Application of Deep Learning Techniques to Depth Images for Person Tracking and Detection

Velton Cardoso Pires[1], Eduardo Silva Palmeira[2], Felipe Antunes dos Santos[3]

[1] Universidade Federal do Rio Grande do Norte vellpires@gmail.com
[2] Universidade Estadual de Santa Cruz espalmeira@uesc.br
[3] Public University of Navarre fante.antunes@outlook.com

Abstract. Nowadays, using neural networks for image processing and tracking of individuals/objects is a highly popular subject that can be applied to various real-world issues. However, for such cases, the image often needs to possess good quality and exhibit distinct features to aid in object detection, posing challenges in environments with low or no illumination. In our work, we present a comparative study on the performance of leading convolutional neural networks for the detection and tracking of individuals through depth color images generated by infrared sensors. Additionally, we aim to demonstrate the usability of the YOLO (You Only Look Once) architecture as an alternative for identifying objects in images generated by sensors that do not rely on illumination. Experimental results showcase that the approach using YOLO Tiny improves accuracy by approximately 9% and processes around 8 times more frames per second (FPS).

Keywords: Person Detectiion, Neural Networks, Image Processing

1 Introduction

In the current world, with greater availability and accessibility to video surveillance camera systems in real-time and given the large amount of data generated by this monitoring, there is a need for the use of technologies capable of managing this information and managing the data, to instruct and assist human beings in decision-making processes. Especially speaking of environments with a large flow of people, these technologies serve, in addition to other purposes, to monitor the behavior of people who circulate in a given place.

In this context, computer vision techniques for processing the information generated from camera systems has raised as a very important alternative to solve many problems. Strategies for extracting characteristics and detecting patterns are widely studied by several researchers but in more different areas of knowledge (see [1]). We highlight the use of Convolutional Neural Networks (CNN) for the treatment and classification of digital images.

Here we would like to point out YOLO and RESNET convolutional neural networks. The first one is very known in the literature and is useful for object detection in real time. The other one is very efficient in handling more sophisticated deep-learning tasks and models.

The main goal of this paper is to propose a comparison study about the performance of different CNN for person detection and tracking from depth color images gotten from a specific device. Images were get from a public space and a dataset with 1000 images was created for the training step.

Section II is devoted to recalling the concept of convolutional neural networks and image processing. Section III brings a deep search on some important related works. The methodology is presented in Section IV and experimental results are shown in Section V. In the final section, a comprehensive conclusion of the work is provided, outlining challenges, results, and ideas for future work.

2 THEORETICAL FOUNDATION

In recent years, with the continuous improvement of the computational power of the hardware, the deep learning algorithms have developed rapidly which have enabled better performances in the field of computer vision. Person and object detection algorithms are mainly based on the extraction of features and shapes and its representative algorithms include a Combination of Histogram of Oriented Gradients (HOG) [2] and Support Vector Machine (SVM). However, these algorithms have some deficiencies, such as the large volume of computational work, speed of data processing, the inability to adapt to the data, such as rotation, and difficulties in dealing with occlusion, limiting their applicability.

Advances in deep learning in computer vision applications (in particular, convolutional neural networks) in the accuracy of classification and object recognition have achieved an impressive improvement. The evolution of Graphic Processing Units (GPUs) also contributed significantly to the adoption of CNN in computer vision.

There are two main types of algorithms for image detection: The first one is a two-step algorithm where firstly it generates a series of sample regions of the possible candidates and then performs the detection, through the convolutional neural network. These algorithms have high precision however it requires a high processing cost making them difficult to execute in real-time (see R-CNN [3], Mask R-CNN [4], Faster-RCNN [5], Fast-RCNN [6]); the second type is the only one step algorithms which are based on regression methods and the detection problem is treated as a regression problem. These kinds of algorithms directly predict categories and locate objects without needing to generate regions to identify trained objects. This type of algorithm has better efficiency than two-step algorithms, however, its accuracy tends to be worse. Examples of these algorithms are YOLO [7], YOLOv2 [8], YOLOv3 [9], SSD [10].

In which follows in this section it is discussed two particular CNN are very important for the development of this work.

2.1 RESNET

Residual network (ResNet) was proposed by [11] and it is based on the VGG-Net [12] network style, with 152 layers architecture which is 8 times deeper than

VGG. This new network has become more efficient in different image and video processing tasks, including image classification, object detection, segmentation, location, and so on. Using this new structure [11] managed to train a network with up to 1001 layers, which was practically impossible before, exhibiting excellent performance in computer vision.

The ImageNet is a dataset with over 14 million images and more than 21 thousand classes. In addition to this dataset, the challenge of image classification and object detection was proposed, which is named Large Scale Visual Recognition Challenge[4] (ILSVRC). Some of the products of said challenge are the ability to evaluate and compare the progress of detection algorithms and computer vision [13].

Regarding the ILSVRC, one of the main metrics to evaluate the performance of the algorithms is the classification error percentage, of which in 2012 was the first time a CNN won, with the AlexNet, followed and surpassed by ZFNet in 2013, VGG and GoogLeNet in 2014 and ResNet in 2015. The ResNet CNN was the first winner at this challenge to outperform the human classification error for the dataset [13], [12], [14].

In neural networks in general, when the NN goes past a certain complexity, with an excessive number of layers for example, the classification begins to degrade instead of improving, given that the model might be too complex for the problem, decreasing generalization capabilities. In this thinking, the ResNet presents a deeper architecture than its previous ILSVRC winning counterparts, with the idea that a CNN more complex than the shallower fine-tuned one for the problem should not produce inferior results than a deeper one and, for that end, the identity mapping and residual function and connection are applied, in which skip connections are created, as seen in Figure 1[5] [12], [14], which illustrates how a residual block operates, we have a vector of values, and we apply a Rectified Linear Unit (RELU) function, where for any input value x, the output of the RELU function is the input value itself if it's positive, and 0 if it's negative. In other words, the RELU function "activates" positive outputs, leaving them unchanged, while nullifying negative outputs.

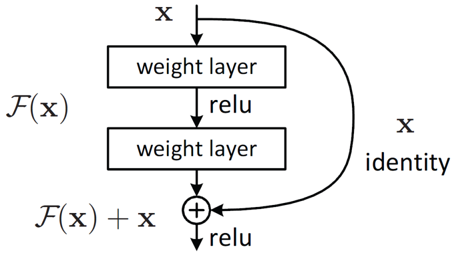

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

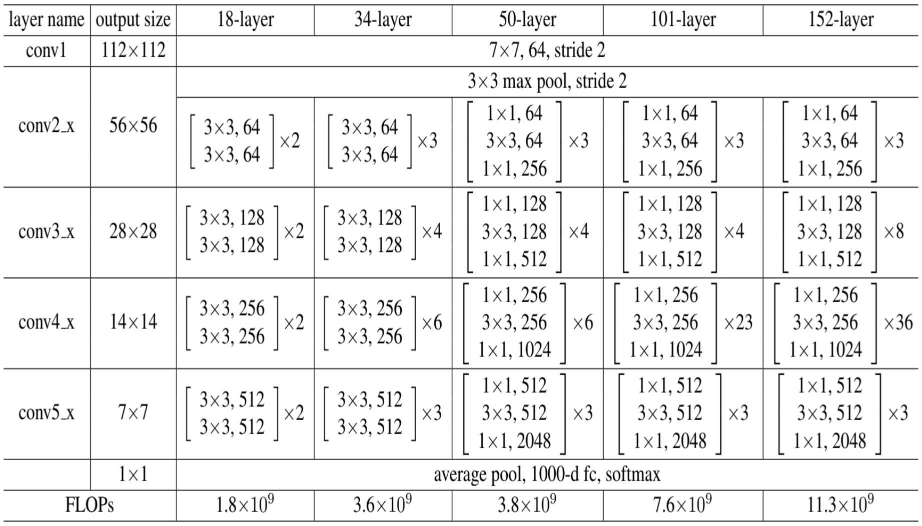
Fig. 1. ResNet residual block

[4] Challenge site: http://www.image-net.org/challenges/LSVRC/

[5] Image source at https://neurohive.io/en/popular-networks/resnet/

Regarding the ResNet, this name stands for Residual Network and was proposed by [11], based on the VGG-Net network style, with 152 layers of architecture which is 8 times deeper than VGG. This new network has become more efficient in different image and video processing tasks, including image classification, object detection, segmentation, location, and so on. Using this new structure, [11] managed to train a network with up to 1001 layers, which was practically impossible before, exhibiting excellent performance in computer vision [12].

From a technical view, most convolutional layers have 3×3 filters and share two main characteristics: (1) for each output with equal map sizes, the layers have the same number of filters; and (2) if the characteristics of the map size are halved, the number of filters is doubled to preserve complexity time per layer. The ResNet network (Fig 2) has fewer filters and less complexity than the VGG model [12], there is a summary of the output size at every layer and the dimension of the convolutional kernels at every point in the structure.



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Fig. 2. ResNet architectures [12]

2.2 YOLO and TINY YOLO

YOLO [7] is a one-stage detection network that works for detects objects, dividing an image into a network unit. The characteristic map of the output layer of the YOLO network is designed to return the box coordinates in which it returned the detection points that delimit the rectangle, the class number, and its accuracy. Thus it allows the detection of several objects with a single inference.

YOLO has 24 convolutional layers followed by 2 fully bonded layers. Some convolutional layers use size 11 convolutions to reduce the dimension of depth feature maps. Therefore, the detection speed is much faster than conventional methods such as HOG. However, due to the processing of the network layers, location errors can occur and the accuracy can be low, making it unsuitable for detecting people in real-time. To solve these problems, the YOLOv2 [8] network has been proposed, improving detection accuracy, if compared to the YOLO network, using batch normalization for convolution, multi-scale training, and anchor box layers. However, the detection accuracy of objects is still low for small and dense objects. Another characteristic of the YOLOv2 network is the use of dimension clusters. Using K-means Clustering in the training set to automatically find the best regions for detection.

To circumvent the difficulties of the YOLOv2 network, the v3 version of this network was proposed by [9] using the Darknet-53 network to detect the incoming video sequence. This network is very complex, requiring a lot of computational power for its processing. Due to the structure of the network, the detection speed is also affected. YOLOv3 consists of layers of convolution, built from a deep network, to obtain better accuracy than previous networks. This network uses a logistical classifier to calculate the object's probability of being assigned a specific label.

Previous versions use the softmax function to generate the probabilities of the labels. To minimize the loss of classification, it uses binary cross-entropy for each label instead of the mean square error used in previous versions. The network also has multi-scale detection capabilities, using the concept of pyramid resource networks. This allows YOLOv3 to detect objects of various sizes. In more detail, when an image of three channels of R, G, and B is introduced into the YOLOv3 network it is emitted from three detection layers. YOLOv3 predicts regions with 3 different scales and then extracts the characteristics from those scales.

The result of the network predictions is a 3D tensor, which outputs the detected region, the accuracy of detection, and the class of the detected object due to the change of the dimension of the final tensor in comparison with the previous version, as follows:

$$N \times N \times (3 * (4 + 1 + C)) \tag{1}$$

$N \times N$: number of cells grind of the system

3: to decode the characteristics extracted from each of the 3 scales

$4 + 1$: to decode the bounding box displacements along with the scores of the identified objects

C: and the number of network input classes

It is important to point out that the detection speed of YOLOv3 is as fast as YOLO [19] and YOLOv2 [20]. Therefore, in terms of the trade-off between accuracy and speed, YOLOv3 is suitable for real-time object detection applications. However, in general, it still has an accuracy less than a two-phase detector using a delimitation phase of the region to be detected. To solve that problem Tiny-yolov3 is a simplified version of YOLOv3.

The structure of the Tiny-yolov3 network has only seven convolutional layers and 6 layers of grouping. The network structure of Tiny-yolov3 the simplified network improves the detection speed, but also loses part of the detection accuracy [16].

The algorithm proposed in this paper uses a similar methodology as [15], using YOLOv3 is shown Fig. 3 detector to detect the incoming video sequence, assigns the tracker to the detection result of the first frame, and then uses the Kalman filter to predict the movement of all tracking targets from frame to frame and then calculates the Intersection over Union (IoU) distance from the target between the two frames using the Hungarian method to obtain the best correlation results, record the target of the successful game and the target that is not successfully matched. For the target unsuccessful to extract the appearance feature from the depth, the Hungarian method is used again to obtain the correlation result so that the target loss due to the long-term occlusion of the target can be avoided to some extent by maintaining the high frame rate.

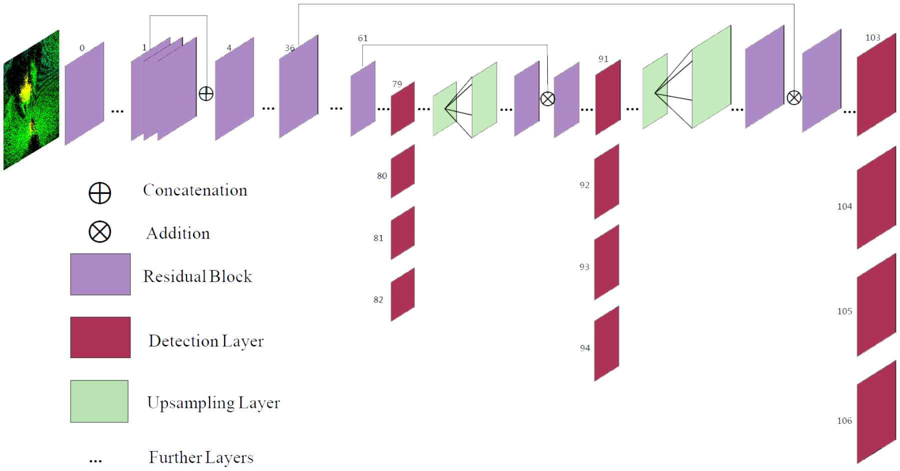


Fig. 3. Yolov3 architecture

2.3 DEPTH IMAGES

A Depth Image is an image that contains information about the distance between the surface of objects taking as a starting point the sensor that is capturing the transmitted waves. The sensor that captures distances from objects is a TOF sensor. TOF means ”Time of Flight”, in show Figure 4, which is a technology for measuring the distance between objects by measuring the time a sound wave, light, microwave, or other wave takes to travel to the object and return. It has been widely used for sonar, research, and motion capture.

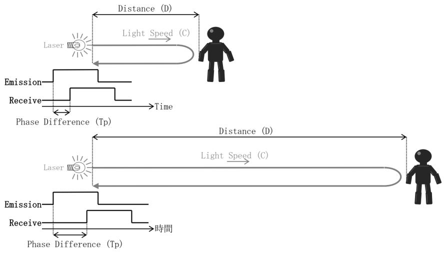


Fig. 4. TOF Sensor Operation

Color combinations define the Depth Image’s perception of distance, generating an image similar to thermal images. The combination of thermal images and depth has recently been used for various applications, such as human detection, 3D maps of environments, and detection of the proximity of a certain object to the sensor, or from the object to another object. These images, in combination with other sensors, map the environment and serve as a basis for generating 3D images. In this process, a combination of depth images and distance sensors is performed, resulting in the creation of an additional dimension in the image. This interpolation ultimately generates 3D images.

The use of two types of chambers implies a procedure to calibrate their relative posture (that is, extrinsic parameters). The generated images have a color gradient that personifies the distance of the objects to the sensor, as we can see in Fig. 5.

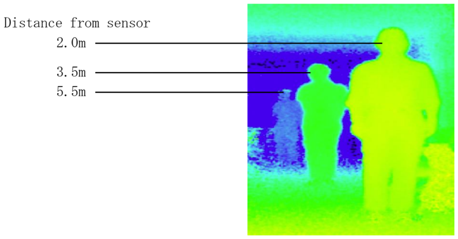


Fig. 5. Image generated by the TOF sensor

3 RELATED WORKS

The CNNs have been used in several person-detection and tracking applications that have similarities with this work and, in this section, some of these preexistent researches are listed, that is:

- In [16] a system for real-time detection, based on low-resolution thermal infrared images and CNNs, is proposed. To this end, the authors used the Maximally Stable Extremal Regions (MSER) and a self-designed CNN to perform the classification. The CNN was trained to differentiate the human hot spots detected through the MSER from others, such as dogs and cats. Regarding the results, the architecture proposed was applied to several different datasets of these images and reached up to 80 percent in error reduction when compared with other approaches based on MSER, Support Vector Machines, Decision Trees (DTs), and local image descriptors.
- The small target detection on infrared images is explored in [17]. In this work, the Point Spread Function (PSF) is used to generate training instances from the images, and a CNN architecture designed by the authors is used for the classification. The proposed method was compared against the Max-Median filter and the Top-hat methods and resulted in a decrease in false positives and, as described by the authors, an efficient and effective method that can be used for target detection systems.
- People tracking using CNN features is explored in [18]. In this work, the authors performed a comparison of two approaches. The first one used the Faster-RCNN framework with the VGG16 CNN, inserted a Region Proposal Network (RPN) between specific layers of the CNN, and used the 4096 features outputted from the *fc7* layer at the Euclidean Distance calculation as a similarity metric between individuals in different frames. The second one used the Siamese Neural Network (SNN) to distinguish different people in different frames by the twin functions and thus perform the tracking. Related to the results, the Euclidean approach outperformed the SNN one also, as the authors described, the Euclidean method depends heavily on the robustness of the individual frame detection. Despite this work having a scope very close to our own, this study was conducted with a different dataset. This dataset contains images of individuals with all colors, textures, and environmental characteristics. Therefore, its application is carried out in a less restricted manner. In our case, due to the nature of the sensors, our images lack most of these characteristics that can be captured by regular cameras.
- The ResNet-50 is customized in [14] to a specific dataset to perform facial recognition. A dataset with 400,000 images from 1200 individuals, with each image accompanied by its equivalent 3D model, is used, where this data is processed and combined in RGB-Ds images and the ResNet variant is trained with the dataset resulting dataset. As a result, there are the specific CNN architecture and the 82% accuracy reached in the given experiment.
- In [19], the multiple-target vehicle tracking, using compressive measurements in low-resolution infrared videos generated from a moving camera is ex-

plored. The study takes three different approaches, being: a Gaussian Mixture Model-based tracker, the STAPLE tracker, and the ResNet-18. As the research's conclusion, the first two classifiers had serious issues given the quality of the data whereas the CNN approach performed better than the other two but, had to use several data to train which, given the amount of existing data at the work, was considered a limitation.

4 METHOD

Our main goal is to propose a comparison analysis of the performance of YOLOv3, YOLOv3 Tiny, and RESNET Convolutional Neural Networks in the face of processing 3D-depth image got from a device to provide human detection, count, and tracking.

It is important to point out that the device used in this work produces a 3D matrix from a 3D-depth image with 3-coordinates (x, y, d) where (x, y) are position coordinates into the xy-plane and d is the distance of the device to the person detected.

Thus, the method used follows the following steps:

1. The 3D images and their corresponding matrices are obtained from the device in real-time. After obtaining the 3D image, it is stored in a variable as a tensor with the format (x, y, d), where two sensors work together, one capturing a 2D depth image and another sensor capturing the distance from the sensor to objects present in that area of interest, thus generating a 3D image.

2. The CNN is applied to count the detected humans. The CNN is applied to the 3D image, converted to 2D, identifying the people contained in that area of interest, performing the count and creating bounding boxes, showing the type of object identified and the accuracy of that identification.

3. Each detected human receives a label (detection framing rectangle DFR) and a reference pixel is marked at the center of the DFR; For each bounding box created, people receive a label, and some points of interest are also calculated within that area identified by the CNN, creating mapping points for people and performing tracking using the method person using the algorithm *Simple Online and Realtime Tracking with a Deep Association Metric* (Deep SORT) [20].

4. From the image matrix and the reference pixel, a position of the detected human is generated and stored; For each detected person, data is generated containing information about their position in the environment, points of interest within the bounding box, detection accuracy, and trajectory traveled by the person. In this way, the CNN algorithm, together with the Deep SORT method, can perform tracking and if the identified person leaves the identification area, the network method in combination with the stored characteristics can re-label the person with the same identification and continue

storing data.

5. The information stored, a detected human tracking, and its trajectory in the scenario is created.

The Figure 6 represents the result of the method used.

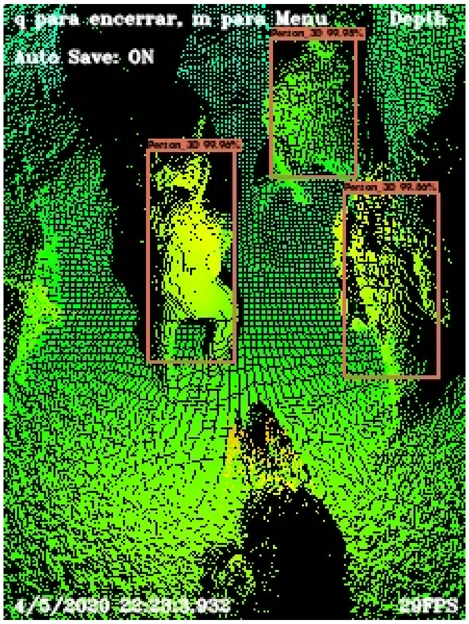


Fig. 6. Example of a 3D Deep image, with its detections

In this way, we can identify people even when there is no light in the environment of interest, and also perform tracking at alert points in the environment. Thus, this method can be used as a security mechanism in Banks, Hospitals, or any place that requires an effective monitoring tool that operates even when there is no presence of light in the environment.

5 EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed human detection algorithm is tested on our collected dataset from a specific scenario. The data collection for the developed algorithm was done in partnership with a confidential technology company, where it was agreed that the used data could not be disclosed to the community due to potential business interests.

All experiments presented in this section were carried out on a core i5 notebook computer with a 2.60 GHz CPU and 8GB RAM.

The images were acquired using a TOF sensor, which measures the distance between objects by calculating the time that a sound wave, a light, a microwave, or another wave takes to travel to the object and return.

The principle of a TOF sensor can be explained using an Infrared light beam as an example. After a pulse is emitted there is a phase difference between the emitted pulse and the pulse that was reflected on the surface of an object. The farther the distance the larger the phase difference.

Since the light speed is constant and quite stable and with the calculated phase difference TOF sensors are able make to precise measurements even during short periods. Such technology allows the generation of Depth Images.

In our experiments, we compared the algorithms, Resnet, Yolov3, and Yolov3 tiny. We used a dataset with over 500 image samples and 5 videos with a duration of 1min each, to compare the probability of detection, and the Fps, of each method. We use both image and video to check the performance and accuracy of the detections.

The results of each method, is presented in Table 1.

Table 1. Result of model contrastive experiments

Models	Mean Average Precision	Variance Precision	Fps	Weights
ResNet	0.848	0.551	2.02	150 MB
Yolo-v3	0.975	0.002	3.31	240 MB
Yolo-v3-Tiny	0.911	0.007	18.16	33 MB

In our tests, we noticed a very high variance in the detections, using Resnet, and this variance occurred more frequently when people were overlapping, i.e., when one person was in front of or behind the other, making it difficult for the network to identify with total precision whether it was, in fact, one person.

Yolov3 and Yolov3 Tiny, on the other hand, showed a low variance and a high probability of detection, showing their efficiency. Note that Yolov3 tiny presents a lower stability, if compared to Yolov3, despite its lower probability, in our data set, it presented a low variance, showing a better method to be used in the industry, for presenting good stability, with a great detection range, presenting a superior performance in computational time.

Also, note that Yolov3 Tiny has a much higher FPS compared to the other networks. Due to its simpler architecture, on computers that do not use a dedicated GPU for processing, this network proves to be very efficient for detections, however, it presents a lower hit probability when compared to Yolov3.

6 Conclusion

In this work, we present a method to identify individuals in deep and three-dimensional images generated through deep imaging. This method proves to be efficient and effective for detecting and tracking people in environments devoid of illumination. In addition to proposing an approach for identifying and tracking individuals using well-established network architectures and methods from the literature, it is intended for potential application across various industries, particularly in locations with minimal or no illumination, where more precise security measures are necessary.

To demonstrate the method, we employed LIDAR sensors for image generation, which lack many characteristics such as texture and color, thus posing a challenge in feature extraction through neural networks. We trained neural networks for object detection to showcase their usability even in scenarios where images possess limited features. This study illustrates that even with sparse information, these networks maintain stability and scalability within industrial environments. The detection results highlight the robustness of our approach compared to other related methods, using YoloV3 Tiny.

Nevertheless, several unresolved challenges persist, such as low accuracy when images of individuals overlap, the presence of excessive noise in 3D deep images, as well as tracking failures when the object exits the field of view, resulting in its identification as a new person. In future endeavors, we plan to leverage these 3D data alongside distance sensor data, which support image generation, to achieve more effective tracking. Furthermore, we intend to employ advanced image processing mechanisms to enhance quality and consequently improve detection accuracy.

Acknowledgment

References

1. M. N. Murty and V. S. Devi, *Introduction to Pattern Recognition and Machine Learning.* Co-Published with Indian Institute of Science (IISc), Bangalore, India, 2015. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/8037
2. T. Surasak, I. Takahiro, C. H. Cheng, C. E. Wang, and P. Y. Sheng, "Histogram of oriented gradients for human detection in video," *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, pp. 172–176, 2018.
3. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.
4. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.
5. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
6. R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1440–1448, 2015.

7. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 779–788, 2016.
8. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6517–6525, 2017.
9. ——, "YOLOv3: An Incremental Improvement," 2018. [Online]. Available: http://arxiv.org/abs/1804.02767
10. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.
11. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 2016.
12. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
14. T. Gong and H. Niu, "An implementation of resnet on the classification of rgb-d images," pp. 149–155, 2019.
15. Z. Yi, Y. Shen, and Q. Zhao, "Multi-Person tracking algorithm based on data association," *Optik*, vol. 194, no. April, 2019.
16. C. Herrmann, T. Müller, D. Willersinn, and J. Beyerer, "Real-time person detection in low-resolution thermal infrared imagery with mser and cnns," in *Electro-Optical and Infrared Systems: Technology and Applications XIII*, vol. 9987. International Society for Optics and Photonics, 2016, p. 99870I.
17. D. Zhao, H. Zhou, S. Rang, and X. Jia, "An adaptation of cnn for small target detection in the infrared," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 669–672.
18. D. Chahyati, M. I. Fanany, and A. M. Arymurthy, "Tracking people by detection using cnn features," *Procedia Computer Science*, vol. 124, pp. 167–172, 2017.
19. C. Kwan, B. Chou, A. Echavarren, B. Budavari, J. Li, and T. Tran, "Compressive vehicle tracking using deep learning," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2018, pp. 51–56.
20. N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 748–756.