International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Moving Object Detection Using Deep Learning Method

Tamanna Sahoo[1], Bibhuprasad Mohanty[2], Binod Kumar Pattanayak[3*]

Submitted: 16/10/2023 **Revised**: 07/12/2023 **Accepted**: 17/12/2023

Abstract: Moving object detection and tracking is the core traffic surveillance technology (TSS) . It is quite difficult to distinguish some things in a video frame series due to moving object's orientation fluctuation, changing weather circumstances, moving objects' appearance, and non-target objects in the background. Even though numerous techniques for detecting and following moving objects were developed, they were unable to produce the desired results. In this study, we use a deep learning method and trained models to create and deploy real-time object detection systems. Real-time static and moving object detection and object class recognition are capabilities of the system. The main objectives of this study were to examine and create a real-time object detection system using CNN. Since CNN has the limitation of rapid accuracy degradation after being saturated, the Softmax classifier is used in the stack layer for mitigating such performance decay. The simulation results of three video sequence of CDnet database shows almost 90% of average precision, with acceptable visual accuracy of moving object detection.

Keyword: *Traffic surveillance, real-time, CNN and Softmax.*

1. Introduction

The growth of usage of videos and images are increasing now-a-days for the purpose of acquiring and transmitting information as a result of the quick development technology of computers. The traditional manual analysis and retrieving technique is presently futile as considerable volumes in video data are produced every day. It is also susceptible to errors in judgment and visual fatigue brought on by protracted, monotonous employment. Moving object detection, being the most important work and first step towards video analysis. It is also challenging subject in the domain of computer vision research. Its technical concepts primarily reference science and technology in the areas of artificial intelligence [1], computer vision [2] and pattern recognition [3,4].

Military functions, surveillance system, smart transportation, commodities scrutiny[5], and other situations are the main examples of its economic benefit and potential applications. Conventional studies typically employ demanding artificial feature extraction procedures, like directional gradient histogram and scale-invariant feature transformation, which extracts expression information in the original input which is related to the target.

The extracted information is essential for identifying the targets, which is then used for training the classifier. The dynamic changes of the natural environment and the impact made by man-made noise poses a challenge to established approaches although the required detection algorithm [6] has generated favourable outcomes. The algorithm, for example, fails to capture the impact of additional elements because it only extracts one characteristic. Individual scenes needs to be developed with varied features because manual functions are not common, which requires considerable amounts of effort along with innovation. Some features cost a lot to compute. Further research in the field of object detection techniques in complicated situations is still required due to shortage of extremely precise detection technique.

The CNN-based target detection technology has been studied for years and is now being used extensively. In this case, it is undoubtedly essential to examine the moving target detection optimization algorithm based on CNN in this article. In order to develop CNN, a novel network, convolutional procedures and multilayer artificial neural networks (NNs) are merged [7, 8]. Using a convolution technique, the desired qualities are automatically retrieved from the original image to create additional uniformity and features that look realistic. Furthermore, it can endure some distortion.

In terms of accuracy and speed, the identification of targets based on CNN algorithms has recently made

[1]*Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India, Email: tammy.sahoo.18@gmail.com*

[2]*Department of Electronics and Communication Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India, Email: bibhumohamnty@soa.ac.in*

[3*]*Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India, Email: binodpattanayak@soa.ac.in*

enormous improvements. The CNN model can use a variety of nonlinear transformations to convert the input image into an accumulation of feature maps. when convolution and pool layers are configured alternatively. As a result, the feature maps could be classified through fully connected NN that completes the process of image recognition. The CNN network was trained by employing a backward propagation method under a careful guidance. Target detection technology focuses mostly on the challenging endeavor of categorizing, locating, and detecting objects in videos and images. The detection process in the video completes the assignment of tracking the targets after identifying the targets and their position. If convolutional neural network is utilized for learning and training the image's features that can more accurately describe the image, target detection technology can advance significantly. This can take the place of traits that were previously intentionally created using human understanding and intelligent design.

2. Related Work

Numerous target identification techniques have been suggested after years of research. Based on different technological approaches, the detection techniques is broadly split into two distinct categories: template matching and classification of images. A face detection technique that has utilized the CNN model structure, has been suggested by Lei et al. This method have earned a reputation as one of the best face identification techniques in the field of face recognition because of its tremendous resiliency from each and every direction of the face. The approach of detection as well as recognition [9] is also used in practical settings .An inventive CNN-based detection of targets method has been presented in [10]. It utilises a component detection module that minimizes amount of the computation. Thus, dividing the complicated target into various unique elements for detection.

Since convolution as well as pooling procedures are often performed alternately by CNN where the features are bit by bit abstracted from low to high level when several transformations (mostly nonlinear) are applied [11]. In the next paper[12], the researchers have employed SPP-Net which improves the speed and precision of the RCNN's detection. The challenge is that every single connection layer has restricted the size of input which is overcome by applying pooling on the spatial pyramid. In this case, an input image is not required to be cropped or enlarged, and the exactness of detection is improved. SPP-Net significantly accelerates detection by only requiring one extraction of an image's forward CNN feature. According to the paper[13], the authors have proposed the DenseNet classification model to better incorporate image characteristic information into the network. In this process, multiple blocks has been arrayed in the network as well as every single convolution layer of every block has been part of a feature map which helps in propagating through the next convolution layer. According to paper[14], CNN could be utilized to identify pedestrians. This approach uses the training samples to monitor and optimize the entire CNN. According to researchers in [15], have developed a recurrent equation that constantly modifies the weights of every Gaussian function. Hence, the total quantity of Gaussian functions applied to each pixel can be altered according to requirements. In [16], the authors have recommended combining pedestrian detection and context learning activities in order to improve CNN through multitasking training. The learnt context knowledge includes the characteristics of individuals and environments, which dramatically improves CNN's pedestrian recognition and reduces false positives.

According to authors of [17], merging of multilayer feature maps to produce multiscale super features has been recommended. This can be used to enhance the feature expression of targets, HyperNet greatly improves detection's accuracy emphasizing the value of candidate frames in the detecting process by more precise candidate frames. According to paper.[18], a different region-based technique has been proposed. To detect motion, the method looks for statistical cyclic shift moments in the image regions. Initially, due to weak anti-interference skills in results of algorithmic detection, significant deficiencies in targets may have been easily recognized.

In [19], the authors put forth the CanDiag method, a Transfer Deep Learning (TDL)-based approach to cancer detection. Deep learning (DL) is used alongwith fog computing, a real-time automatic method for diagnosing breast cancer. The Mammographic Image Analysis Society (MIAS) library's mammography images were used for the purpose, and transfer learning (TL) methods like GoogleNet, ResNet101, ResNet50, VGG16, InceptionV3, VGG19, and AlexNet were combined by means of CNN to produce better prediction results. These TDLs additionally uses techniques based on feature reduction which are principal component analysis (PCA) as well as support vector machine (SVM) classifier. In a subsequent study[20], the authors made improvements by preparing a deep transfer learning (DTL) model for a self-contained system for the detection of breast cancer using the Cancer Imaging Archive (TCIA). Transfer learning (TL) and CNN were used to pre-process the data. The proposed model's viability is then demonstrated through comprehensive simulations. Fog computing technologies are included into this architecture to further ensure the confidentiality and safety of patient data, which also lessens the demand on centralized servers and boosts output.

This article suggests a optimization strategy consisting of CNN model for detecting moving target. To reduce a great deal of calculation, the component detection module using this method divides the complicated target into many elements and recognizes each one independently. After detecting the labels of concealed variables in unlabeled data, the training model establishes the classification rules for obtaining target objects. The network is assisted in locating the target object, increasing the detection results' precision position.

3. Methodology

The architecture diagram for the suggested detector model is shown in Fig. 1. The number of features and vehicle images from CDnet database [26] are the primary inputs in the model. Then, data augmentation is the next component. Accordingly, the estimation of number of anchor boxes is one of the primary component of this newly introduced model. The detector is the last and final important component. Figure 1 presents the suggested detector model.

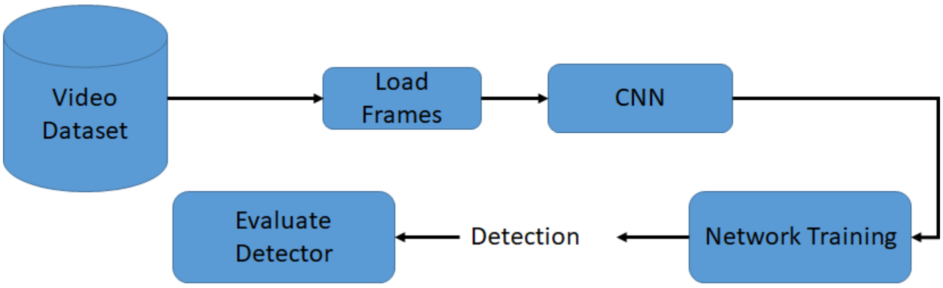

Fig.1: Block Diagram of Proposed Model.

3.1 CNN Model

During the examination of the target's geometrical or statistical characteristics, the target can be specifically located as well as segmented from the picture or video. The researchers have continuously increased their research work in deep learning by utilizing computers for collecting data and training on account of the ongoing advancements in computer hardware and Internet technologies. The best technique for detection and identification, in the opinion of many academics nowadays, is CNN-based structure [21]. The use of CNN, a powerful artificial neural network(ANN) established on the basis of traditional ANNs, has been tremendously advantageous in the domains of digital image processing as well as computer vision. It is simpler to conceive nonlinearity and create models that correspond to NN's inherent which has an advantages over conventional computer systems. Because of its excellent capacity for learning and ability to model the framework of human brain neurons, it offers a strong overall approximation for complicated nonlinear functions. Deep neural networks, often known as "deep NN," are basically multilayer perceptron networks that provide input-output mapping. The CNN structures are subset of deep NN. The use of local connections as well as weight sharing gives an advantage of network optimization while also partially refraining from overfitting and simplify training the parameters of network . The processing method of a mammalian cat's visual system acted as an example for multilayer CNN network architecture. The whole peripheral vision has been guided by the local receptive field which is next to the convolution process. The overall amount of weights in the CNN model[22] can be further decreased by performing a convolution operation on the input image that extracts image features which uses a convolution kernel that shares values depending on the local receptive field. The computer is capable of identifying the convolution kernel essential for extracting the desired features which is provided from input image by implementing theory of deep learning. However, CNN can endure some alterations and distortions. Numerous detection systems have had remarkable success finding and detecting objects according to these criteria. Using backward and forward propagation, convolutional network models are constructed. The forward propagation method is left alone as the network's output is calculated. Every single neuron analyzes its layer's outcome and transfers the layer that lies underneath it. This continues till it arrives at the output layer, when it computes the output result for the network. The weighting parameters of this network are then modified via backpropagation error. Further, iterative training has been utilized constantly for enhancing the accuracy while gradient descent is employed to minimize the set loss function [23]. The classification model flow is often depth-based. A huge number of stacked activation function layers, convolution layers, and pool layer modules are used to process the input data using CNN in order to essentially recover the image's features. The expected loss is then fed to the objective function and complete connection layer. Then, the network parameters have been modified to reflect the loss back propagation of the objective function. In the below figure 2, the NN configuration has been presented.

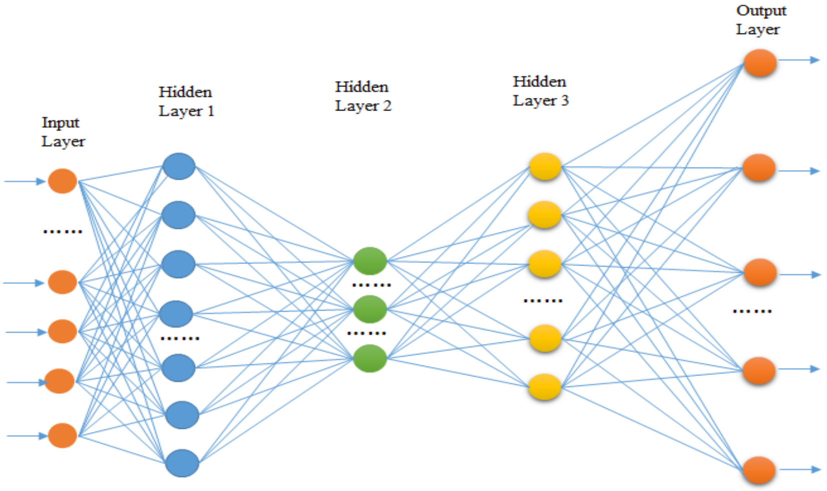

Fig.2: Configuration of NN.

The amount of time-consuming preparation steps on the original image are decreased because CNN model may input the actual image at once. Additionally, this could facilitate work and increase output. Convolution neurons have a feature called weight sharing that reduces the overall quantity of nodes and parameters. The local receptive field regulates the range, lessening the model's complexity. With the down sampling layer, the output image is obtained, and the features extracted minimizes the dimension.. The method has been created expressly for images in which weight sharing is employed to create the features of the convolution layer from the local properties of the layer that came before. By correctly modifying its structural design, it is possible to construct CNN with the regression skills that is needed for a number of different applications.

This model frequently adopts to simultaneously organize the sampling and convolution layer in the lowest part of the network that applies feature extraction along with input image's compression. The full connection layer shall be chosen as the highest level of the CNN model as it conducts various operations on the features of image that are extracted from the preceding each layer individually by processing and also by mapping, such as regression classification. An updated version of the source image that results from combining the extracted image features with feature vectors. Since artificial NN incorporates convolution and down sampling processes into its fundamental design, it is better suited for detecting images than regular NN. The features that CNN obtained have certain spatially constant characteristics.

The presented CNN-based model consists of two components. In the first stage, the model automatically recognizes the images and features. A feature identification module containing the convolutional and pooling layers is present after the first stage [24]. Utilizing the convolutional filter (kernel) provided in the convolutional layer, the feature map is generated by the convoluting earlier layers. A pass through the activation function is performed after the convolution process that is needed to obtain the activation map for the subsequent layer. By raising the pooling (subsampling) layer, the activation map lowers. The activation function uses this activation (or feature) map and adds bias to it before using it to create an activation map for the subsequent layer.

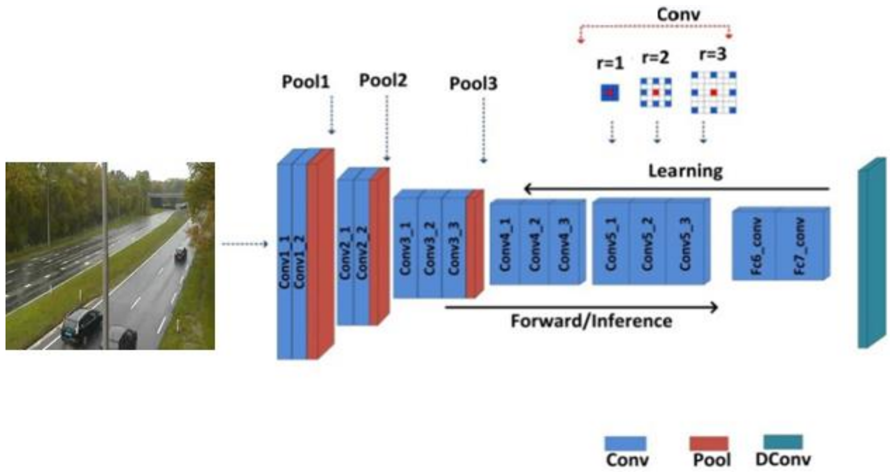

Fig. 3: CNN Architecture.

The widely used CNN image classification methods now stack many convolution layers to extract visual information in each module. Convolution computation and sampling are introduced by CNN[25] to let the computer learn on its own the target attributes from the input image. The computer benefits from good recognition results for a variety of targets in addition to good robustness to certain distortions and other changes in the environment.

The low-level CNN structure is made up of a convolution and a pool sampling layer alternatively, whereas the high-level CNN structure is made up of a complete connection layer alongwith the associated classifier. Following feature extraction, the lower structure's feature picture is sent into the first fully connected layer as its input, and the classifier is employed as the final layer's output layer. Finally, logistic regression, support vector machines, and softmax regression can all be used for classifying the images. To create a full feature graph, the model structure makes use of characteristic weight sharing to apply the same convolution kernel operation across the entire image.

This strategy will significantly decrease the range of weight parameters needed for network development. The implementation of the weight sharing operation that prominently decreases the total number of weight parameters under similar circumstances, was one of CNN's most significant developments. CNN positions the convolution layer underneath its down sampling method. After down sampling the features in the convoluted output layer, it combines the data. And it drastically lowering the hidden layer, which consists of the amount of units in between the layer of inputs and the output. Thus, feasible to minimize the overall computational complexity. The model also demonstrates some characteristics of spatial invariance. The convolution kernel's parameters is comparable to conventional NN weight parameters because they are associated to local pixels in the associated feature graph. The CNN classification's individual convolution layers are composed of a convolution kernel with several channels, therefore when convolution kernel parameters grow, the differences between large and small convolution kernels will develop substantially. Two processes are involved in the calculation of each feature map in the convolution layer: first, the feature map from the preceding layer is convolved with a variety of convolution kernels; second, the quantity of feature mappings presented in convolution layer is multiplied by the result achieved. When the outcomes are combined using a nonlinear activation function, the convolution layer's characteristic map is the final product. After network convolution, the feature space of the hidden layer is mapped to the input image. Thus, fully connected layer plays the role of a classifier by translating the representation of features in the hidden layer towards sample label space. Further, convolution layer has processed the input, the activation function must nonlinearly change the outcome, that is comparable to convolving the input following by a nonlinear connection layer. As a result, a key part of deep CNN is the activation function. The sigmoid function, presented below is defined as:

$$\text{Sigmoid(x)}=\frac{1}{1+e^{x}} \qquad (1)$$

It is common practice to utilize the continuous, smooth, monotonic function curve as the threshold function of NNs as its value occurs from 0 to 1. As a result, classification issues can be used it as a probability distribution function. The Tanh function, whose convergence rate is quicker than that of the Sigmoid function, solves the issue of the nonzero center of the Sigmoid function. It is explained as follows:

$$\text{F(x)=tanh(z)}=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}} \qquad (2)$$

Following equation (3) is a definition of the ReLU function:

$$\phi(x)=\text{max(x,0)} \qquad (3)$$

The results of this above activation function is completely zero when the input signal is less than 0. These input and output are linearly identical when the signal is bigger than 0. In order to simplify the model's computations and increase the resolution of image properties, the CNN network model's sample layer, sometimes denoted as the pool layer, is essential. This features are helpful in one section of an image making it suitable for various other parts of the image since the images are static. The pool layer's unique operation method aggregates the features of small locations on the preceding feature map in order to explain information in the big picture.

In most cases, the sampling layer follows the convolution layer and statistical procedures are applied to the tiny collection of extracted visual information. Since the back propagation is static, the features of the sample layer have no impact on how it changes. By deleting unnecessary samples from the feature graph, the pool layer is able to down-sample as well as minimize the number of parameters. Maximum pooling and mean pooling are the two most used pooling methods in the CNN model. In general, the maximum pool operation is mostly used by the CNN classification model. CNN steadily transforms from low to high level features while using a range of convolution layers and pool layers to extract the features from the input image.

Weights connect every neuron between layers, and the entire link layer is often placed at the end of a CNN. The way neurons are connected in conventional NN is the same as this. The connection layer incorporates the classifier for every component of the model. By connecting each node's input in the complete connection layer to every node's output in the layer prior to it, the features of image acquired in the preceding convolution and pool layer are transferred to the sample's mark space. Every neuron in the preceding layer is connected to every neuron in the entire connected layer after feature extraction, which identifies the complete connected layer. These high-level properties of the connection layer can all be mapped to the specific functions of the output layer.

The primary objective in the connecting layer is to integrate the two-dimensional feature map's data into a single one-dimensional feature vector. The classifier uses the retrieved one-dimensional feature vector to classify the image. The ultimate output structure is called the output layer., which is identified as the deep learning network's output, and the regression function is also used to assess the picture classification category. The training procedure is convoluted and redundant, and the scale of the usual CNN parameters is tremendous. The following instances demonstrate how to utilize the Softmax classifier:

$$\mathrm{Y}(\mathrm{x_i})=\frac{exp(x_i)}{\sum_{i=1}^{M}(exp(x_i))} \qquad (4)$$

4. Result and Discussion

Throughout this research, various trials have been carried out to assess the efficiency of the proposed method to locate the moving objects. This work presents the outcomes for three video sequences from CDnet [26] database.

The system that the proposed model was put into use on has the following features: GPU learning was accomplished using MATLAB, and the Deep Neural Network library (CuDNN). Up to 60% of the dataset was divided into training images, 10% into validation images, and 30% into testing images. The modified CNN model with an initial learning rate of (σ) and 60 epochs. The detector's mini-batch size is set to 64.

Figure 4 depicts the detection results for skating sequence for frame 1953 in which major challenges like dynamic background (snow) is present as well as a tree is occluding the path of moving objects. The results shows that proposed method provides significantly better output than ViBe[27], TD-2DDFT[28], and TD-2DUWT[29] and BSUV-NET[30], as it only detects the moving objects while other techniques are detecting the snow as well as a false detection of tree and other human beings. This makes the detection results very unclear about the moving objects of current frame in case of state-of-art methods.

Three people wearing caps and gloves are clearly detected as moving objects by the proposed technique, which correctly detects trees as stationary objects. It is intriguing to note that the suggested technique can get rid of the dynamic background caused by the constant snowfall throughout the entire video sequence. The two people are difficult to tell apart against the black backgrounds because of the low contrast. As a result, in Fig. 4(g), the proposed technique likewise fails to display both people's whole contour.

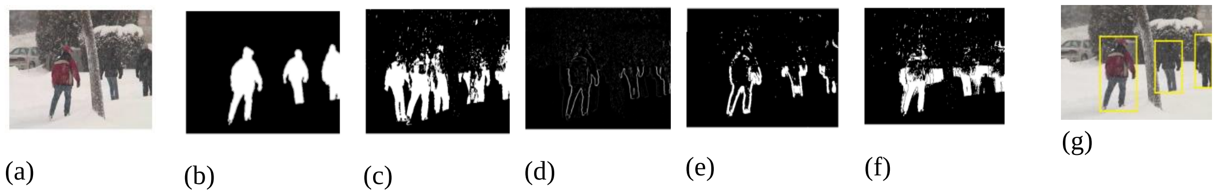

Fig. 4: Visual Comparison of Skating video sequence for frame number-1953:(a)Input frame(b) Ground-truth frame; (**c**) result of ViBe; (**d**) result of TD-2DDFT; (**e**) result of TD-2DUWT; (f) result of BSUV-NET; (g) result of proposed method.

The outcomes of frame 502 from the tramcrossroad sequence are shown in Figure 5. The original image of frame 502 is shown in fig. 5(a), and its reference image is shown in 5(b). The video frame that was selected is a complicated one since it shows a large number of moving small-sized cars against a static background of high-rising structures and leveling roadways (see Fig. 5(a)). A close examination of the frame discloses that the small moving cars are actually impossible to tell apart from the background. However, the dataset's ground truth reference frame only includes the three moving autos that are easily visible and recognisable. According to the ViBe results in Fig. 5(c), it incorrectly depicts the road and the cars.

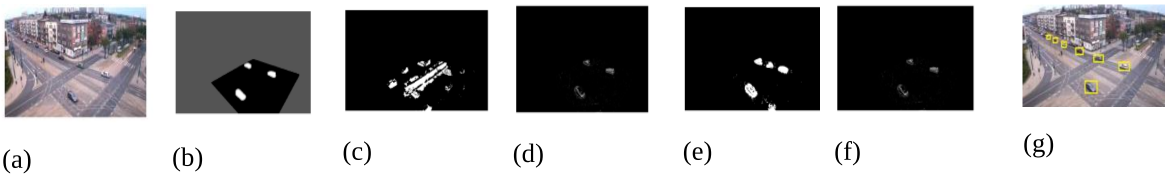

Fig. 5: Visual Comparison of Tramcrossroad video sequence for frame number-502:(a)Input frame(b) Ground-truth frame; (c) result of ViBe; (d) result of TD-2DDFT; (e) result of TD-2DUWT;(f) result of BSUV-NET; (g)result of proposed method.

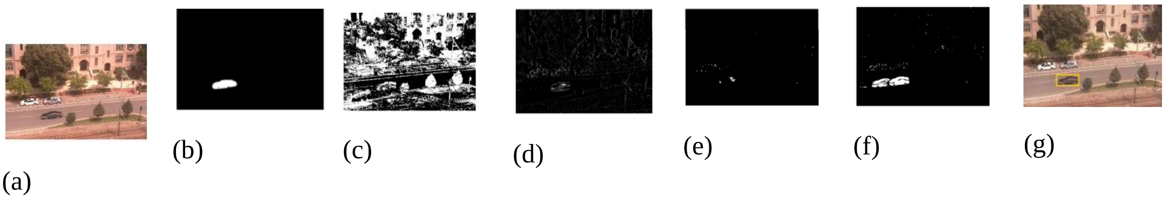

Fig. 6: Visual Comparison of Intermittent pan video sequence for frame number-1581: (a)Input frame(b)Ground-truth frame; (c) result of ViBe; (d) result of TD-2DDFT; (e) result of TD-2DUWT ;(f) result of BSUV-NET; (g) result of proposed method.

The above figure 6 represents the detection results of intermittent pan sequence where the camera is moving on its own. The frame background is very complex; only one black car is the moving object. The detection of this moving object is very challenging and almost all algorithms fail to detect the objects. The proposed methods: ViBe, TD-2DDFT, BSUV-NET; and IDS-2DUWT suffer a lot from the presence of static lines and ghostly appearance. There are several false detections or the presence of static lines and ghostly appearances in ViBe, TD-2DDFT. This disadvantage of the state-of-art background subtraction method has been overcome in the proposed method.

The performance of the proposed method on two video sequences is evaluated using four standardized measures: recall, precision, F1-measure, and specificity.

The equations are given below:

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \qquad (5)$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \qquad (6)$$

$$\text{F1-Measure}=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}} \qquad (7)$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN+FP}} \qquad (8)$$

$$\text{Average Precision}=\frac{\sum precision}{recall} \qquad (9)$$

The quantitative performance measurement of all the above sequences from the CDnet database [24] employed for the experimentation is shown in Tables 1-3. According to the QPM results, the proposed method performs better in both the video sequences.

TABLE 1 Comparison of proposed method and state-of-art methods for skating video.

Method	Recall	Average Precision	F1-Measure	Specificity
ViBe[27]	0.7210	0.7779	0.7394	0.7816
TD-2DDFT[28]	0.7654	0.6205	0.6914	0.6257
TD-2DUWT[29]	0.6266	0.7287	0.8332	0.5999
BSUV-NET[30]	0.8203	0.8113	0.7868	0.9946
Proposed Method	0.8956	0.8984	0.8442	0.8943

TABLE 2 Comparison of proposed method and state-of-art methods for tramcrossroad video.

Method	Recall	Average Precision	F1-Measure	Specificity
ViBe[27]	0.6095	0.8821	0.6692	0.7924

TD-2DDFT[28]	0.7939	0.8300	0.6924	0.7295
TD-2DUWT[29]	0.5940	0.7574	0.8751	0.8036
BSUV-NET[30]	0.8179	0.8319	0.7986	0.9944
Proposed Method	0.8703	0.9084	0.8841	0.8687

TABLE 3 Comparison of proposed method and state-of-art methods for intermittent pan video.

Method	Recall	Average Precision	F1-Measure	Specificity
ViBe[27]	0.6490	0.7487	0.8596	0.8192
TD-2DDFT[28]	0.5875	0.8564	0.5515	0.7807
TD-2DUWT[29]	0.7496	0.6212	0.7394	0.8693
BSUV-NET[30]	0.8113	0.7535	0.7668	0.9904
Proposed Method	0.8345	0.9182	0.8606	0.8941

5. Conclusion

Moving object detection is very difficult method with many challenges like scale differences, local occlusion, attitude changes, and complex scenes will considerably affect the detection of targets. We have utilized the CNN model for moving object detection to get high-performance results. In this paper, the utility of the CNN scheme as a moving object detection technique has been presented in various video sequences of change detection dataset which have above said complexities. The simulation results of three video sequence of CDnet database shows almost 90% of average precision, with acceptable visual accuracy of moving object detection.

In our future work, we intend to use deep learning models to detect various vehicle types in different datasets and also by cascading numerous CNNs in the model that may gradually classify images from complex to simple, and also simplifying the training.

Reference

[1] Y. Lin, F. Fan, J. Zhang et al., "DHI-GAN: improving dental based human identification using generative adversarial networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 1, pp. 1–13, 2022.

[2] Z. Zhang, Y. Ding, X. Zhao et al., "Multireceptive field: an adaptive path aggregation graph neural framework for hyperspectral image classification," Expert Systems with Applications, vol. 217, Article ID 119508, 2023.

[3] W. Cai, X. Ning, G. Zhou et al., "A novel hyperspectral image classification model using bole convolution with three direction attention mechanism: small sample and unbalanced learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–17, 2023.

[4] Y. Ding, Z. Zhang, X. Zhao et al., "Unsupervised self correlated learning smoothy enhanced locality preserving graph convolution embedding clustering for hyperspectral images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–16, 2022.

[5] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang, and B. Yang, "A real-time ATC safety monitoring framework using a deep learning approach," IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 11, pp. 4572–4581, 2020.

[6] J. Zhang, Z. I. Ye, X. Jin, J. Wang, and J. Zhang, "Real-time traffic sign detection based on multiscale attention and spatial information aggregator," Journal of Real-Time Image Processing, vol. 19, no. 6, pp. 1155–1167, 2022.

[7] X. B. Jin, Z. Y. Wang, J. L. Kong et al., "Deep spatio-temporal graph network with self-optimization for air quality prediction," Entropy, vol. 25, no. 2, p. 247, 2023.

[8] X.-B. Jin, Z.-Y. Wang, W.-T. Gong et al., "Variational bayesian network with information interpretability filtering for air quality forecasting," Mathematics, vol. 11, no. 4, p. 837, 2023.

[9] C. Lei, D. H. Ma, H. Q. Zhang, and Lm Wang, "Moving target network defense effectiveness evaluation based on change point detection," Mathematical Problems in Engineering, vol. 2016, no. 6, Article ID 6391502, 11 pages, 2016.

[10] H. R. Roth, L. Lu, J. Liu et al., "Improving

computer-aided detection using CNNs and random view aggregation," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1170– 1181, 2016.

[11] D. P. Tran and V. D. Hoang, "Adaptive learning based on tracking and Re-Identifying objects using convolutional neural network," Neural Processing Letters, vol. 50, no. 1, pp. 263– 282, 2019.

[12] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specifc object detection: a survey," IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 84–100, 2018.

[13] T. Long, Z. Liang, and Q. Liu, "Advanced technology of high resolution radar: target detection, tracking, imaging, and recognition[J]," Science China Information Sciences, vol. 62, no. 4, pp. 1–26, 2019.

[14] Z. J. Ruf, D. B. Lesmeister, C. L. Appel, and C. M. Sullivan, "Workflow and convolutional neural network for automated identification of animal sounds," Ecological Indicators, vol. 124, no. 2, Article ID 107419, 2021.

[15] F. Tang, X. Zhang, S. Hu, and H. Zhang, "Convolutional features selection for visual tracking," Journal of Electronic Imaging, vol. 27, no. 3, p. 1, 2018.

[16] V. Andrearczyk and P. F. Whelan, "Using filter banks in Convolutional Neural Networks for texture classification," Pattern Recognition Letters, vol. 84, no. 11, pp. 63–69, 2016.

[17] M. Kumar and C. Bhatnagar, "Hybrid tracking model and GSLM based NN for crowd behavior recognition[J]," Journal of Central South University, vol. 24, no. 9, pp. 147–157, 2017.

[18] J. Li, G. Li, and H. Fan, "Image refection removal using end-to-end convolutional neural network," IET Image Processing, vol. 14, no. 6, pp. 1047–1058, 2020.

[19] A.Pati, M. Parhi, B.K Pattanayak, B. Sahu, S. Khasim; "CanDiag: Fog Empowered Transfer Deep Learning Based Approach for Cancer Diagnosis." Designs. 2023 Apr 23;7(3):57.

[20] Pati, M. Parhi, B.K Pattanayak, D. Singh, V. Singh, S. Kadry, Y.Nam, B.G Kang, "Breast Cancer Diagnosis Based on IoT and Deep Transfer Learning Enabled by Fog Computing. Diagnostics." 2023 Jun 27;13(13):2191.

[21] N. Liu, Y. Xu, Y. Tian, H. Ma, and S. Wen, "Background classification method based on deep learning for intelligent automotive radar target detection," Future Generation Computer Systems, vol. 94, no. MAY, pp. 524–535, 2019.

[22] J. Wang, X. Wang, K. Zhang, Y. Cai, and Y. Liu, "Small UAV target detection model based on deep neural network," Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University, vol. 36, no. 2, pp. 258–263, 2018.

[23] C. Wang, J. Zheng, and J. Bo, "Deep NN-aided coherent integration method for maneuvering target detection," Signal Processing, vol. 182, no. 9, Article ID 107966, 2021

[24] V. Patnaik, M. Mohanty, and A. K. Subudhi, "Identification of healthy biological leafs using hybrid-feature classifier," *The Imaging Science Journal,* vol. 69, pp. 239-253, 2021.

[25] Y. Liang, W. Peng, Z.-J. Zheng, O. Silvén, and G. Zhao, "A hybrid quantum–classical neural network with deep residual learning," *Neural Networks,* vol. 143, pp. 133-147, 2021.

[26] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 387-394.

[27] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing,* vol. 20, pp. 1709-1724, 2010.

[28] V. Crnojevic, B. Antic, D. Culibrk, "Optimal wavelet differencing method for robust motion detection," In Proceedings of the 16th IEEE International Conference on Image Processing (ICIP 2009), Cairo, Egypt, 7–12 November 2009; pp. 645–648.

[29] F.Cheng, Y.Chen, "Real time multiple objects tracking and identification based on discrete wavelet transform," Pattern Recognit., 39, 1126-1139,2006.

[30] M. O. Tezcan, P. Ishwar and J. Konrad, "BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Videos," *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, USA, 2020, pp. 2763-2772, doi: 10.1109/WAC V45572. 2020.9093464.