HUMAN DETECTION BASED ON A SEQUENCE OF THERMAL IMAGES USING DEEP LEARNING

X. Wang [1], S. Hosseinyalamdary [2]

[1] Faculty of Geo-information Science and Earth Observation (ITC), x.wang-2@student.utwente.nl

[2] Department of Earth Observation Science (EOS), Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, Enschede, s.hosseinyalamdary@utwente.nl

KEY WORDS: Human detection, Temporal consistency, Deep learning, Thermal images

ABSTRACT:

Human detection has been playing an increasingly important role in many fields in recent years. Human detection is a still challenging task because, for the group of people, each individual has his unique appearance and, body shape.
Compared with the traditional method, the deep learning neural network has the advantages of shorter computing time, higher accuracy and easier operation. Therefore, deep learning method has been widely used in object detection. The current state of art in human detection is RetinaNet. Among all the deep learning approaches, RetinaNet gives the highest accuracy of human detection (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018).The temporal component of video provides additional and significant clues as compared to the static image. In this paper, the temporal relationship of the images is utilized to improve the accuracy of human detection. Compared to using only an image, the accuracy of human detection is 21.4% higher when a sequence of images is applied.

1. INTRODUCTION

1.1 Background Information

Human detection is a useful tool in many research fields. This technology provides a solid technical foundation for these problems and guarantees their development. The importance of human detection in autonomous driving, Post-disaster rescue, automated surveillance, military and robotics services has become significant (Gajjar, Gurnani, & Khandhediya, 2017).
After the disaster, a device equipped with a human detector could help the rescue team find out where the survivors are (Doherty & Rudol, 2007). If the specific location of the trapped people can be detected and reported to the search and rescue personnel, it will greatly enhance the efficiency of the rescue team. In autonomous driving, human detection technology ensures the safety of both drivers and pedestrians (Ballas, Larochelle, & Courville, 2015). Human detection plays a key role in automated surveillance (D, Manjunath, & Abirami, 2012 ; Moore, 2003). Human detection technology can help to monitor some suspicious activities. In the military, the human detection device can help to monitor the enemy's action.
No matter which application is mentioned, how to obtain high precision human detection algorithm is the primary consideration. The increased precision achieved in the field of human detection will play a huge role in other applications that rely on human detection technology. So how to improve the accuracy of human detection is a problem worth studying.

1.2 Problem Statement

No matter which application is mentioned, how to obtain high precision human detection algorithm is the primary consideration. The increased precision achieved in the field of human detection will play a huge role in other applications that rely on human detection technology. So how to improve the accuracy of human detection is a problem worth studying.
The task of human detection can be divided into two parts. We need to determine if there are people in that image; if a human is detected in the image, and we still need to get their corresponding coordinates to show his location.
So far, many scholars and scientists have invested a lot of efforts in human detection and made some achievements. Human detection is still a very challenging problem. It can be affected by occlusions, blurred backgrounds and poor visibility at night. Human detection is still a challenging task because of the different appearances and postures of each person (Gajjar et al., 2017).
As well as camera capture at various views to the human body. This can cause humans to be obscured by other people or objects. The contours of people far from the camera are blurred, which can also affect the accuracy of human detection.

1.3 Characteristics of visual and thermal images

Visual images and thermal images are the two major information sources used in human detection researches (Fan, Xu, Zhang, & Chen, 2008).
Visual images are more widely used than thermal images. Visual image refers to an image synthesized using RGB channels. In the visual images, the human detection has been extensively studied and has achieved good results. Most human detection tasks are still based on visual images (Hwang, Park, Kim, Choi, & Kweon, 2015). The visual image has the disadvantage of being sensitive to light changes. As a result, they are vulnerable to insufficient exposure or excessive exposure during a sudden change in illumination. Visual images are affected by poor lighting condition. Because they need plenty of light. Therefore, when light is insufficient, such as in the night, at dusk and shadow area, visual image quality drops.
Thermal images are the representation of the amount of infrared energy emitted, transmitted and reflected by the object in terms of brightness (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). The amount of radiation emitted by the object increases as the temperature of the object increases. Due to this, the thermal camera can measure the temperature of an object. The body temperature is different from the temperature of the environment. Thus, the thermal image can distinguish between human and other objects, particularly at night or in shadow.

Advanced thermal detectors also have the ability to detect infrared radiation behind thin walls or other obstructions.
The disadvantage of thermal images is that the thermal detector is susceptible to non-human factors when the outside temperature is high (Kim et al., 2017; Baek, Hong, Kim, & Kim, 2017). They provide limited performance in human detection. The closer the body temperature to the external environment temperature, the worse human detection accuracy in the thermal image. In desert areas, for example, thermal images perform poorly because the temperature of sand changes frequently and human detection can be a challenge. In addition, many other things that generate heat automatically may interfere with human detection.
Besides, the resolution of the thermal image is low, and it's very difficult to identify the distant human body in a thermal image (Fan et al., 2008). Thermal light sources may also have an effect on the quality of the image at night. So, the thermal information is not as detailed as the visual images.

1.4 Temporal consistency

Compared with the single image, video analysis provides more information for the identification of the task. It adds the time component and therefore, it improves the accuracy of human detection by adding movement information and trajectory (Ng et al., 2015).
The temporal component of video provides additional and significant clues as compared to the static image classification since many actions can be reliably identified based on motion information (Simonyan & Zisserman, 2014). Compared to using only an image, the accuracy of human detection is higher when a sequence of images is applied. In other words, we can use the temporal relationship of the images to improve accuracy of the human detection.
All objects are in constant movement. The occurrence of motion takes time. The time is added to the neural network while motion information is added. Nowadays, there are two main ways to add time information into convolutional neural networks. One is to process two or more frames simultaneously, while the other works with optical flow and corresponding frames at the same time.
The essence of the second method is to extract the motion information between different frames. The temporal information is converted into movement of detected objects and added to the convolutional neural network.
If a human being is identified in a series of continuous frames, it will be more confidently detected as human.

1.5 Research objective

The main objective of this study is to apply deep learning method and temporal information to convolutional neural networks to do human detection with a sequence of thermal images.

1. What is the state of art in human detection? How accurate is the state of art?
2. How much temporal CNN can improve human detection using the state of art?

1.6 Innovation aimed at

Many methods of deep learning have already been applied to object detection and have made remarkable achievements. Temporal information is of great importance here because it enforces temporal consistency among thermal images and improves human detection accuracy.

Here, I propose a novel idea to add temporal information to the thermal images and apply Temporal convolutional neural network (T-CNN). Later, I am going to compare the results of these approaches to human detection.
The key contribution of this study is to find out which type of deep learning methods perform better in human detection by using a sequence of thermal images. This is the first time to use a sequence of thermal images to do human detection based on my own knowledge. We are the first group to use this innovative approach to do human detection.

2. RELATED RESEARCH

2.1 Deep learning

The concept of deep learning stems from research on artificial neural networks. Multiple-hidden layer perceptron is one of the structures of deep learning (Lecun, Bengio, & Hinton, 2015). The main idea of depth learning is to incorporate low-level features into a more abstract level of advanced presentation attribute categories or features. The distributed characteristic representation of the data is found.
Deep learning is a kind of machine learning method based on data representation. Observations (such as images) can be expressed in a variety of ways. For example, a vector of the intensity value of each pixel, or more abstractly represented as a series of edges, areas of a particular shape, and so on (Ramachandran, Rajeev, Krishnan, & Subathra, 2015).
A reference CNN consists of two main processes: feature extraction and classification (Zhu et al., 2017). The purpose of feature extraction is to extract different parts of each object, such as human head, arms, and legs. Classification refers to calculating the degree of certainty of a person at this location. If the certainty is high, then the outcome is going to be one person in that position and vice versa. The basic principle of feature extraction is the detection of features from low to high levels. Low-level features include edges and colours. A high-level feature is an object, such as a cat, a tree, and a table.

Nowadays, more and more deep learning approaches are applied to human detection.
R-CNN (region-based convolutional neural network method) combine region proposals with CNNs (Girshick, Donahue, Darrell, Berkeley, & Malik, 2012) (Uijlings, Sande, Sande, & Smeulders, 2012). First, they applied high-capacity CNNs to bottom-up regional proposals to locate and segment objects. Then, when marked as insufficient training data, the supplementary task is pre-trained with supervision, and then a domain specific fine adjustment is performed, which can significantly improve performance. The drawback of R-CNN is that it is slow at training-time. Because it needs to run full process of CNN for watch region proposal. The other drawback is that CNN features are not updated in response to regressors.
Soon afterwards, a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection (Fast R-CNN) was proposed (Girshick, 2015)(Hosang, Omran, Benenson, & Schiele, 2015 ; Li et al., 2017)). Compared with R-CNN, Fast R-CNN not only improves the training and testing speed but also improves the detection accuracy. Region proposal method is then implemented on the feature map. Fast R-CNN trains deeper neural networks than R-CNN. F-CNN is faster than R-CNN and it can achieve higher accuracy in testing task (Girshick, 2015).

Ren et al. came up with an object detection algorithm that eliminates the selective search algorithm and allows the network

learn the region proposal (Ren, He, Girshick, & Sun, 2015). This method called Faster R-CNN.Similar to fast R-CNN, the image is provided as input to the convolution network, which provides the convolution feature graph. Instead of using a selective search algorithm on the feature map to identify regional Suggestions, it's using individual networks to predict regional Suggestions. Then use the RoI pool layer to predict regional suggest refactoring, the layer is used for classifying suggest area of the images and predict the offset value of the bounding box.

2.2 Temporal Convolutional Neural Network

Ng et al. propose two deep learning methods for handling long video classification (Ng et al., 2015). The first approach explores the various convolutional time property pool architectures and finds the design choices that need to be made to adapt to CNN for this task. The second way they used a cyclic neural network to model video as an ordered frame sequence.

Simonyan applied the motion information in two adjacent images to the deep convolutional network of video action recognition (Simonyan & Zisserman, 2014).

In action recognition, both Ng et al. and Simonyan & Zisserman suggested applying optical flow to better enforce temporal consistency.

In the detection of human abnormal activities, Zhou et al. proposed a method to detect and locate abnormal activities in the video sequence of crowded scenes (Zhou et al., 2016). The main novelty of this method lies in the coupling of anomaly detection and spatial-temporal CNN). By performing space-time convolution, the architecture captures features from the spatial and temporal dimensions while the appearance and motion information contained in the continuous frames is extracted. Experimental results

Ballas and his colleagues proposed a method which successfully considered the local and global time structures of video to generate the description (Ballas et al., 2015). This method combines the representation of temporal and spatial 3D convolutional neural network (3D-CNN) to short-time dynamics. 3D CNN represents training through the video action identification task, giving a representation that is compatible with human action and behavior.

Karpathy studies multiple ways to extend CNN connectivity in the time domain to take advantage of local spatiotemporal information and proposes a multi-resolution, centrally structured architecture as a promising approach to accelerate training (Karpathy et al., 2014). Lea proposed temporal CNN which is a unified approach to capturing relationships in a hierarchical manner at low, medium and high time scales.

3. METHODOLOGY

3.1 Data Preparation

To add temporal information into neural network, three frames are stacked into one image to train the neural network. The principle of stacking images is shown in Figure 1. Then these stacked images are used to train temporal convolutional neural networks. The time span between the two adjacent frames is 0.1 seconds.

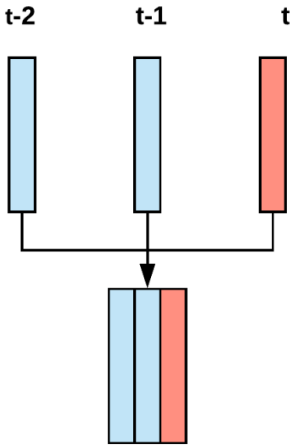

Figure 1 The method to stack images (The red rectangle is key frame and the blue rectangles are attached frame)

For reference convolutional neural network, images are used separately to train the network and do validation and testing part. This means that the model is trained with single images. Only one image is processed at a time. Temporal information is not added to the neural network in this case. This network is just a reference convolutional neural network.

3.2 Training and testing part

Unlike other stationary objects, humans move all the time. The essence of adding temporal information to the convolutional neural network is to add motion information to the model in order to improve the accuracy of human detection. In this study, I will compare the results of the reference CNN and temporally-consistent CNN for human detection. It is then concluded that there is a conclusion about that effect of temporal consistency on the accuracy of human detection.

A parameter needs to be set during the testing process, score threshold. “score” is a value which represents the level of confidence that the retinal network identifies an object as a human. A bounding box with a score higher than the score threshold will be recorded.

On the accuracy of the model evaluation mainly based on the classification accuracy, localization accuracy, and computational complexity in three aspects. The way to evaluate classification and positioning accuracy is mean average precision(mAP) (Han, Zhang, Cheng, Liu, & Xu, 2018).

In this paper, the average accuracy is used to find the best results for each model among different score thresholds. Recall, precision, F1 score and average Intersection of Union (IOU) will be the four criteria to evaluate the outcome of a human detection model.

Precision tells you how many of the detected objects were correct. It is a measure of completeness (Powers, 2007). Recall, also known as sensitivity, tells you how many of the objects that should have been detected were actually selected. It is a measure of exactness (Powers, 2007). F value, on the other hand, is the evaluation index integrating these two indicators and is used to comprehensively reflect the overall index. After calculating the value of precision and recall, F1 value can be obtained, which is the harmonic average of the two value(Sasaki, 2007).

Intersection over Union (IoU) is equal to overlap region divided by union region. IoU evaluates the geometric relation between labeled bounding box and predicted bounding box (Han et al., 2018).

3.3 Validation part

Validation part is done with the training of the model. Validation dataset is ten percent from training dataset. Because the KAIST dataset does not have independent validation data. In order to make the model training more complete and to monitor the overfitting problem, I extracted the last ten percent of each folder of the training dataset used for validation.

4. IMPLEMETATION

4.1 RetinaNet

All the models are trained using RetinaNet network. RetinaNet, introduced by Lin et al, achieves a higher accuracy compared with most of neural networks on COCO dataset (Lin et al., 2018). Among all the deep learning approaches, RetinaNet gives the highest accuracy of human detection.
The current state of art in object detection is RetinaNet. It is a robust one stage object detector (Lin et al., 2018). RetinaNet is composed of a backbone network and two subnetworks. One of the subnetworks is used for classification, called classification subnet. The second subnet called box regression subnet which performs convolution bounding box regression. RetinaNet has a specific loss function, which can be used to address imbalance between foreground and background classes during training.

4.2 Dataset

In this project, a total of three datasets were applied. They are COCO, KITTI and KAIST dataset.
As mentioned in introduction, human detection is still a challenging task. Because of the diversity of human clothing, posture, and appearance. Compared with other kinds of object detection, human detection needs a large number of samples to get an ideal result. The thermal image dataset is much smaller than the visual image dataset. The KAIST dataset is far from sufficient for human detection. To deal with this problem, we adopted the idea of “domain adaptation”. Here, COCO and KITTI images are introduced to pre-train the model.
COCO has 330k images 80 different categories. For KITTK dataset, the camera is fixed on a vehicle. The images are specialized for autonomous driving, which is closer to KAIST dataset.
A set of thermal images used in this thesis is from the website: KAIST (Korea Advanced Institute of Science and Technology) Multispectral Pedestrian Detection Benchmark. The KAIST Multispectral Pedestrian Dataset consists of 95000 colour-thermal pairs. They are captured by a vehicle which carries a colour camera and a thermal camera. The images are captured during day and night time. The dimension of these images is 640*512 pixels. Both horizontal and vertical resolutions are 96 dpi.
In this thesis, I only use thermal images part. The dataset used a long-wave infrared (7.5~13μm, also known as the thermal band) camera. The data is available online. The KAIST dataset has 95328 images in total. The training part contains 50187 images, and testing part has 45141 images. The testing annotation files used in this study are provided by Ms. Jingjing Liu due to plenty of errors occurred in the original one. This is a subset of KAIST testing dataset.

4.3 Fine tuning model

In order to get better results and solve the overfitting problem, a fine-tuning model was used. I prepared two kinds of fine-tuning models. One of them is based on a pretrained COCO model and then trained on KITTI data. The COCO model is found from GitHub. Another one only uses KITTI training data to train a model. The accuracy of these two models are shown in Table 1. The COCO-KITTI model has a higher accuracy than KITTI model, so in this study, COCO-KITTI model was chosen as fine-tuning model to train four models mentioned above.

COCO_KITTI Model	KITTI Model
Pedestrian with average precision: 0.4436 Cyclist with average precision: 0.4563	Pedestrian with average precision:0.3432 Cyclist with average precision: 0.3910

Table 1 The comparison of COCO_KITTI model and KITTI model

It is considered that both COCO and KITTI dataset are colour images. And in this project, I will use thermal images to do human detection. The different data types may have a negative effect on the model. Because of this, I trained another model only using the KAIST thermal dataset. This means this is a network without fine-tuning model.
I used RetinaNet evaluation code to evaluation these two models. The result can be seen in Table 2. The evaluation results of COCO_KITTI model and KITTI model. The result shows that fine tuning model can help to get a higher accuracy.

COCO_KITTI_KAIST (thermal) Model	KAIST (thermal) Model
average precision: 0.0149	Average precision: 0.0140

Table 2 The evaluation results of COCO_KITTI model and KITTI model.

The result shows that fine tuning model can help to get a higher accuracy
I tested these two models using testing dataset. Here I use mAP to evaluate the results. I calculated the Mean IoU, precision and recall. The results are shown in Table 3.

COCO_KITTI Model	KITTI Model
Mean IoU is 0.4556 Precision is 0.1085 recall is 0.0975	Mean IoU is 0.4494 Precision is 0.0904 recall is 0.0818

Table 3 The testing results of two model

As can be seen from the above results, the COCO, KITTI and KAIST thermal model has a lower loss and a higher testing accuracy. Here, the testing dataset is the original dataset instead of the improved one. So, the output is very disappointing. This shows that the fine-tuning model performs better in overfitting than the model without fine-tuning.
The COCO-KITTI model is chosen as fine-tuning model to train other networks.

5. RESULTS

The results show that the convolutional neural network which added temporal information has a better performance than the neural network trained using single images (as shown in Figure 2). The biggest point in each line shows the best performance

with the maximum rectangle area of each model (blue dot at 0.5, red dot at 0.7). The recall-precision lines of temporal CNN (red) are obviously above the recall-precision line of single images neural network. This means that the accuracy of human detection using temporally-consistent network has been greatly improved.

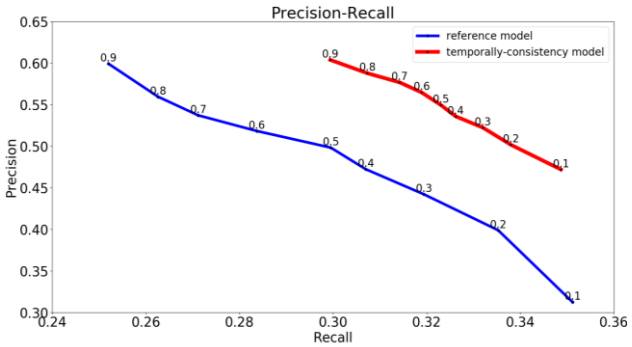

Figure 2 Precision-recall graphs of two models

The areas of rectangle of each model on every score threshold can be seen in **Error! Reference source not found.**. The highlights are the maximum value of precision and recall products of each model. The corresponding score is considered as the "best" score threshold. By comparing the "best" results of these models, we can find that continuous later neural network is better than the other three models. The largest area of the rectangle (score threshold on 0.7 continuous later network) is 21.4 percent larger than the smallest (score threshold on 0.5 single images network).

According to this table, the "best" performance of the temporally-consistent network will occur with a higher score threshold. This also means that the quality of the detected human is also higher than that of the single images network.

The F1 score-score threshold graphs of two models are shown in Figure 3. From this figure, we can see that F1 score expresses the same information as the precision-recall diagram. The line for temporal CNN is above the line of reference CNN.

Recall reflects the model's ability to detect positive samples. The higher the recall is, the better the model can recognize positive samples. Precision reflects the ability to tell negative samples apart. F1-score is a combination of these two criteria. The higher the F1-score is, the more robust the model is.

It can be seen from this graph that graph for temporally-consistent CNN is nearly parallel to the horizontal axis. This indicates that the networks containing time information play a positive role in the stability of the human detection model under the extreme value of score threshold that is too large or too small.

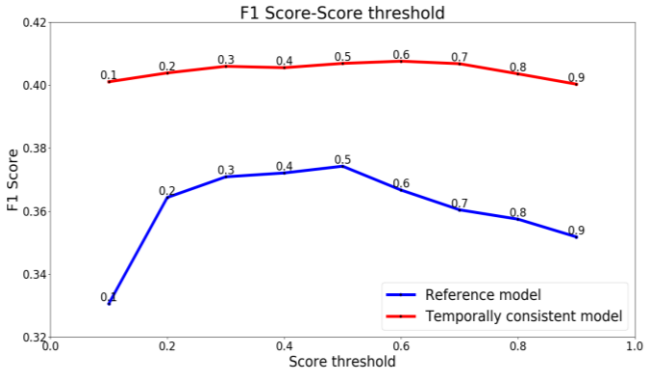

Figure 3 The F1 score- score threshold graph of two models.

Each score threshold has a corresponding average IOU, as shown in Figure 4. The average IOU will grow with the score threshold. This is because, with the higher the score threshold, we can become more confident to say the detected human is correct. As you can see from this graph, the reference model grew faster than the other one. This also shows that the networks with temporal information are more stable than the ordinary network.

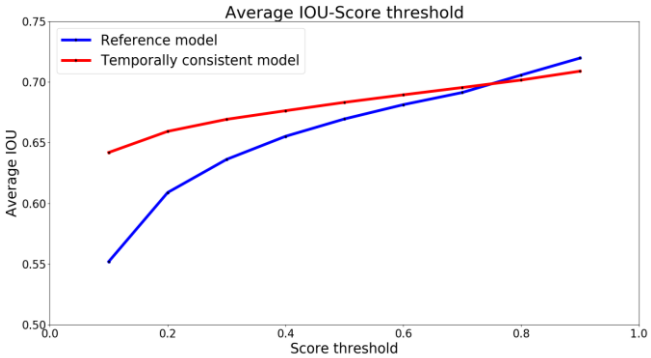

Figure 4 Average IOU-Score threshold graphs of two models

6. CONCLUSION

In this thesis, we discussed the role of temporal consistency in human detection. This study trained two different types of convolutional neural networks. By comparing the testing results of different models, we know that the human detection accuracy of the convolutional neural network with temporal information is better than that of the reference convolutional neural network. The accuracy of the temporally consistent model for human detection is 21.4% higher than that of the reference model, according to the product of recall and precision. Temporal information also could increase the stability of human detection model.

REFERENCES

Baek, J., Hong, S., Kim, J., & Kim, E. (2017). Efficient Pedestrian Detection at Nighttime Using a Thermal Camera. *Sensors*, *17*(8), 1850. https://doi.org/10.3390/s17081850

Ballas, N., Larochelle, H., & Courville, A. (2015). Describing Videos by Exploiting Temporal Structure, 4507–4515. https://doi.org/10.1109/ICCV.2015.512

Correa, M., Hermosilla, G., Verschae, R., & Ruiz-del-Solar, J. (2012). Human Detection and Identification by Robots Using Thermal and Visual Information in Domestic Environments. *Journal of Intelligent & Robotic Systems*, *66*(1–2), 223–243. https://doi.org/10.1007/s10846-011-9612-2

Doherty, P., & Rudol, P. (2007). A UAV Search and Rescue Scenario with Human Body Detection and Geolocalization. In *AI 2007: Advances in Artificial Intelligence* (pp. 1–13). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-76928-6_1

Fan, X., Xu, L., Zhang, X., & Chen, L. (2008). The Research and Application of Human Detection Based on Support Vector Machine Using in Intelligent Video Surveillance System. In *2008 Fourth International Conference on*

Natural Computation (pp. 139–143). IEEE. https://doi.org/10.1109/ICNC.2008.315

Gajjar, V., Gurnani, A., & Khandhediya, Y. (2017). Human Detection and Tracking for Video Surveillance A Cognitive Science Approach. *ArXiv:1709.00726v1 [Cs]*, 2805–2809. https://doi.org/10.1109/ICCVW.2017.330

Girshick, R. (2015). Full-Text. *IEEE International Conference on Computer Vision (ICCV 2015)*, 1440–1448. https://doi.org/10.1109/iccv.2015.169

Girshick, R., Donahue, J., Darrell, T., Berkeley, U. C., & Malik, J. (2012). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2–9. https://doi.org/10.1109/CVPR.2014.81

Gowsikhaa D, Manjunath, & Abirami S. (2012). Suspicious Human Activity Detection from Surveillance Videos. *(IJIDCS) International Journal on Internet and Distributed Computing Systems*, *2*(2), 141–149.

Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection. *IEEE Signal Processing Magazine*, *35*(1), 84–100. https://doi.org/10.1109/MSP.2017.2749125

Hosang, J., Omran, M., Benenson, R., & Schiele, B. (2015). Taking a Deeper Look at Pedestrians. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4073–4082.

Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 1037–1045. https://doi.org/10.1109/CVPR.2015.7298706

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. F. (2014). Large-scale video classification with convolutional neural networks. *Proc. IEEE CVPR*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

Kim, J. H., Hong, H. G., & Park, K. R. (2017). Convolutional Neural Network-Based Human Detection in Nighttime Images Using Visible Light Camera Sensors. *Passaro VMN, Ed. Sensors (Basel, Switzerland)*, *17*(5), 1065. https://doi.org/10.3390/s17051065

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2017). Scale-aware Fast R-CNN for Pedestrian Detection. *IEEE Transactions on Multimedia*, 1–10. https://doi.org/10.1109/TMM.2017.2759508

Lin, T., Goyal, P., Girshick, R., He, K., & Piotr Dollar. (2018). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2858826

Moore, D. (2003). A real-world system for human motion detection and tracking.

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101

Powers, D. M. W. (2007). Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation.

Ramachandran, R., Rajeev, D. C., Krishnan, S. G., & Subathra, P. (2015). Deep learning – An overview. *International Journal of Applied Engineering Research*, *10*(10), 25433–25448. https://doi.org/10.1016/j.neunet.2014.09.003

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *ARXIV*, 1–14.

Sasaki, Y. (2007). F-measure.pdf, 1–5.

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos, 1–9. https://doi.org/10.1017/CBO9781107415324.004

Uijlings, J. R. ., Sande, K. E. . Van De, Sande, T., & Smeulders, A. W. M. (2012). Selective Search for Object Recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5

Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., & Zhang, Z. (2016). Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, *47*, 358–368. https://doi.org/10.1016/j.image.2016.06.007

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017, December). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*. https://doi.org/10.1109/MGRS.2017.2762307

Revised January 2019