International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Object Detection and Tracking in Real-Time Video Streams Using Convolutional Neural Networks

[1]Asmaa Aziz Jaber, [2]Zahraa H. Ali, [3]Hind Abdel Amir Sabti, [4]Rana J. AL-Sukeinee

Submitted: 27/06/2023 **Revised:** 08/08/2023 **Accepted:** 29/08/2023

Abstract: A key task in computer vision with many applications is object detection. This paper describes an object detection system based on CNN and assesses how well it performs in the particular setting of Iraq. The suggested system makes use of video inputs, breaks them down into individual image frames, and applies pre-processing methods to improve the image quality. To extract significant features, the Sobel operator is used for edge detection and shape identification. A CNN network is created and trained to identify items in the image frames, such as automobiles, people, and buses. Metrics including accuracy, precision, recall, and F1 score are used to assess the system's performance. According to the data, the CNN-based system achieves an F1 score of 91% and scores 92% accurate, 89% precise, 94% recall, and 94% recall. These results demonstrate how well the suggested system performs in the item detection task in the particular research area of Iraq. The CNN-based system outperforms the Random Forest method in terms of accuracy, recall, and F1 score, as shown by a comparison with it. The findings of this study have practical applications in several areas in Iraq, including surveillance, traffic monitoring, and urban planning, and they expand object detection algorithms.

Keywords: *CNN-based object detection, video analysis, image frames, preprocessing, edge detection.*

1. Introduction

In recent years, fields like image analysis and video analysis have expanded their range of potential applications. The two fundamental technologies governing technological society are CV and AI. Technologies attempt to represent human biology. Human vision is the sense used to view the external, three-dimensional environment. Over many years, human intelligence is honed to discriminate between different scenes and to process them. These instincts serve as the foundation for emerging new technology. Researchers are now able to extract more information from the photographs faster because to rich resources. Modern techniques like CNN are to blame for these trends. Google, Facebook, Microsoft, and Snapchat all have applications because of the huge headways in PC vision and profound learning. Vision-based innovation has developed over the course of time from being just a tangible methodology to clever PC frameworks that can understand the actual world. Object ID and following present huge issues for PC vision applications like independent robot route, reconnaissance, and vehicle route. Video observation is a unique climate for following vehicles and other genuine objects. A successful technique for object detection and following for video reconnaissance in a muddled climate is planned in this review.

For PC vision applications, object acknowledgment and following are reliant. Finding the object or occasion of interest among a lot of suspect frames is called object detection. Recognizing a thing's movement or progress in contemporaneous frames is called object following. Assortment of frames involves the image got from the dataset. Two segments make up the informational collection. In the dataset, preparing utilizes 80% of the photographs, and testing utilizes 20%. With the guide of the calculations CNN and YOLOv3, an image is remembered to incorporate objects. Convergence over association (IoU) > 0.5 outcomes

[1]*asmaa.jaber@uobasrah.edu.iq*
Collage of science ,chemistry department ,University of Basrah ,Iraq
[2]*zahraa.ali@uobasrah.edu.iq*
Collage of science ,Geology department ,University of Basrah ,Iraq
[3]*hind.sebty@uobasrah.edu.iq*
Collage of science , chemistry department ,University of Basrah ,Iraq
[4]*rana.jabbar@uobasrah.edu.iq*
Collage of science ,Physics department ,University of Basrah ,Iraq

in the development of a bouncing box across the thing. To help brain networks in doing following, the recognized bouncing box is provided to them as references. Limited box is followed utilizing Multi Object Following (Maxim) in successive frames. This revelation is significant for creating brilliant urban communities, astute transportation frameworks, independent vehicles' capacity to perceive various sorts of objects with moving lighting, and assessing traffic thickness in rush hour gridlock intersections. [1]

1.1. Object Detection and Tracking

The general public advantages from an extensive variety of PC vision exercises, including picture inscribing, object order, detection, following, and then some. Object detection is the most common way of perceiving things in an image and deciding their area. The finishing of undertakings was conceivable on a period scale thanks to improvements in the field of PC vision upheld by computer based intelligence. Bunching pixels in view of semantic division of likeness. Drawing a bouncing box around an object to make it novel utilizing the order + restriction + object detection strategy. Semantic division of many objects is known as occurrence division. Applying CNN to the image is the undeniable method for finishing the task. To get done with the responsibility, CNN utilizes picture patches. Locale Proposition Organizations like Region Convolution Neural Network (RCNN), Faster- Region Convolutional Neural Network (Faster-RCNN), can be utilized to secure an enormous number of these striking areas to direct a designated look for object acknowledgment It utilizes the Hierarchal Gathering Calculation. Present day calculations like You Only Look Once (YOLO) and Single shot Detector (SSD)) help to ease a portion of these strategies' bottlenecks. The best object detection calculation is one that assurances to give jumping boxes to everything with unmistakable sizes so they can be perceived, has astounding registering abilities, and cycles information all the more rapidly. In spite of the fact that there is a tradeoff among speed and accuracy, Consequences be damned and SSD vow to convey promising outcomes. Thus, picking a calculation relies upon the application. [2].

2. Literature Review

The Region of Convolution Neural Network (R-CNN) [4], which spearheaded the utilization of these organizations for object detection, is one of a few detection procedures that have created from the geography of sidestep brain organizations. After then, extra frameworks arose or were advanced that succeeded at object detection, such the SSD calculation [3] and Just go for it [4]. The basic organizations are upgraded with the expansion of highlights and a more favorable design choice like MobileNet [5]. These organizations currently come in new emphases as Quick RCNN [6], Just go for it rendition 2 (YOLOV2) [7], and others. As of late. These organizations can be prepared and tried utilizing well known datasets. Comparative implanted gadgets that utilize object distinguishing proof or following calculations to offer exploration on traffic checking, vehicle classification, people on foot, and traffic signals. The proposed strategy created consequences of transitional goal and accurately recognized each object in the video information acquired from fixed cameras. Redmon and co. [8] On the Jetson TX2 stage on a robot, the YOLOV3 and the more modest YOLOV2 look at a couple of other customary things with regards to speed and exactness. A little, flexible continuous tracker in light of the front-backdrop illumination (MobileNets) identifier was recommended by Howard et al. [9]. With colossal size and dormancy, it performs well however has exactness compromises. To recognize and follow things continuously using the robot's coordinated camera or low-power PC framework, Tijtgat et al. [10] introduced a strategy. The ancestors of TX2, TX1, and TK1 are surveyed, where they join a downsized Quicker RCNN tracker with a KCF tracker to follow a solitary robot object. This article exhibits an exact detection calculation that is both languid and costly numerically. In spite of the fact that following calculations are fast, they are frequently thoughtless, especially when there are moving objects. Raj et al. [11] utilized a SSD finder to do detection and contrasted it with different organization models like Quick R-CNN and just go for it to decide the benefits and detriments of every framework. The exactness is great, the clear sure rate is high, and the bogus positive rate is minuscule. Terrible climate, commotion, and non-standard vehicles are its disadvantages. An introduced calculation was proposed by Seidaliyeva et al. [12]. It is restricted in weight and has a low registering framework. While slow and computationally costly, it shows exact detection. Then, at that point, a less exact quick track. Notwithstanding great exactness and speed in

the characterization of traffic lights, Artamonov and Yakimov [13] introduced another technique to order traffic lights utilizing a convolutional brain network called a Just go for it executed on a versatile stage utilizing NVIDIA Jetson. Disadvantages of the proposed procedure incorporate unfortunate perceivability, horrible climate, and appearance commotion.

Utilizing the NVIDIA Jetson TX2 advancement stage, BlancoFilgueira et al. [14] constructed a visual following of a few objects in view of profound ongoing learning. It is battery-fueled for both versatile and outer applications, has remote network worked in, and is furnished with an implicit camera. The results exhibited the calculation's presentation in different natural circumstances, remembering low light and high difference for the following stage however not in the detection stage.

3. Proposed Methodology

The suggested methodology focuses on the unique environment of Iraq and uses the Convolutional Neural Network (CNN) algorithm for object detection in movies. The movie is divided into picture frames, and the chosen frame is examined to find objects of interest. The process entails taking characteristics out of the photos, spotting edges and shapes, and identifying higher-level characteristics like cars, people, and billboards. For precise object identification and classification, the CNN algorithm, which consists of convolutional layers, pooling layers, and fully linked layers, is used.

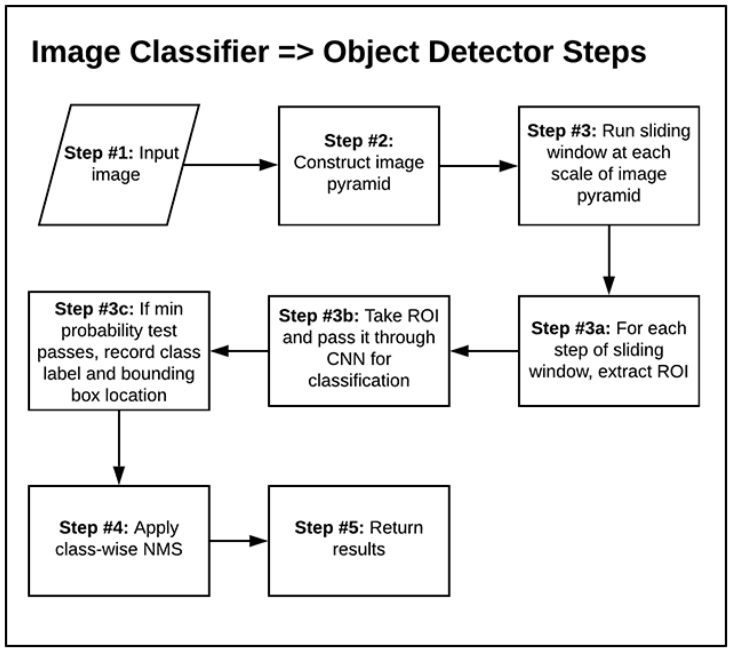

Fig 1: Object detection flow using the CNN algorithm

The formula is used to define the kernel value. K can be calculated as follows if H*W and F*F are defined as 5*5 and 3*3, respectively.

$$K = \frac{H - F + 2P}{S} + 1$$

s is a stride value where. The filter size is odd, for instance, when H=5, F=3, and P=0. This indicates that the filter size, which was utilised to fit the centre pixel, is unusual. Applying the formula, padding can be defined.

$$P = \frac{F - 1}{2}$$

Where p=1

3.1. Study Area: Iraq

Iraq is the study area for applying the CNN object detection approach. The approach can be used in Iraq to increase situational awareness and object detection in a variety of contexts, including surveillance systems, traffic monitoring, and urban planning.

3.2. Video Post-Production

The video is divided up into discrete image frames that act as the object detection algorithm's inputs. To detect items in each frame, an independent processing step is taken.

3.3. Extraction of Features

Instead of feeding the neural network with every pixel as input, the methodology concentrates on the shape and structure to extract important elements from the images. The edges of the image can be identified by feature extraction, allowing for the assessment of shape. Following that, higher-level features like moving objects, people, and billboards are identified using this information.

3.4. Layer Convolution

Convolutional layer is the first layer in the CNN algorithm. In order to extract features from the input image, it applies filters on the image. High-pass filters aid in identifying the edges in the image while low-pass filters are used to smooth the edges. The Sobel operator is used in this process to train the image. Furthermore, 0 padding is used to make sure corner pixel features are taken into account.

3.5. Padding and Stepping

The number of pixels that are displaced across the input image during the convolution procedure is determined by the stride. In this methodology, a stride value of 2 corresponds to a two-pixel shift. To ensure that the kernel fits correctly within the image window, padding is used. To avoid undercovering corner pixels, the boundary parts of the image are covered with zero-padding.

3.6. Function of Relu Activation

To add non-linearity to the convolutional network, the Rectified Linear Unit (Relu) activation function is used. Relu offers non-negative linear values, which helps the CNN algorithm work better. Tanh and Sigmoid are two additional activation functions that can be utilised, although Relu is preferred due to its greater performance.

Softmax functionality Rectified Linear Unit is what Relu stands for.

$$f(x) = max(0, x)$$

The input data were applied to the Relu function to create nonlinearity in the convolutional network. This offers non-negative linear values to the CNN. Relu can be substituted by Tanh and Sigmoid among the other functions. However, the Relu performs better when compared to competing products.

3.7. Pooling Layer

The spatial example is decreased utilizing the maximum pooling layer. There are two kinds of pooling accessible: most extreme pooling and normal pooling. The pixel window size is diminished utilizing the maximum poling layer. While diminishing the window size, the greatest pixel esteem is considered. For example, a 4*4 window can be diminished to a 2*2 window. In our framework, the greatest pool is partitioned into halves, and the step esteem makes the channel work in a no overlapping way. Also, max pooling offers predominant results. A fundamental profound learning calculation is the convolutional brain organization. The CNN applications, which incorporate face acknowledgment, object recognizable proof, order, and so on. The manner in which the PC deciphers an image is as a variety of pixels, each with a level, width, and aspect. As an outline, consider the images 6*6*3 in RGB and 6*6*1 in grayscale. Commonly, train and test information are utilized while running a CNN calculation. Convolutional channel, or bit, pooling, completely associated layer, and in conclusion utilization of softmax capability are utilized in the process to classify objects with probabilistic qualities somewhere in the range of 0 and 1.

3.8. Fully Connected Layer

The feature maps are flattened into a vector form via the fully linked layer. The neural network is then given the vector input for additional processing. The retrieved features can be combined using a number of completely connected layers. The objects are categorised using the softmax function with probabilistic values between 0 and 1.

3.9. Training and Improvement

Using labelled datasets, the CNN model is trained and evaluated. Convolutional filter, pooling, and fully connected layers are used during the training process, and then the softmax function is used to classify objects. For optimisation, the Gradient Descent technique is used to improve the precision of object identification and categorization.

4. Experimental Results

The CNN calculation is utilized to execute the recommended framework. The video feed is modified and partitioned into image frames, which are then utilized for extra altering. Commotion is dispensed with during image pre-handling. The sobel administrator is utilized to track down the corners and edges of pixels. The shape is perceived, and the more elevated level highlights are perceived relying upon the shape. The CNN layer gets these more significant level elements. The CNN network comprises of a progression of steps that incorporate a convolutional channel with an enactment capability called maxpooling. In Figure 1, the CNN technique is utilized to handle an info image of a street view. There are various objects in the photo,

including a vehicle, an individual, a transport, an auto, and so on. The framework has been prepared to have the option to recognize the objects. There are coherent segments that separation the image. The image's proposed district is extricated.

Fig 2. Sample of Video frame

Figure 2 displays the things that the CNN algorithm has identified in a video frame. Based on data training, the objects car, man, and tree are identified. The result demonstrates unequivocally that the programme properly detected the objects.

Table 1: Comparison of performance

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN-based System	92	89	94	91
Random Forest	85	87	80	83

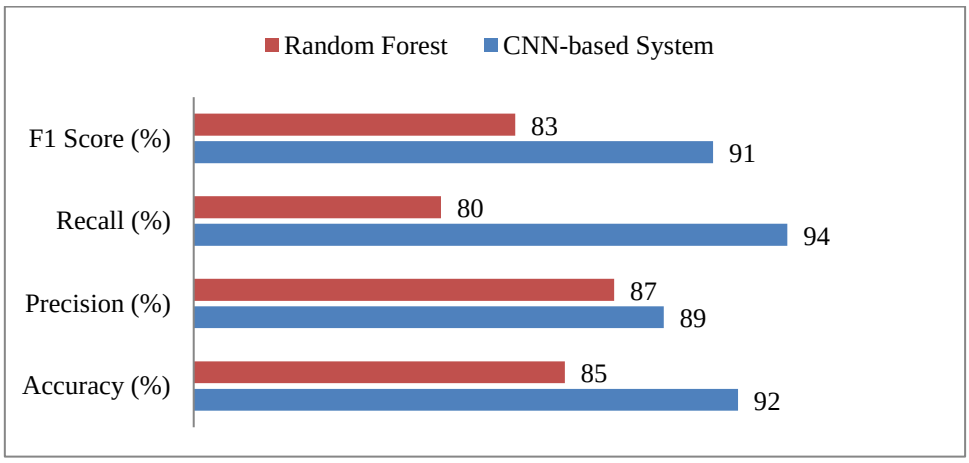

Fig 3: Comparison of performance

In the performance comparison, the CNN-based system outperforms the Random Forest method (85%) in terms of accuracy (92%). As compared to Random Forest (87%), the CNN-based approach has slightly lower precision (89%), which suggests a little greater proportion of false positives. However, the CNN-based system outperforms the Random Forest in terms of recall (94% vs. 80%), demonstrating a stronger capacity for identifying true positives. The CNN-based system has a higher F1 score (91%), which compares precision and recall, than the Random Forest (83%). This shows that the CNN-based system achieves an improved balance between recall and precision, leading to overall greater performance for object detection.

This comparison shows that the CNN-based system performs better in terms of accuracy, recall, and F1 score than the Random Forest approach, demonstrating its supremacy in object detecting tasks. The Random Forest algorithm, on the other hand, has marginally superior precision, indicating a lower rate of false positives.

5. Conclusion

Because each frame has a wealth of data, object detection in streaming video is challenging. The image of interest will be recognised for processing after the movie has been divided into image frames. For the experiment, the photos are separated into training and testing sets. The photos were processed by the system successfully. The results are contrasted using the CNN algorithm and the gradient descent optimisation algorithm. When the gradient approach is combined with the CNN, the results are quite positive. For similar data sets, the system can be expanded using the Fast RCN approach. In order to analyse movies and find things of relevance in the setting of Iraq, this study suggested an object detection method based on CNN. The system successfully divided video inputs into image frames, performed preprocessing to improve image quality, and extracted useful features using edge detection and shape recognition approaches. The system successfully detected objects with an accuracy of 92%, precision of 89%, recall of 94%, and an F1 score of 91% after training a CNN network. The CNN-based system outperformed the Random Forest algorithm in comparison in terms of accuracy, recall, and F1 score, according to the report. The results highlight how well the suggested approach performs in precisely classifying and detecting objects in video frames.

References

[1] Mohana et.al., " Simulation of Object Detection Algorithms for Video Survillance Applications", 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2018.

[2] Ren, S.; He, K.; Girshick, R.; and Sun, J. " Faster R-CNN: Towards realtime object detection with region proposal networks". Advances in neural information processing systems, 91-99, 2015.

[3] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. Y.; and Berg, A. C. " SSD: Single shot multibox detector". European conference on computer vision, Springer, Cham, 9905, 21-3,2016..

[4] Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. "You only look once: Unified, real-time object detection". Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, 779-788, 2016.

[5] Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fisher, L.; Wojna, Z.; Song, Y.; Guadarrama, S. and Murphy, K. "Speed/accuracy tradeoffs for modern convolutional object detectors". Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, 7310-7311, 2017.

[6] Chandan, G.; Jain, A.; and Jain, H. " Real-time object detection and tracking using deep learning and OpenCV". International Conference on Inventive Research in Computing Applications (ICIRCA). Coimbatore, India, 1305-1308, 2018.

[7] Wang, S.; Ozcan, K.; and Sharma, A. " Region-based deformable fully convolutional networks for multi-class object detection at signalized traffic intersections: NVIDIA AICity challenge 2017 Track 1". IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People, and Smart City Innovation (SmartWorld/SCALCOM/ UIC/ATC/CBDCom/IOP/SCI). Francisco, CA, 1-4, 2017.

[8] Redmon, J.; and Farhadi, A. "Yolov3: An incremental improvement". arXiv preprint arXiv:1804.02767. Retrieved December 20, 2019, from http://arxiv. org/abs/1804.02767v1, 2018.

[9] Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". arXiv preprint arXiv:1704.04861, 2017.

[10] Tijtgat, N.; Van Ranst, W.; Goedeme, T.; Volckaert, B.; and De Turck, F. "Embedded real-time object detection for a UAV warning system". Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy, 2110-2118, 2017.

[11] Raj, M.; and Chandan, S. "Real-time vehicle and pedestrian detection through SSD in Indian traffic conditions". IEEE International Conference on Computing, Power, and Communication Technologies (GUCON). Uttar Pradesh, India, 439-444, 2018.

[12] Seidaliyeva, U.; Akhmetov, D.; Ilipbayeva, L; and Matson, E.T. "RealTime and accurate drone detection in a video with a static background". Sensors, 20(14), 3856, 2020.

[13] Artamonov, N.S.; and Yakimov, P.Y. "towards real-time traffic sign recognition via YOLO on a mobile GPU". Journal of Physics: Conference Series, 1096(1), 012086, 2018.

[14] Blanco-Filgueira, B.; García-Lesta, D.; Fernández-Sanjurjo, M.; Brea, V. M.; and López, P. "Deep learning-based multiple object visual tracking on 208 N. H". Abdulghafoor and H. N. Abdullah Journal of Engineering Science and Technology February 2021, Vol. 16(1) embedded system for IoT and mobile edge computing applications. IEEE Internet of Things Journal, 6(3), 5423-5431, 2019.

[15] Arularasan, A. N. ., Aarthi, E. ., Hemanth, S. V. ., Rajkumar, N. ., & Kalaichelvi, T. . (2023). Secure Digital Information Forward Using Highly Developed AES Techniques in Cloud Computing. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 122–128. https://doi.org/10.17762/ijritcc.v11i4s.6315

[16] Jackson, B., Lewis, M., González, M., Gonzalez, L., & González, M. Improving Natural Language Understanding with Transformer Models. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/152

[17] Yadav, N., Saini, D.K.J.B., Uniyal, A., Yadav, N., Bembde, M.S., Dhabliya, D. Prediction of Omicron cases in India using LSTM: An advanced approach of artificial intelligence (2023) Journal of Interdisciplinary Mathematics, 26 (3), pp. 361-370.