

Investigating the effect of distance on the implementation of RCNN automatic detection technique to the human body

Majed Kamil Qetheth⁽¹⁾

Department of physics, College of science, Al Mustansiriyah University, Iraq
majedkamil@uomustansiriyah.edu.iq

Ali A. D. Al-Zuky⁽²⁾

Department of physics, College of science,
Al Mustansiriyah University, Iraq
Prof.alialzuky@uomustansiriyah.edu.iq

Basaad Hadi Hamza⁽³⁾

Department of physics, College of science, Al Mustansiriyah University, Iraq
bassaadhadi@uomustansiriyah.edu.iq

Abstract:

The identification of humans constitutes a crucial component of monitoring systems, given the significance of the timely detection of individuals. Despite advancements in people detection systems, detecting humans at long distances remains challenging. In this study, we employed the Region-based Convolutional Neural Network (RCNN) approach to training a system on images captured at varying distances between the camera and individuals. The results demonstrate promising outcomes, with the system achieving a maximum detection recall of 1 for identifying people at distances of up to 40 meters and maximum precision of 1 for identifying people at distances of up to 50 meters.

Keywords: Human Detection , RCNN , ACF, Recall, Precision, True positive, False negative.

Note: The research is based on a PhD dissertation.

Introduction:

In recent years, detecting people in video scenes in surveillance systems has found a wide range of applications, such as detecting anomalous phenomena, characterizing human gait, people counting in a dense crowd, face identification, gender, and fall classification. Attention is due to the detection of the elderly, etc[1]. The detection of humans has been a significant concern in the research on computer vision. This application of

computer vision technology involves identifying whether a human body is present in a given image or footage to promptly and precisely locate individuals. It has found widespread application in fields such as intelligent surveillance, security-assisted driving, and other domains. However, the effectiveness of human detection is impacted by several factors, including the complexity of the background, variations in lighting conditions, differences in clothing, posture, viewing angle, and other such variables. Consequently, it is often challenging to acquire high-quality image feature information, which lowers the recognition rate and detection speed. Therefore, there is a need for improvement in this area [2]. In recent years, it has become possible to detect people for observation in difficult environments, and in particular, the accuracy of people detection using Convolutional Neural Networks (CNN) has improved significantly[3].

In the current study, a region-based convolutional neural network (RCNN) technique was used to investigate the effect of different distances on human detection in video scenes and compared the results with those of an approved method (ACF).

Previous Studies:

In 2010, Chern-Horn Sim and his group proposed a scheme for detecting people in random image frames of a video sequence showing a dense scene against a cluttered background. The method used only spatial information, and a trained Viola-Jones-type local detector was used in the first image pass to identify people in a dense scene. This resulted in numerous false positives. Therefore, in the second stage, they sought to reduce the number of false positives. They presented their results in the form of receiver performance curves. For example, with a detection accuracy of 79.0%, the false positive rate is 20.3% [4].

In 2018, Tattapon Surasak and his group expanded their research on video people detection method which is directional gradient histogram or HOG by developing an application to import and detect people from videos. They used the HOG algorithm to analyze each frame of the video to find and count people. After analyzing the video from start to finish, the program creates a histogram showing how many people were found and how long the video played [5]

In 2020, Ejaz Ul Haq and his group published a robust framework for detecting and tracking people in noisy and closed environments using data augmentation techniques. In addition, they used softmax layers and the built-

in loss function to improve the detection and classification performance of the proposed model. The main attention was paid to fulfilling the tasks of detecting a person in unrestricted conditions[6] .

in 2021, A. Haider et.al. They proposed a new regression-based method for human detection from thermal infrared images. The convolutional regression network was fully designed to map the anthropogenic heat signature in the thermal image input to spatial density maps. The regression intensity map was then subsequently processed to detect and localize the human in the image. The regression-based method can detect humans with an accuracy of 99.16% and a retrieval of 98.69% [7].

in 2022, Pei-Fen Tsai et.al. They proposed the use of a thermal imaging camera (TIC) along with a deep learning model as an intelligent approach for detecting humans during emergency evacuations in scenarios with low visibility caused by smoke and fire. Using YOLOv4 technology for real-time object detection. Detection accuracy has been obtained greater than 95% for locating people in a low visibility smoke scenario was achieved at 30 frames per second (FPS) [8].

2. Methodology:

In this section will address the essential stages of building a human body detection system in digital images captured at varying distances, utilizing the RCNN technique.

2.1. The introduced human detection system:

In this study, data was captured by several videographers at various distances from 10 to 70 meters, the videos were converted to frames using the VLC media player, and a large number of frames (240 frames) (10m, 20m) were selected. , 30 m and 40 m) and introduced for system training using RCNN technology.

RCNN is an object detection model that uses large-capacity CNN to propose upstream regions for object localization and segmentation. Selective search is used to identify a large number of candidate regions (“regions of interest”) for bounding box objects, extract features from each region separately, and classify [7]. RCNN takes an input frame and uses feature maps generated by convolutional layers to suggest where features might be located [3] . This means that RCNN runs a classifier based on each sentence, checks the object existence probability, if the probability exceeds a threshold, the sentence is flagged and RCNN runs the network as a pair. means to handle another part. Using the generated feature map, we extract the bounding boxes and identify

the class with the highest probability of objects matching each bounding box [8], as shown in figure (1).

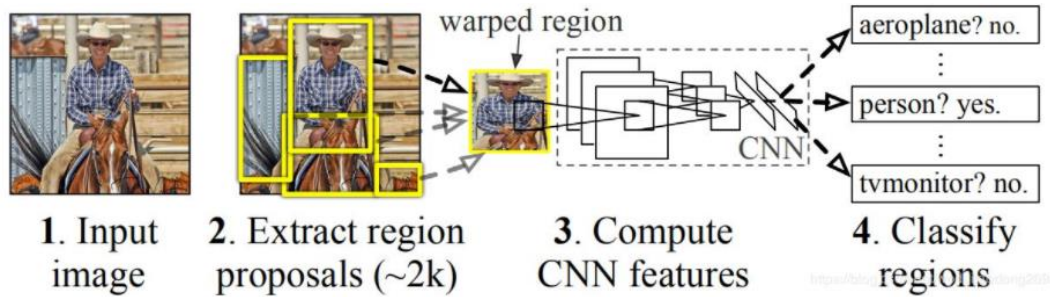



Figure 1: An illustrative diagram of the stages of R-CNN

2.2. Data collection:

In this study, the data was captured by several videographers at different distances from 10 to 70 meters, and the videos were converted into frames using VLC media player, and a large number of frames (240 frames) were selected at distances (10m, 20m, 30m and 40m) and were introduced to train the system using the RCNN technique.

Example of extracted frames as showing table (1).

Table 1: Samples of Extracted frames from videos at different d (between persons and camera)

d(m)	Images
10 meters	

<p>30 meters</p>	
<p>50 meters</p>	

2.3. Labeling and Training algorithm: human body detection:

The image labeling tool in Matlab was used to label 240 images to get a database to use in training the system.

This database was used in the process of training the system using RCNN technology to obtain a trained model for human detection.

Training is designed using the layers and parameters shown in the table (2).

Table 2: Layers and options used in training the system

Layers		Options
1	'Image Input' 50x50x3 Images with 'zerocenter' normalization	Gradient Decay Factor: 0.9000 Squared Gradient Decay Factor: 0.9990 Epsilon: 1.0000e-08 Initial Learn Rate: 1.0000e-06 Learn Rate Schedule: 'none' Learn Rate Drop Factor: 0.1000 Learn Rate Drop Period: 10 L2 Regularization: 1.0000e-04 Gradient Threshold Method: 'l2norm' Gradient Threshold: Inf Max Epochs: 20 Mini Batch Size: 32 Verbose: 1 Verbose Frequency: 50 Validation Data: [] Validation Frequency: 50 Validation Patience: Inf Shuffle: 'once' Execution Environment: 'auto' WorkerLoad: [] OutputFcn: [] Plots: 'none' Sequence Length: 'longest' Sequence Padding Value: 0 Sequence Padding Direction: 'right' Dispatch In Background: 0 Reset Input Normalization: 1
2	'Convolution' 32 3x3 Convolutions with stride [1 1] and padding [1 1 1 1]	
3	'Max Pooling' 3x3 Max pooling with stride [1 1] and padding [0 0 0 0]	
4	'ReLU'	
5	'Convolution' 32 3x3 Convolutions with stride [1 1] and padding [1 1 1 1]	
6	'ReLU'	
7	'Average Pooling' 3x3 Average pooling with stride [1 1] and padding [0 0 0 0]	
8	'Convolution' 64 3x3 Convolutions with stride [1 1] and padding [1 1 1 1]	
9	'ReLU'	
10	'Average Pooling' 3x3 Average pooling with stride [1 1] and padding [0 0 0 0]	
11	'Fully Connected' 64 Fully connected layer	
12	'ReLU'	
13	'Fully Connected' 2 Fully connected layer	
14	'Softmax'	
15	'Classification Output' Crossentropyex	

The steps of the training algorithm are as follows and as showing in figure (2):

Input: image (I_i); where $i=1, 2, \dots, N$.
 Out put: Training model (HB_Det).
 Start algorithm:
 1. **Step1:** load the input image (I_i).
 2. **Step2:** Create the people detector object (HB_Det) in (I_i) by employing the Histogram of Oriented Gradient (HOG) features and a trained Support Vector Machine (SVM) classifier, to detect incompletely visible humans in an upright posture.
 3. **Step3:** Detect people using the people detector object (HB_Det).
 4. **Step4:** Annotate detected people.
 5. **Step5:** Go to step1 and load next image.
 6. **Step6:** End algorithm.

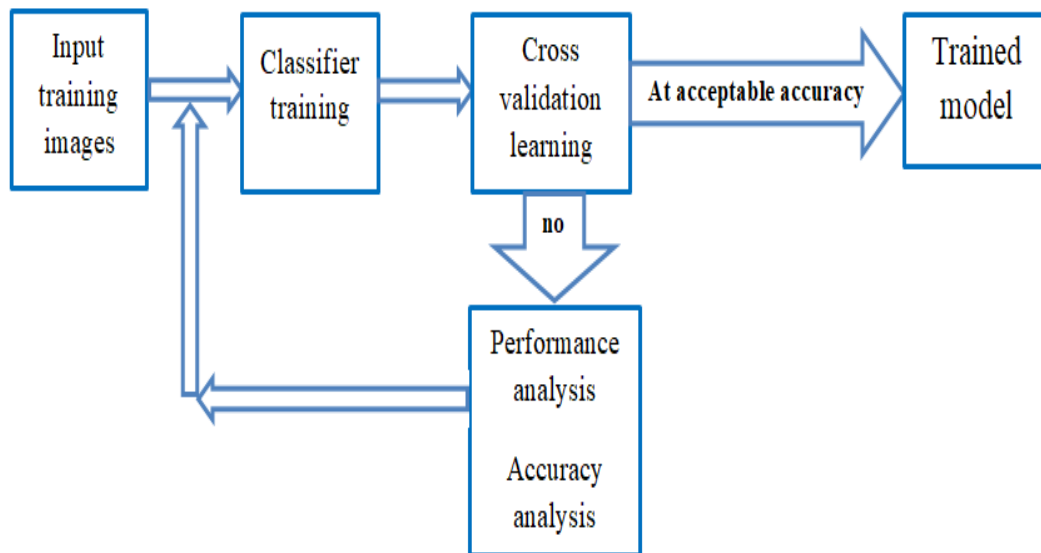


Figure 2: Trained model Scheme

2.4. Human body detection using the RCNN model algorithm

After a Human Body training model was created (HB_Det) , it was tested by applying it to detect human body within an image captured by the camera. The steps of the test algorithm are as follows and as showing in figure (3):

Input: Images (I_i); where $i=1, 2, \dots, N$.

RCNN human body detection model(HB_Det)

Output: Detect the human body (I_b: extracted body image) form (I_i).

Start algorithm:

1. Step1 for $i=1$ to N

2. Step2: Detect the human body in the (I_i) by using the detector (HB_Det).

Annotate the image with the bounding boxes BBox for the detections and

the detection confidence scores, and extracting human body images I_b using BBox and I_i.

3. Step3: end for.

4. Step4: End algorithm.

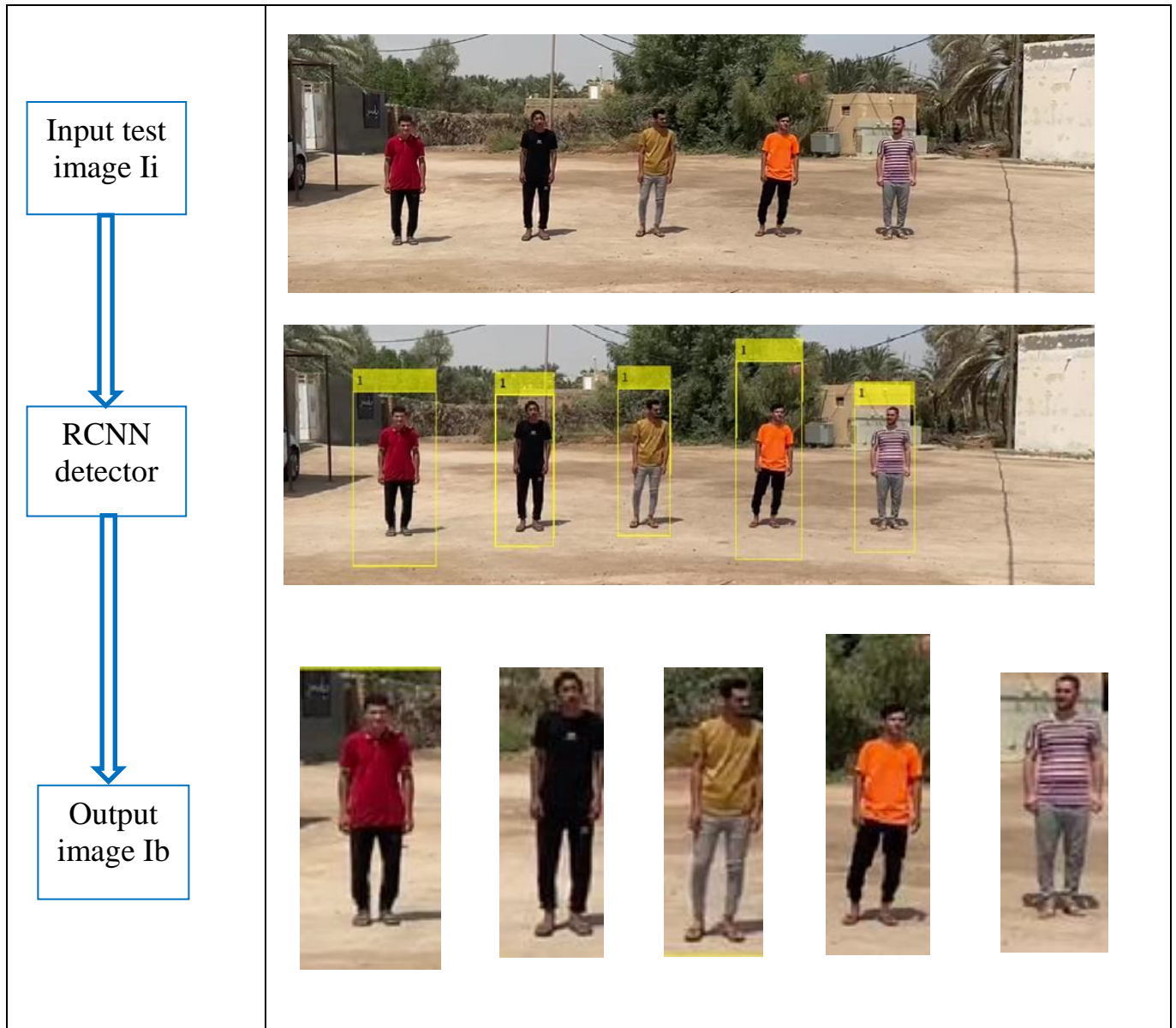


Figure 3: Testing model Scheme

2.5 Performance Evaluation: Metrics for evaluating success

In deep learning and data analysis, various metrics are used to evaluate the performance of models. Two important measures were used, Recall and Precision. Below are the equations for calculating this metrics.

$$\text{Recall} = \frac{T_p}{T_p + F_n} \quad (1)$$

$$\text{Precision} = \frac{T_p}{T_p + F_p} \quad (2)$$

Where:

- True positive (Tp): Correct detection of targets.
- False negative (Fn): Non detection of targets.
- False positive (Fp): An incorrect detection of targets.

3. Results:

After three hours of training, a model was trained using 240 images in which each image consisted of 5 individuals. The trained model was tested for its detection accuracy at distances of 10m, 20m, 30m, and 40m, which resulted in a 100% detection rate. However, as the distance increased, the detection rate decreased gradually, reaching a low of 12% at a distance of 60m. Nevertheless, the performance of the model was superior to that of the ACF method, particularly at long distances, as shown in table (3) and figures (4-6) :

Table 3: Recall for R-CNN and ACF in human detecting using 50 images for each distance.

Distance (m)		10	20	30	40	50	60	70
Recall	R-CNN	1	1	1	1	0.64	0.12	0
	ACF	1	0.98	0.39	0	0	0	0
Precision	R-CNN	1	1	1	1	1	0.94	0
	ACF	1	1	1	0	0	0	0

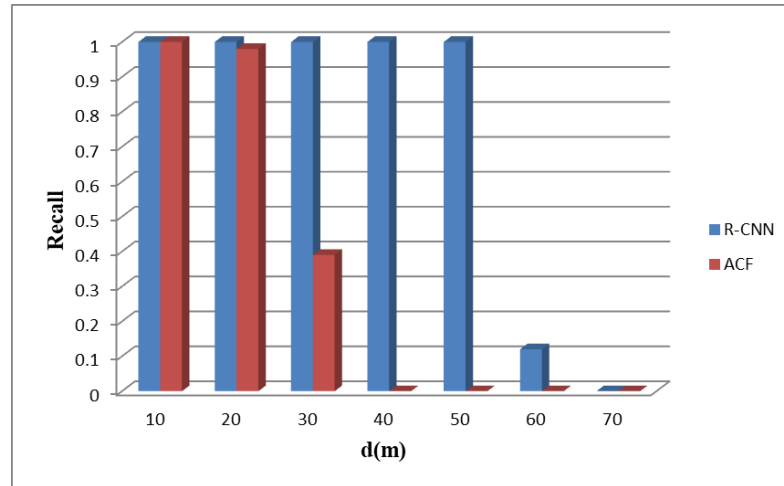


Figure 4: Comparison between recall of human R-CNN detector and recall of human ACF detector as a function of distance

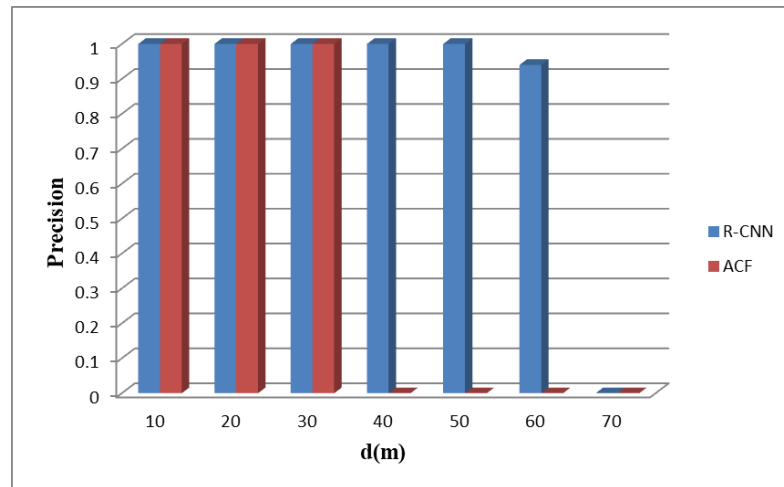
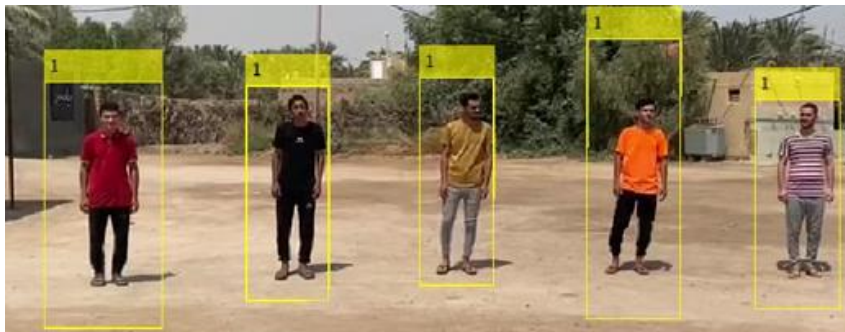


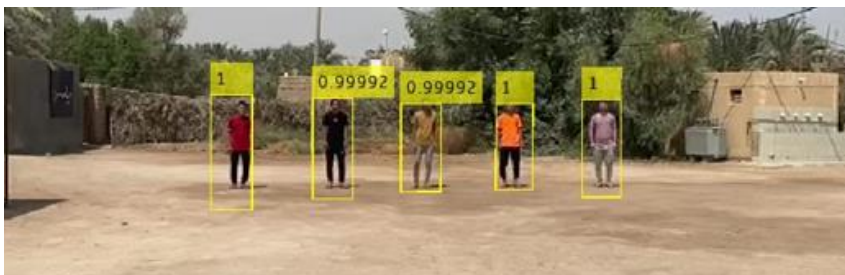
Figure 5: Comparison between precision of human R-CNN detector and precision of human ACF detector as a function of distance



10 m



20 m



30 m



40 m



50 m



60 m

Figure 6: Samples of human detection for each distance using RCNN technique.

Conclusions:

The detection results used showed that the use of images with different distances in training gives promising and good results in detection. Where images with distances (10m-40m) were used to train the system using RCNN technique, the obtained detection results were compared with the detection results of a pre-designed system, which is ACF, so the results of the used system were much better for the detection distance.

References

1. M. Paul, S. M. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications-a review," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 1-16, 2013.
2. Y. Wang, J. Wu, and H. Li, "Human detection based on improved mask R-CNN," in *Journal of Physics: Conference Series*, vol. 1575, no. 1: IOP Publishing, p. 012067, 2020.

3. J. H. Kim, H. G. Hong, and K. R. Park, "Convolutional neural network-based human detection in nighttime images using visible light camera sensors," *Sensors*, vol. 17, no. 5, p. 1065, 2017.
4. A. M. Cheryadat and R. J. Radke, "Detecting dominant motions in dense crowds," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 568-581, 2008.
5. T. Surasak, I. Takahiro, C. h. Cheng, C. e. Wang, and P. y. Sheng, "Histogram of oriented gradients for human detection in video," in *2018 5th International conference on business and industrial research (ICBIR)*, 2018: IEEE, pp. 172-176.
6. E. U. Haq, H. Jianjun, K. Li, and H. U. Haq, "Human detection and tracking with deep convolutional neural networks under the constrained of noise and occluded scenes," *Multimedia Tools and Applications*, vol. 79, pp. 30685-30708, 2020.
7. A. Haider, F. Shaukat, and J. Mir, "Human detection in aerial thermal imaging using a fully convolutional regression network," *Infrared Physics & Technology*, vol. 116, p. 103796, 2021.
8. P.-F. Tsai, C.-H. Liao, and S.-M. Yuan, "Using deep learning with thermal imaging for human detection in heavy smoke scenarios," *Sensors*, vol. 22, no. 14, p. 5351, 2022.
9. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.
10. S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B. Y. Choi, and T. R. Faughnan, "Real-time human detection as an edge service enabled by a lightweight cnn," in *2018 IEEE International Conference on Edge Computing (EDGE)*, 2018: IEEE, pp. 125-129.

دراسة تأثير المسافة على تطبيق تقنية الكشف الآلي RCNN لجسم الإنسان

بسعاد هادي حمزة⁽³⁾
قسم الفيزياء، كلية العلوم، الجامعة
المستنصرية، العراق
07704535670

علي عبد داوود الزكي⁽²⁾
قسم الفيزياء، كلية العلوم، الجامعة
المستنصرية، العراق رقم الهاتف
07706040619

ماجد كامل غثيث⁽¹⁾
قسم الفيزياء، كلية العلوم، الجامعة
المستنصرية، العراق
07806436086

bassaadhadi@uomustansiriyah.edu.iq

Prof.alialzuky@uomustansiriyah.edu.iq

majedkamil@uomustansiriyah.edu.iq

مستخلص البحث:

يشكل التعرف على البشر مكونًا حاسمًا لأنظمة المراقبة ، نظرًا لأهمية الكشف عن الأفراد في الوقت المناسب. على الرغم من التقدم في أنظمة الكشف عن الأشخاص ، فإن اكتشاف البشر على مسافات طويلة لا يزال يمثل تحديًا. في هذه الدراسة ، استخدمنا نهج الشبكة العصبية الالتفافية المستندة إلى المنطقة (RCNN) لتدريب نظام على الصور الملتقطة على مسافات متفاوتة بين الكاميرا والأفراد. تُظهر النتائج نتائج واعدة ، حيث حقق النظام أقصى استرجاع للكشف قيمته 1 لتحديد الأشخاص على مسافات تصل إلى 40 مترًا وأقصى انضباط قدره 1 لتحديد الأشخاص على مسافات تصل إلى 50 مترًا.

الكلمات المفتاحية: كشف الإنسان ، RCNN ، ACF ، الاسترجاع ، الانضباط ، إيجابي حقيقي ، سلبي كاذب.

ملاحظة : البحث مستل من اطروحة دكتوراه.