UNICITY: A depth maps database for people detection in security airlocks

Joël Dumoulin[†], Olivier Canévet[‡], Michael Villamizar[‡], Hugo Nunes[∓], Omar Abou Khaled[†], Elena Mugellini[†], Fabrice Moscheni[∓], and Jean-Marc Odobez[‡]

[†]HumanTech Institute, HES-SO Fribourg, Switzerland, firstname.lastname@hes-so.ch
[‡]Idiap Research Institute, Martigny, Switzerland, firstname.lastname@idiap.ch
[∓]Fastcom Technology SA, Lausanne, Switzerland, name@fastcom-technology.com

Abstract

We introduce a new dataset, dubbed UNICITY[1], for the task of detecting people in security airlocks in top view depth images. If security companies have been relying on computer systems and algorithms for a long time, very few are trusting artificial intelligence and more specifically machine learning approaches in production environments. We are confident that the recent advances in these domains, especially with the democratization of deep learning, will open new horizons for security systems. We release this dataset to encourage the development of such approaches in the scientific community.

*UNICITY consists of 58k images collected from 65 recorded sequences with one or two people performing different behaviors including attacks and trickeries (*e.g. *tailgating[2]). It also provides full annotation of people such as the location of head and shoulders. As as result, UNICITY is perfectly suited for training and adapting machine learning algorithms for video surveillance applications. This paper presents the data collection, an evaluation protocol, as well as two baseline methods for attack detection.*

1. Introduction

In this paper, we focus on people detection inside security airlocks. An airlock is a space separating a restricted area from a non-restricted area, in which a person needs to authenticate to be granted access to the restricted area. For a company active in this field, the need for benchmark data is growing as more and more features and flexibility are expected by customers, but such data is lacking. Expert systems are usually sufficient for standard use cases like anti-tailgating solutions as long as the conditions are controlled (*e.g.* specific room size, specific sensor-person relation, minimal lighting variations). But more advanced approaches taking advantage of artificial intelligence are needed to bring more flexibility and cope with more complicated use cases, like for instance granting access to a secure zone only if certain number of authorized persons are present in a delimited sub-part of the monitored zone.

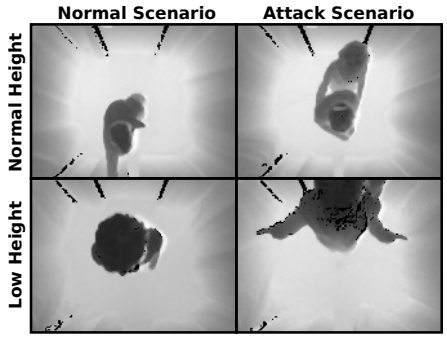

Figure 1: The UNICITY dataset (Bluetechnix Argos).

To meet the demand of these even more challenging purposes, we introduce a new publicly available dataset of depth maps (see Figure 1). For building this dataset, named UNICITY, we recorded 3.4 millions of images using 6 color and depth cameras (4 Kinects and 2 industrial cameras) located at different positions inside an airlock constructed to simulate a building access room.

The UNICITY database is a selected part of the full data collection in order to focus in the problem of detecting people from top view images acquired by a depth and industrial sensor. These particular choices are motivated by *(1)* data protection regulations to maintain people's privacy (use of depth data), *(2)* increase people visibility (camera in zenithal position to cope with occlusions), and *(3)* robustness (industrial sensors for daily applications).

Specifically, UNICITY has three main contributions: the first one is the recorded data that has about 58k images

[1]The dataset is released on https://www.idiap.ch/dataset/unicity
[2]When a person walks very close to another to get into a restricted area.

from two top view industry-oriented cameras (Bluetechnix Argos3D-P220 and Fotonic G-series) at two different airlock heights. It comprises 65 sequences recorded with or two persons for two given scenarios: one person entering inside the airlock to swipe their badge to be granted access (normal scenario), and two persons entering the airlock, the second either tailgating or attacking the first one to fool the system and enter to the restricted area (attack scenario).

The second contribution are the ground truth annotations of all persons in the 58k images. Each annotation was manually done and includes the location of head and shoulders in the image and the degree of visibility of persons to distinguish between easy and difficult cases.

Lastly, the third contribution is that the dataset comes along with an evaluation code which computes performance criteria, to allow fair comparison between different users benchmarking their methods on this dataset.

We present the related work in section 2 while section 3 describes the data acquisition and overcame challenges. Section 4 introduces the UNICITY dataset and its main features. Finally, two baseline methods and their performance for detecting attacks are shown in sections 5 and 6.

2. Related work

The dataset targets a security application: the verification of the presence of a single person in an airlock using depth data. As such, it indirectly relates to two main topics: person counting, and depth processing for similar purpose (person detection and tracking). Below we review previous works in these topics and present existing datasets.

People Counting. People counting can be considered as a straightforward extension of person detection, but due to its applications in video analytics for surveillance, it has been a subject of important research on its own [8, 13, 2, 5, 10, 1]. The main use case is concerned with occupancy and flux analysis in open spaces, but other applications like anti-tailgating in metro turnstiles, presence verification or intrusion detection have been considered as well. As such, it has a multitude of challenges: variability in illumination conditions, viewpoints, person shapes, and depending, on scenarios, crowd density and proximity between people, *etc*. As an early work, Liu *et al*. [8] proposed a segmentation algorithm to improve robustness of people counting systems using one or more cameras. Histograms of oriented gradients combined with local binary patterns have shown good robustness to partial occlusion by detecting head and shoulders of people in real time [13], but this technique only allows an estimation of the number of people, not a precise count.

The fusion of different data sources is another approach studied in the field of surveillance. As a representative case, Schreiber *et al*. [10] used a trinocular configuration based on two monochrome cameras (stereo configuration) for analyzing the depth information as well as a color camera placed in the middle to count the number of people passing a door or ensure that only one person is present in a room. The method has demonstrated good performance in real environments (*e.g*. airports).

Depth maps for counting and tracking. In recent years, real-time depth cameras have brought new opportunities, including in surveillance, and have been prefered over color cameras for counting people [5]. Approaches generally consists on foreground/background segmentation to segment persons before using depth information to localize head candidates, as in Bondi *et al*. [1], allowing real-time counting in crowded environments. In [9], a method to track people from top view depth maps is proposed. The head detection step relies on the extraction of viewpoint specific features that are classified with a SVM framework in order to estimate the location of head in the image. Interestingly, a 93% true positive rate for a 99% true negative rate performance is reported for the head classification step.

Datasets. Li *et al*. [6] released a public dataset named CASIA Pedestrian Counting Dataset. They recorded over 1 TB of both RGB images and videos from more than ten cameras, capturing scenes through several seasons and weather conditions and with various crowd densities. While the images are meant to be used for training and testing pedestrian detectors, the videos can be used to evaluate pedestrian counting systems. Liciotti *et al*. [7] created a dataset for person re-identification with a RGB and depth camera in top view configuration and fixed camera height. This dataset comprises 23 video sequences (around 60k images) where 100 people pass under the camera wearing different seasonal outfits. Recordings were made in indoors and illumination conditions were not controlled, thus not constant. Del Pizzo *et al*. [3] provide a top view dataset for people counting using a RGB camera and a Kinect depth sensor. It includes 17 sequences recorded in indoors and outdoors with a variable number of persons (from one and up to four persons) walking in the same and/or opposite directions. Yet, this dataset is weakly annotated including only the number of people crossing a virtual line in the scene. Therefore, it cannot be used neither for benchmarking nor for training machine learning approaches for localizing people and body parts (*e.g*. head and shoulders). Moreover, the depth maps are only coded on 8 bits out of the 11 available.

Compared to these datasets, the UNICITY dataset consists of the original depth data from two industry-oriented sensors placed at two different airlock heights. UNICITY has 65 sequences (about 58k images) where 26 participants, wearing diverse outfits, performed attack and standard scenarios in the airlock (Figure 1). Full annotation of people along with evaluation code are also provided. This dataset is then appropriated for training relatively small approaches from scratch, or for adapting (*e.g*. fine tuning) more advanced methods such as deep networks for people detection.

Feature	Kinect	Argos	Fotonic
Framerate (fps)	30	40	40
Depth camera x	512×424	160×120	640×480
Depth range (m)	0.5 – 4.5	3.5	0.15 – 5
Field of view (h-v)	70° – 60°	90°	80° – 64°
RGB camera	1920×1080	–	–
Protection	No	IP65	IP65, IP67

Table 1: Sensors features.

3. Data acquisition

In this section we describe the process of data acquisition. We describe the sensors used, the recording architecture, the considered scenarios, and the challenges faced.

3.1. Sensors and data formats

Three types of sensors have been used for the data acquisition process: *(1)* Microsoft Kinect for Windows v2, *(2)* Bluetechnix Argos3D-P220 and *(3)* Fotonic G-series. We will refer to them as Kinect, Argos and Fotonic in the rest of this paper. The features of the sensors are detailed in table 1. Thanks to its very good value for money, the Kinect is commonly used for academic research purposes, but it is not an adequate choice in industrial applications, due to its design: lack of robustness to various lighting or low/high temperatures, no International Protection Marking, and the robustness of the plastic construction is questionable in the long term. Moreover, it has been announced that the Kinect would not be manufactured anymore. For this reason, the Argos and the Fotonic – which are two industry-oriented sensors – have been used in addition.

Four kinds of data have been recorded: depth, infrared, HD color and registered color. The registered image means that each pixel of the color image is corresponding to a pixel of the depth sensor. Depth and infrared data are OpenCV matrices saved as binary files. HD color data (1920×1080) and color data (512 × 424) are saved as JPEG files. For Argos and Fotonic sensors, only depth data is recorded. For Kinect sensors, the four kinds of data have been recorded.

3.2. Recording architecture

A dedicated structure, see Figure 2, was manufactured. The structure was made adjustable to ease data recordings and allow different base configurations and heights. On the top, there are three sections where the sensors can be screwed to a mounting bracket: a fixed central section and two sliding sections. To reduce clutter (recording devices, room furniture, *etc*.) as well as to limit the ambient infrared noise (outdoor lighting), cloth was added all around the recording structure.

Physical settings. The position of each sensor is schematized in Figure 2 representing a top view of the recording structure. One Argos and one Fotonic are placed in the middle of the structure, next to each other, facing down. Four

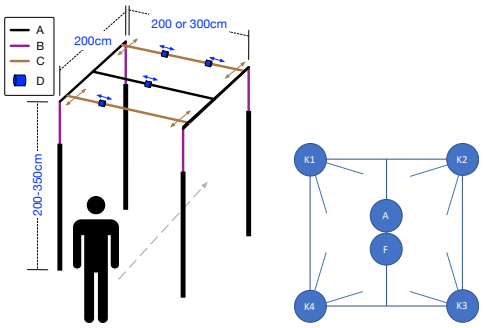

Figure 2: Schema of the recording structure (left). A: Fixed sections, B: Graduated telescopic sections, C: Sliding sections, D: Sliding mounting brackets. Top view of the recording structure (right). K1-K4: Kinects, A: Argos, F: Fotonic.

Kinects are placed on each corners, with an angle allowing to cover a maximum of the airlock. The physical structure configuration used is a 200 cm square, with two distinct heights: normal height (250 cm) and low height (210 cm).

Recording hardware. Four recent laptops (with Intel Core i7-6700HQ and 16GB RAM) were used as recording devices. Both the Argos and the Fotonic sensors are connected to them through Ethernet, while the Kinects are connected through USB3. One recording device has been used to simultaneously record one Kinect and the Argos, one is in charge of one Kinect and the Fotonic, and the last two recording devices are connected to one Kinect each.

Softwares. Several softwares have been specifically developed to ease the whole process, constituting a complete acquisition system. The recording configurations (participants, scenario, *etc*.) are edited and saved in database with a Django application. An Angular web application allows to control (start and stop) the recording of all sensors at the same time by communicating with the recording devices through sockets. The whole system is running inside a Docker container, thus facilitating its deployment.

3.3. Scenarios

During the recording process, the participants were instructed to perform different behaviors, but following three scenarios established to build the dataset. The scenarios are:

• ***Normal:*** one person enters the room, crosses and exits. The person can act in a natural way, but he can also adopt a strange behavior.

• ***Standard attack:*** this is the most common attack in airlock security. One assailant is trying to force an authorized person to let him enter the secure zone. In this scenario, a first person enters the room, then a second person forces the first one to let him enter as well. Finally both persons enter the secure zone.

• ***Tailgating:*** tailgating is a fairly common airlock security attack. Two persons enter, one following the other very closely to fool the system into detecting only one person.

3.4. Challenges

Several challenges have been encountered, a major one being the synchronized recording. Indeed, six sensors connected to four different recording devices have to be controlled so the saved frames are synchronized. Even if all sensors could be controlled to start the recording at the exact same time, they can all have a variable framerate: even if the framerate is known, for example 40 fps for the Argos, there can be small variations, mainly due to the load of the recording device. For this reason, it is not possible to only rely on frame indexes, but it is also needed to know the exact time when a frame has been recorded. Our solution is to retrieve the UNIX timestamp when the frame is grabbed, and store it directly in the frame filename. In order to have synchronized recording devices clocks, the NTP protocol[3] has been used. Due to the variable framerate of the sensors, the delays introduced by the messaging protocol to control the start and the end of the recording of each sensor – between 5 ms to 40 ms – has been considered acceptable as it introduces an offset of one frame only in the worst case.

Another challenge has been finding an efficient way to write the frames to disk. At first, writing a single frame was taking more time than the delay between two consecutive available frames. We came up with a solution based on an asynchronous queue and a dedicated thread for handling this problem. This way, we can continue to grab frames without having to wait until the previous frame has been written. A second problem was the size of the files. At the beginning, we were saving depth maps in YAML format, but it was taking too much space. We ended up saving them as OpenCV Mat structures in binary files. Regarding the RGB images, we first save them as Bitmap files, then convert them to JPEG in batch once the recording is finished, thus reducing the write time.

4. The UNICITY dataset

This section describes the UNICITY dataset that we release[1]. Out of the 3,409,137 frames recorded from the 6 cameras, we extracted a relevant part which was fully annotated and allows studying people counting and attack detection under different conditions.

4.1. Data

Some statistics of the data we release are provided in table 2. In essence, the data are 65 video sequences collected with the Argos and Fotonic sensors (top view, see Figure 2 for their respective location) under the following conditions:
• ***Normal scenarios:*** people enter the airlock, stop in the middle, swipe their badge, and leave the airlock.
• ***Attack scenarios:*** another person enters the airlock, either to force the first one to let him/her enter, or by tailgating the first one to fool the system.

[3]https://help.ubuntu.com/lts/serverguide/NTP.html

# Frames	# Sequences	# Participants
58,404	65	26
# Heads	# Left Shoulders	# Right Shoulders
41,292	39,687	39,445

Table 2: UNICITY dataset features.

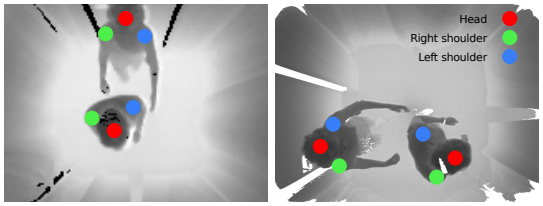

Figure 3: Synchronized views of Argos (left) and Fotonic (right) sensors (rotation of 90° due to physical constraints).

4.2. Dataset annotation

We manually annotated all the frames of the UNICITY dataset with several elements. This process took around 50 hours for the 58k frames of both sensors.

Body landmarks. Figure 3 shows two synchronized Argos and Fotonic frames in an attack scenario, with the body landmark annotation: head, right and left shoulders. When one of the limbs is occluded (*i.e.* not visible), the annotation is not present.

Number of people, attack, and visibility labels. In addition to the location of the three body landmarks, we have also annotated for each frame the number of people in the airlock. As a by-product, when two persons are present, the frame corresponds to an attack.

Given the airlock size and camera field of view, detecting attacks may depend on the visibility of the person(s). Hence, we have also annotated the degree of visibility for each person in the airlock using four visibility tags:
• ***Full:*** the person is fully visible (landmarks are visible).
• ***Partial:*** the person is partially visible and at least one landmark (head or shoulder) is visible.
• ***Truncated:*** a large portion of the person is visible but not any landmarks (*e.g.* lower body).
• ***Difficult:*** similar to the truncated label but it only applies for a small portion of the person (*e.g.* a leg or a hand).
• ***Invisible:*** the person is not visible in the airlock.

These visibility tags allow to measure the performance of the security system for different levels of difficulty (refer to section 4.3).

4.3. Task and evaluation protocol

This section describes the evaluation protocol and criteria for the task of detecting attacks in airlocks. Importantly, note that an evaluation code is released alongside the dataset for fair comparison with future publications on this dataset. Also, note that other tasks could be considered for evaluation as well, like the detection of body landmarks, body orientation estimation, or explicit people counting.

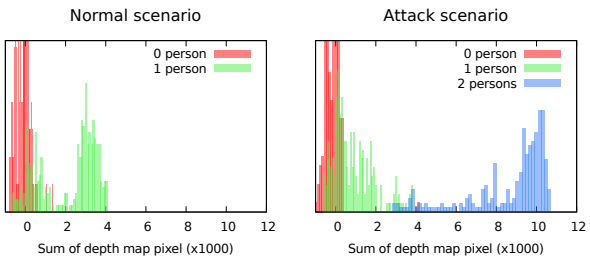

Figure 4: Distribution of the sum of depth-map pixels (*i.e.* volume inside airlock). A simple threshold on this value can discriminate between attack (*i.e.* more that one person), and non-attack (*i.e.* less than one).

Evaluation difficulty levels. For evaluation, we have proposed four different levels of difficulty according to the degree of visibility of people in the airlock. Note that it is a common practice [4, 12] to perform the evaluation on different subsets of the dataset, and be able to analyse how the methods perform on "easy" and "hard" cases.

Specifically, the level 1 only has frames with $full$ visibility tag. Level 2 includes level 1 and frames with $partial$ visibility tag. Likewise, level 3 includes level 2 and frames with $truncated$ visibility tag. Finally, level 4 has all frames in the dataset. All levels also have frames with $invisible$ tag which act as negative samples during evaluation.

Performance criteria. Standard ones were used. As we are interested in attack detection, a true positive (TP) is correctly detecting an attack (*i.e.* more than one person in the airlock), a true negative (TN) is correctly detecting one person or an empty airlock. We note FP for false positives, and FN for false negatives. Using these values, we compute four metrics for evaluation and comparison: Recall (R), Precision (P), F-measure (F) and Accuracy (A)[4].

5. Baseline Methods

This section presents two baseline methods to detect attacks in an airlock of a given size. The first is a simple method based on the estimated volume of the person(s) inside the airlock. The second method uses a deep convolutional network to detect and count people through the localization of body landmarks [11].

Volume-based Method. The depth sensors provide a depth map: a pixel value represents how far (in millimeters) the object is from the sensor. If we call B the airlock background (*i.e.* when the airlock is empty), and I a depth map, then $B - I$ represents how high the objects are from the ground, and $\sum(B - I)$, the sum of all the depth map pixels, is proportional to the "volume" of the object in the airlock.

Figure 4 depicts the histograms of the "volume" inside the airlock across the frames of a given sequence. The left figure corresponds to a normal scenario, with only one person in the airlock, and the right one to an attack (*i.e.* two persons in the airlock). The red histograms correspond to the case where the airlock is empty; the green ones to the case where there is one person, and the blue one to the case where there are two persons inside. Obviously, the volume inside the airlock is larger with one person inside than when empty; and larger with two than with one.

Based on this observation, we introduce a simple attack detector f (*i.e.* an alarm is raised when more than one person are in the airlock) with the following rule:

$$\begin{cases} 1 & \text{if } \sum(B - I) > \tau \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where τ is a threshold selected to meet a specific precision or recall score.

Network-based Method. As second baseline, we use the network proposed in [11] for people detection. This network, named WatchNet, comprises a feature extraction subnetwork and a series of prediction stages that progressively refine the localization of human body landmarks in the image. Particularly, the network predicts the head, left and right shoulders as body landmarks and estimates from them the body centers to count the number of people inside the airlock. For further details about the network architecture and its training procedure using both synthetic and real depth data, please refer to [11].

6. Experiments

Settings. To evaluate and compute both baseline methods, the UNICITY dataset, consisting of 65 recorded sequences, was split in two parts: one for training, and the other for testing. The training set has 33 recordings while the test set has 32 recordings. The split was done such that a given participant does not appear in both training and test sets. This was done to prevent the algorithms to overfit on some given shape, corpulence, or height.

The parameters of both methods were computed on the training set: the threshold τ for the volume-based method and the network weights for WatchNet. The network was trained for 50k iterations using synthetic data and for 5k iterations with real data (*i.e.* training set) for fine tuning [11].

Detection Results. Figure 5 shows the detection performance (via ROC curves) of the volume-based method on the UNICITY test set and for the Argos data. Solid lines are for the scenarios with low height whereas dash lines are for the normal height. We show one curve for each level of difficulty. We see that this method works very well on low height, and perfectly in the case of level 1, and there is still room for improvement for the normal height.

The discrepancy between the two can be accounted for by the fact that the recording structure (Figure 2) was moved from low to height several times, and that it has not been put back in the exact same position between recordings. The

[4] $R = TP/(TP + FN)$ $P = TP/(TP + FP)$
$F = 2 \times R \times P/(R+P)$ $A = (TP+TN)/(TN+TP+FN+FP)$

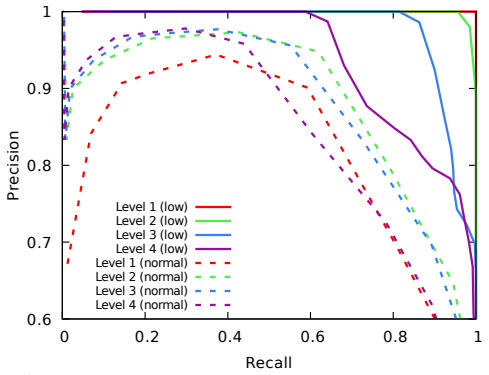

Figure 5: Detection performance of the volume-based baseline method in terms of the level of difficulty and the airlock height.

Difficulty	R	P	F	A	R	P	F	A
	Volume-based				Network-based			
Level 1	0.97	0.55	0.70	0.90	0.99	1.00	1.00	1.00
Level 2	0.96	0.74	0.84	0.91	0.96	1.00	0.98	0.99
Level 3	0.88	0.79	0.83	0.91	0.82	1.00	0.90	0.95
Level 4	0.72	0.81	0.76	0.87	0.63	1.00	0.77	0.89
Average	0.88	0.72	0.78	0.90	0.85	1.00	0.91	0.96

Table 3: Detection rates for both baseline methods.

	R	P	F	A	R	P	F	A
	Without UNICITY				With UNICITY			
Average	0.72	0.98	0.82	0.92	0.85	1.00	0.91	0.96

Table 4: Detection rates for the network-based method according to the use of the UNICITY dataset.

background image (empty background) may thus be different. Moreover, this method does not take into account the corpulence of people and share the same threshold across all recordings of the same height.

Table 3 shows quantitative results of the two baseline methods for the different levels of difficulty and for the Argos data[5]. These rates are computed for all scenarios and airlock heights using the evaluation code. We see the superior performance of the network-based method since it is a more stringent method that focuses on learning and detecting body-part patterns instead of performing simple foreground/background segmentation. As a consequence, WatchNet is a more robust detector with much lower numbers of false positives (see the precision rates).

Finally, table 4 reports the average detection rates of WatchNet according to whether the UNICITY dataset is used to train the network or not. Left side of the table shows the scores without using the dataset. The network is trained using artificial data only, refer to [11]. Right side shows the rates after fine tuning the network with the UNICITY dataset (real data). Note that the use of the proposed dataset improves the detection rates of the network, showing that it allows to train and adapt complex machine learning models for video surveillance applications.

[5]Results for the Fotonic data will be available at the dataset website.

7. Conclusion

The UNICITY dataset[1] is a collection of depth map images taken from industry-oriented cameras, Bluetechnix Argos3D-P220 and Fotonic G-series, that is introduced to push the state-of-the-art further for the task of detecting people in security airlocks. An evaluation code is provided alongside the data for fair comparison with all future methods benchmarked on this dataset.

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.

References

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

[2] 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.

[3] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento. A versatile and effective method for counting people on either rgb or depth overhead cameras. In *ICMEW*, 2015.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. In *PAMI*, 2012.

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

[6] J. Li, L. Huang, and C. Liu. Robust people counting in video surveillance: Dataset and system. In *AVSS*, 2011.

[7] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti. Person re-identification dataset with rgb-d camera in a top-view configuration. In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. 2016.

[8] X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever. Detecting and counting people in surveillance applications. In *AVSS*, 2005.

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

[10] D. Schreiber, A. Kriechbaum, and M. Rauter. A multisensor surveillance system for automated border control (egate). In *AVSS*, 2013.

[11] M. Villamizar, A. Martinez, O. Canevet, and J.-M. Odobez. Watchnet: Efficient and depth-based network for people detection in video surveillance systems. In *AVSS*, 2018.

[12] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, 2016.

[13] C. Zeng and H. Ma. Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *ICPR*, 2010.