features at a quarter of the resolution of the input image, the convolutional layer $C5$ computes features at an eighth of the resolution. Then, features from $C5$ are upsampled and combined with $C4$ via concatenation. The output is fed to layers $C6$ and $C7$ to compute the features F for people prediction. See Figure 3.

The main reason for this configuration is to increase the robustness and accuracy of the network to detect people at multiple scales. This is an important aspect in video surveillance systems since the height at which the camera is located varies depending on the room, and depth measures have a different semantic nature than color images.

Prediction sub-networks: Every prediction stage is composed by four convolutional layers with filters of different sizes, keeping low the numbers of filters. The final layer provides prediction maps ($W/4 \times H/4 \times 4$) for three body landmarks and their center. In this work, we consider the head, left and right shoulder as body landmarks.

The first convolutional layer has filters of size of 5×5 in order to capture larger image spatial context and to encode the spatial relationships among the body landmarks. This spatial/feature co-occurrence has been shown to play an important role to refine the network output predictions [3].

General Settings: All our convolutional layers are computed in combination to batch normalization and Rectified Linear Units (ReLU), showing good experimental results and faster training.

[3] http://mocap.cs.cmu.edu/

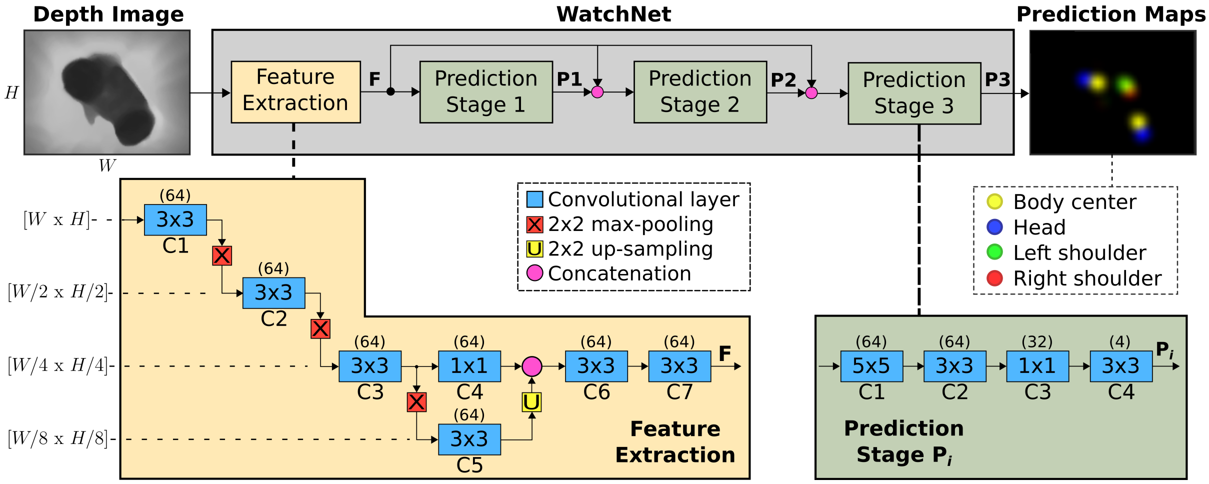

Figure 3. General scheme of the proposed network for people detection. WatchNet consists of a feature extraction module and a series of prediction stages that sequentially refine the prediction maps for human body landmarks (head and shoulders).

Training and Ground Truth: The training loss for WatchNet is calculated as a linear combination of partial losses across the network. We define the global loss by $L = \frac{1}{N}\sum_{i=1}^{N} L_i$, where N is the number of prediction stages and L_i is the loss for the prediction stage P_i. Specifically, the partial loss for a prediction stage i is defined as the mean squared distance between the prediction maps provided by stage i and the dataset's ground truth. Ground truth is computed by Gaussian blobs placed on annotated landmarks locations. We use Adam [10] as optimizer.

Counting People in Images: At test time, our learned WatchNet is applied to each image to compute predictions. We remove predictions whose confidence level is below a threshold β. The choice of β is done accordingly to the user needs (*e.g.* high recall vs high precision).

We use the number of predicted body centers to count the number of people inside the airlock. We observed this selection to be robust in cases when other landmarks lie outside the scene.

3. Experiments

This section evaluates WatchNet for the task of detecting attacks in building access rooms.

Real Dataset: For evaluation we use the Unicity database[4] introduced in [7]. It is composed of 65 recorded sequences of people passing through an airlock giving access to a restricted area. The sequences are organized according to three different scenarios: the first one is a normal scenario with a single person walking and accessing the restricted area; the second scenario comprises two people trying to fool the surveillance system (*e.g.* tailgating); in the third scenario, two people enter, and one of them attacks and forces the other to get into the restricted area (see Figure 1).

For training and evaluation, the dataset was split into 33 and 32 sequences, respectively. In total, the dataset has $29,045$ images (11,372 images for testing) acquired by an industrial depth sensor with a resolution of 120×160 pixels. All sequences were recorded at two different camera heights: 2.1 and 2.5 meters.

The Unicity dataset uses four levels for evaluation defined according to the degree of visibility of people in the airlock. So, for example, level 1 comprises all images where people's landmarks are full visible (head and shoulders). In level 2 at least one body landmark is visible (level 1 is thus a subset of level 2). Similarly, level 3 contains level 2 plus all those images where a portion of people is visible but not their landmarks. Finally, level 4 is all the images in the test set including hard cases (*e.g.* a leg is visible only). All levels also have empty room images which act also as negative samples (positive samples being attacks, i.e. images with two people) during evaluation [7].

Default Parameters: Unless otherwise stated, WatchNet is trained with the synthetic dataset for 50k iterations and is fine tuned with the real training data for 5k iterations. We use three prediction stages and a batch of five samples. To remove noise from depth maps, we resort to inpainting with a filter size of 5.

To select β, we run the network in the training set for varying detection thresholds and take the one that achieves the highest F-measure score, being the operation point with the best compromise between recall and precision.

Training Data: Table 1 shows the detection rates given by WatchNet according to the training data. We can see that the use of synthetic and real images (*i.e.* fine tuning) significantly improves the results, especially the recall rate that corresponds to the detection of attacks and intrusion.

Besides, the table reports the scores for the four evaluation levels mentioned above. Note that the proposed net-

[4] https://www.idiap.ch/dataset/unicity

	R	P	F	A	TP	TN	FN	FP	R	P	F	A	TP	TN	FN	FP
	Synthetic Data								Synthetic+Real Data							
Level 1	0.92	1.00	0.96	0.99	531	4098	49	0	0.99	1.00	1.00	1.00	576	4097	4	1
Level 2	0.83	0.98	0.90	0.95	1367	4867	274	34	0.96	1.00	0.98	0.99	1574	4894	67	7
Level 3	0.64	0.97	0.77	0.90	1512	6649	865	48	0.82	1.00	0.90	0.95	1953	6688	424	9
Level 4	0.48	0.97	0.64	0.85	1543	8083	1698	48	0.63	1.00	0.77	0.89	2050	8122	1191	9
Average	0.72	0.98	0.82	0.92	1238	5924	721	32	0.85	1.00	0.91	0.96	1538	5950	421	6

Table 1. Alarm detection rates provided by WatchNet in the Unicity database according to the training data: synthetic images or synthetic plus real images (*i.e.* fine tuning). The evaluation is done using the recall (R), precision (P), F-measure (F) and accuracy (A) rates, and the numbers of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP).

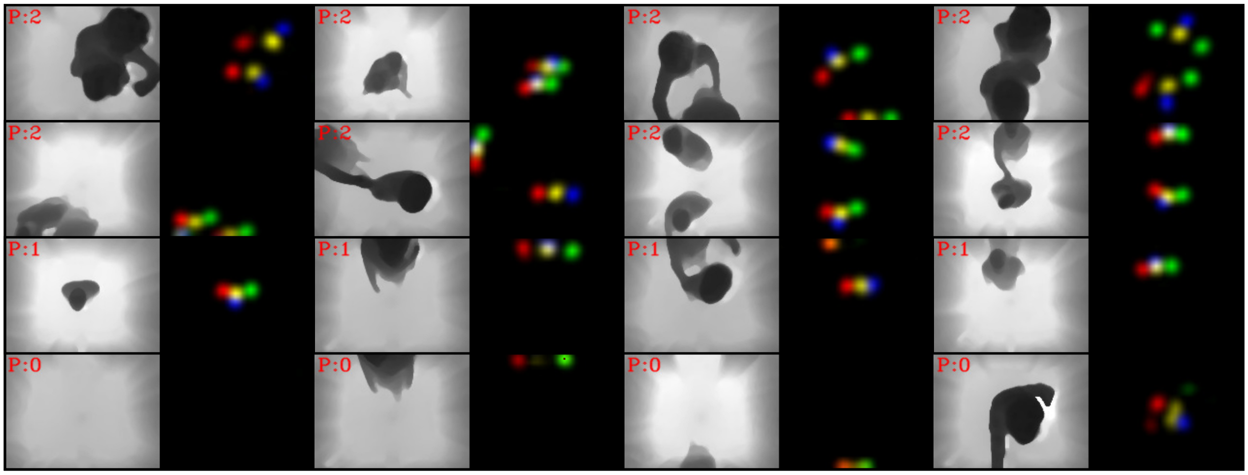

Figure 4. Some image examples with the output of the proposed system for people detection in depth images. The system predicts the location of head and shoulders as well as the body center, all depicted by blue, green, red and yellow spots respectively. The system also estimates the number of people (P) inside the airlock based on counting the number of body centers.

work achieves almost perfect rates for levels 1 and 2 which contain at least one visible body landmark. The scores degrade for levels 3 and 4 since people are not fully visible. Some examples are shown in Figure 4.

Detection Approaches: We compare WatchNet against other approaches in Table 2. The first approach is the baseline provided with the dataset [7]. It is a clever background subtraction method which thresholds the estimated volume inside the airlock: when the volume is larger than a predefined threshold, the method classifies the frame as an attack. The volume is estimated by simply summing up all the pixels of $B - I$, where B is the depth map of the empty airlock, and I the current image. The second approach is a Fully-Convolutional Network (FCN) consisting of 7 convolutional layers, two max-pooling operations and a single final layer for prediction. As a third approach we have the WatchNet without using the layers $C4$ and $C5$ (*i.e.* not multi-resolution features).

The baseline method achieves high precision and accuracy rates, but it does not provide the localization of people in the scene. On the other hand, FCN does detect people but obtains lower results than WatchNet because it does not include the proposed multi-resolution features scheme nor the refinement given by using several prediction stages. Similar case occurs with the third approach, proving again that the proposed network is more robust to detect people at multiple scales. Please compare the recall scores.

Prediction Stages: The detection scores in terms of the number of prediction stages is shown in Table 3. We found that with three stages the method attains the best rates.

Counting People: In Table 4, we evaluate the detection performance of WatchNet using different landmarks for counting the number of people: body center (default), head, and the combination of head and shoulders. Looking at the results we see that the body center attains better results than the landmarks because it is more robust to cases when the person is partially visible.

Use of Synthetic Data: To measure the benefit of using synthetic data, Table 5 shows the alarm detection scores of WatchNet using only real data for training. Table reports scores for different training iterations. For 10K iterations, the network achieves good rates, but they decrease along with the training iterations indicating overfitting. With synthetic data, we perform data augmentation and prevent overfitting, obtaining better results.

People Prediction: Figure 5 shows the confusion matrix for predicting the number of people in the scene. This was computed for evaluation level 1. We see that WatchNet performs remarkably well for people detection.

	Baseline				FCN				WatchNet [Not multi-scale]				WatchNet			
	R	P	F	A	R	P	F	A	R	P	F	A	R	P	F	A
Level 1	0.97	0.55	0.70	0.90	0.92	0.99	0.96	0.99	0.95	0.99	0.97	0.99	0.99	1.00	1.00	1.00
Level 2	0.96	0.74	0.84	0.91	0.87	0.98	0.92	0.96	0.89	0.99	0.94	0.97	0.96	1.00	0.98	0.99
Level 3	0.88	0.79	0.83	0.91	0.74	0.98	0.84	0.93	0.78	0.99	0.87	0.94	0.82	1.00	0.90	0.95
Level 4	0.72	0.81	0.76	0.87	0.56	0.98	0.71	0.87	0.59	0.99	0.74	0.88	0.63	1.00	0.77	0.89
Average	**0.88**	0.72	0.78	0.90	0.77	0.99	0.86	0.94	0.80	0.99	0.88	0.95	0.85	**1.00**	**0.91**	**0.96**

Table 2. Evaluation of WatchNet against a baseline method and other standard networks (FCN).

	1 Stage		3 Stages		5 Stages	
	F	A	F	A	F	A
Level 1	0.97	0.99	1.00	1.00	0.97	0.99
Level 2	0.95	0.98	0.98	0.99	0.96	0.98
Level 3	0.86	0.94	0.90	0.95	0.88	0.94
Level 4	0.72	0.87	0.77	0.89	0.74	0.88
Average	0.88	0.95	**0.91**	**0.96**	0.89	0.95

Table 3. Detection performance evaluation in terms of the number of prediction stages.

	Body Center		Head		Head & Shld	
	F	A	F	A	F	A
Level 1	1.00	1.00	1.00	1.00	0.93	0.98
Level 2	0.98	0.99	0.93	0.97	0.96	0.98
Level 3	0.90	0.95	0.81	0.92	0.89	0.95
Level 4	0.77	0.89	0.67	0.86	0.76	0.89
Average	**0.91**	**0.96**	0.85	0.94	0.88	0.95

Table 4. Alarm detection scores provided by WatchNet according to the body lardmarks used for counting people.

	10K		20K		50K	
	F	A	F	A	F	A
Level 1	0.96	0.99	0.97	0.99	0.96	0.99
Level 2	0.95	0.98	0.95	0.97	0.92	0.96
Level 3	0.90	0.95	0.89	0.95	0.86	0.93
Level 4	0.77	0.89	0.77	0.89	0.74	0.88
Average	**0.89**	**0.95**	**0.89**	**0.95**	0.87	0.94

Table 5. Detection rates of WatchNet trained with real data only.

4. Conclusion

In this paper we presented an access surveillance system based on deep learning to deal with security breaches such as tailgating and piggybacking and to detect attacks on people. Our system demonstrated very good results in a new database created for this problem. Our system is based on an efficient and robust network that sequentially locates parts of people such as head and shoulders in depth images.

Acknowledgments: The work was supported by Innosuisse, the Swiss innovation agency, through the UNICITY (3D scene understanding through machine learning to secure entrance zones) project. In addition, we would like to thank Joël Dumoulin and his colleagues from the HumanTech Institute of the HES-SO Fribourg, Switzerland, who dealt with the recording of the UNICITY dataset.

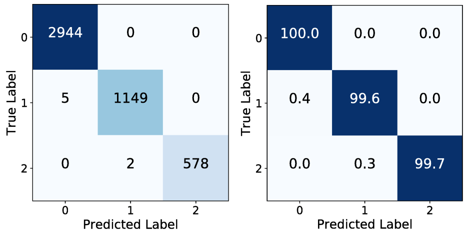

Figure 5. Raw and normalized confusion matrices provided by WatchNet for predicting the number of people in the airlock.

References

[1] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo. Realtime people counting from depth imagery of crowded environments. In *AVSS*, 2014.

[2] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, 2016.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] C. Carincotte, X. Naturel, M. Hick, J.-M. Odobez, J. Yao, A. Bastide, and B. Corbucci. Understanding metro station usage using closed circuit television cameras analysis. In *ITSC*, 2008.

[5] V. Carletti, L. Del Pizzo, G. Percannella, and M. Vento. An efficient and effective method for people detection from top-view depth cameras. In *AVSS*, 2017.

[6] S. Chen, F. Bremond, H. Nguyen, and H. Thomas. Exploring depth information for head detection with depth images. In *AVSS*, 2016.

[7] J. Dumoulin, O. Canévet, M. Villamizar, H. Nunes, E. Mugellini, F. Moscheni, and J.-M. Odobez. Unicity: A depth maps database for people detection in security airlocks. In *Research Report, HES-SO Fribourg*, 2018.

[8] F. Galčík and R. Gargalík. Real-time depth map based people counting. In *ACIVS*, 2013.

[9] R. Hu, R. Wang, S. Shan, and X. Chen. Robust head-shoulder detection using a two-stage cascade framework. In *ICPR*, 2014.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[12] Z. Ma and A. B. Chan. Crossing the line: Crowd counting by integer programming with local features. In *CVPR*, 2013.

[13] J. Nalepa, J. Szymanek, and M. Kawulok. Real-time people counting from depth images. In *International Conference: Beyond Databases, Architectures and Structures*, 2015.

[14] M. Rauter. Reliable human detection and tracking in top-view depth images. In *CVPRW*, 2013.

[15] H. Song, S. Sun, N. Akhtar, C. Zhang, J. Li, and A. Mian. Benchmark data and method for real-time people counting in cluttered scenes using depth sensors. *arXiv preprint arXiv:1804.04339*, 2018.

[16] J. Tu, C. Zhang, and P. Hao. Robust real-time attention-based head-shoulder detection for video surveillance. In *ICIP*, 2013.

[17] P. Vera, D. Zenteno, and J. Salas. Counting pedestrians in bidirectional scenarios using zenithal depth images. In *Mexican Conference on Pattern Recognition*, 2013.

[18] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li. Water filling: Unsupervised people counting via vertical kinect sensor. In *AVSS*, 2012.

[19] L. Zhu and K.-H. Wong. Human tracking and counting using the kinect range sensor based on adaboost and kalman filter. In *International Symposium on Visual Computing*, 2013.