Detection and Recognition of Moving Video Objects: Kalman Filtering with Deep Learning

Hind Rustum Mohammed[1]
Faculty of Computer Science and Mathematics
University of Kufa (UoKufa)
Najaf, Iraq

Zahir M. Hussain[2]
Computer Science and Mathematics, UoKufa, Iraq
Professor, School of Engineering, Edith Cowan University
Joondalup, Australia

***Abstract*—Research in object recognition has lately found that Deep Convolutional Neuronal Networks (CNN) provide a breakthrough in detection scores, especially in video applications. This paper presents an approach for object recognition in videos by combining Kalman filter with CNN. Kalman filter is first applied for detection, removing the background and then cropping object. Kalman filtering achieves three important functions: predicting the future location of the object, reducing noise and interference from incorrect detections, and associating multi-objects to tracks. After detection and cropping the moving object, a CNN model will predict the category of object. The CNN model is built based on more than 1000 image of humans, animals and others, with architecture that consists of ten layers. The first layer, which is the input image, is of 100 * 100 size. The convolutional layer contains 20 masks with a size of 5 * 5, with a ruling layer to normalize data, then max-pooling. The proposed hybrid algorithm has been applied to 8 different videos with total duration of is 15.4 minutes, containing 23100 frames. In this experiment, recognition accuracy reached 100%, where the proposed system outperforms six existing algorithms.**

***Keywords*—Convolution Neural Network (CNN); Kalman filter; moving object; video tracking**

I. Introduction

The problem of detection and recognition of moving objects in deep learning lies in detecting the location of the object and segmentation with removing its background [1]. The recognition model requires object's image without background and a correct categorical label that enables the model to predict the correct location and label the moving object [2].

When addressing the recognition mission, the first important issue to consider is to arrange the class that can be recognized. It is very important to organize knowledge at various levels, and this issue has taken a great interest in Cognitive Psychology, for example in Brown's work, a cat cannot only be thought of as a cat, but a quadruped, boxer, or in general an animated being. Cat is the term in the level of semantic hierarchy that comes to mind most easily, which is by no means accidental [3, 4]. Experimental results revealed that there is a basic level in human categorization. The trouble of object-recognition in digital images have been in the core of computer-vision research a lot of time ago [5]. Past Pascal VOC challenge and ongoing Image-Net wide Scale Visual-Recognition Challenge (ILSVRC) have united significant operations required for the solution of this matter in video scenes. Growing methods in learning with applications in patient's monitoring and health care such as in [6, 7].

This paper is motivated by the need for higher accuracy in the process of finding objects and components within a video, despite obstacles like the distance being detected by the camera or blurring of the image while the object is moving, where these factors contribute to errors in existing techniques.

This paper is organized as follows: Section 2 deals with related work, while the Section 3 focuses on special details required for theoretical background. Section 4 discusses the dataset, and Section 5 explains the proposed work, with results in Section 6.

II. Related Work

The problem addressed consists of two directions: object localization and object recognition. In [8] (2019) the authors presented a deep neural network algorithm based on the effect of visual variation applied on iLab 20M dataset of toy vehicle objects under variations of lighting, viewpoint, background, and focal setting. The experiment results on 1.751 million images from iLab 20M showed significant improvement in accuracy of object detection: DenseNet: 85.6% to 91.6%, 86.5% to 90.71%, AlexNet: 84.5% to 91.6%). CNN improves variation learning as it is capable of noticing special features and has better learning of object representations, where it decreased detection error rate of ResNet by 33%, Alexnet by 43%, and DenseNet by 42. The author in [9] presented a method for recognizing video objects of interest by applying the global Label Distribution Protocol (LDP) then applying the Speeded up Robust Features (SURF) detector. Finally, the objects in the videos are compared and matched with the objects of interest. In [10] the Authors presented the low performers (developmental prosopagnosics (DP)) by the Cambridge Face Perception Test (CFPT) and Cambridge Face Memory Task (CFMT) as signification methods for detection and matching of human face in real time video based on VG factor in visual domain, where they examined the performance of 14 individuals whom were contacted because they are more experienced than their peers at detection and recognizing faces who scored over 90% correct on the online version of CFMT Aus. In [11] the Authors presented a tool for detection and recognition of new object sample in a video by applying the Hybrid-Incremental Learning method (HIL) with Support Vector Machine (SVM), which can improve the recognition ability by learning new object samples and new object

concepts during the interaction with humans. This hybrid technique improves the recognition quality by minimizing the prediction error. In [12] they worked on removing the background of an image for enhancement of detection and recognition using CNN model of recognition.

III. Background

Machine learning is about patterns study that gives computers the capability to learn without being explicitly programmed, where during the training phase the machine learns how to build models and algorithms to predict new data. The most important type of neural networks in deep learning is convolutional neural network (CNN), which is specifically designed for recognition and detection of images.

The author in [13] shows that the performance of various pooling methods used in detecting pictorial objects can be obscured by several confounding factors such as the link between the sample cardinality in the spatial pool and the resolution at which low-level features have been extracted. The authors provide a detailed theoretical analysis of max pooling and average pooling, and give extensive empirical comparisons for object recognition tasks. The author in [14] explains Restricted Boltzmann machines using binary stochastic hidden units, where these can be generalized by replacing each binary unit by an infinite number of copies all having the same weight but have progressively more negative biases. The learning and inference rules for these "Stepped Sigmoid Units" are unchanged. They can be approximated efficiently by noisy rectified linear units (ReLU's). Compared with binary units, these units learn features that are better for object recognition on the NORB dataset and face verification on the Labeled Faces in the Wild dataset. Unlike binary units, rectified linear units preserve information about relative intensities as information travels through multiple layers of feature detectors.

CNN contains many layers of networks that extract features of an image and detect the class for which it belongs; of course, after it is trained with a set of standard images. The architecture and work of CNN is detailed in [15,16].

CNN contains layers, in every layer we convert one size to another through a differentiable procedure. Three types of layers construct a CNN: Convolutional Layer, Pooling Layer, Fully Connected Layer. The first stage is the convolutional layer which processes the image to extract only salient features in it. Filtering the input image to produce the feature map or activation map [17]. Convolution is pure mathematical method achieved in three stages: the first stage is sliding mask feature and image matching patch, second stage is to multiply each input image part (of size equals the mask size) with mask, third stage is the sum of all these multiplications to find the average, then filling the result in a new matrix of features [18].

The second part in the structure of CNN model is the pool operation to shrink the input image matrix for every feature acquired from previous step (convolutional). This is achieved by first selecting appropriate mask size 2 or 3, then selecting a stride moving area of image pixels, usually 2, then sliding the mask over convoluted images [19]; and finally, picking the maximum value form every mask. The third part in the structure of CNN model is the ReLU activation function, through which we pass the pooling result where every pixel that is less than zero will be nulled.

The final part in the structure of CNN model is the classification layer, which is a fully connected layer in which we decide the label of the input data, decided based on the highest voted category [20]. The layers of the convolution network can decrease the error of classification using back propagation to produce best prediction. Layers are passed through multiple times (in iterative fashion) [21].

IV. The Dataset

In this work we build a dataset containing 1000 images from animal and human images from CIFAR-10 database, where each image has the size of 100*100 pixels.

V. The Proposed Approach

In this work we build a dataset containing 1000 images from animal and human images from CIFAR-10 database, where each image has the size of 100*100 pixels.

A. Detection of Moving Object with Kalman Filter

The approach uses repetitive prediction and correction to compute the correct location of the object by comparing the current state with previous state of the object recorded in the history of guessing. Kalman filter algorithm is simply the best tool here as it is based on recursive procedure. Kalman filter has a measurement relation and a state model as shown in the following equations:

$$s(t) = O(t-1)\, s(t-1) + w(t) \quad (1)$$

$$z(t) = H(t)\, s(t) + v(t) \quad (2)$$

where v(t), w(t) are noise processes of Gaussian distribution with zero average, H(t) is measurement matrix, and O(t-1) is the state transition matrix. After collecting videos of multiple objects, the following Algorithm for detection and tracking of objects is presented. Table I clarifies the details of the steps used in this algorithm.

TABLE I. Detection and Background Removal Algorithm

Step1:	Prediction and correction by the Kalman filter.
Step2:	A) Compute the variation of intensity between current frame and next frame. B) Select a threshold.
Step3:	If threshold is less than variation between frames, continue.
Step4:	Compute the centroid of object.
Step5:	Create mask of detecting and acquire the position of match and move to analysis.
Step6:	Move the object vertically or horizontally based on certain dimension, where the movements of object in anticlockwise or clockwise.
Step7:	Return the location of object.

The threshold operation helps in separation of foreground and background form the frame, where it is enabled to build a mask for moving objects. It is a requisite to distinguish between noisy clutters motion and real motion of objects. Object to be tracked and detected have features like shape, edge, color, and boundaries. Procedure of tracking movement of object based on observed point in current frame using the previous frame is shown in Fig. 1.

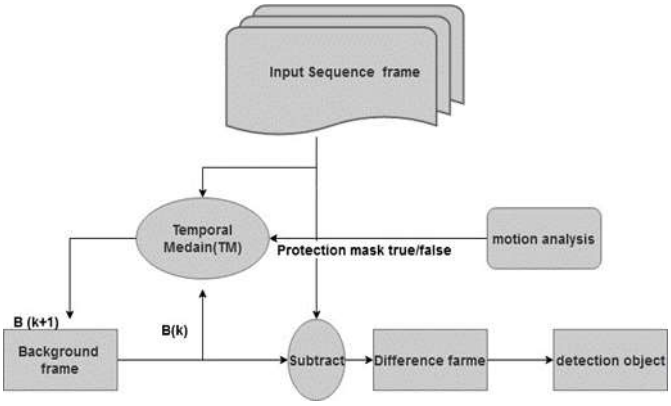

Fig. 1. Detection and Background Removal of Moving Objects.

B. Recognition of Moving Object by CNN

After detecting the object and cropping by Kalman filter, the CNN is applied for training and building a model capable of predicting the new object. In this work each color image consists of three bands (red, green, blue), where the CNN size can be arranged to handle color images; for example, when the color image is of 50*50 size, the hidden layer in CNN model would be of size 50*50*3. When we scale the color image to 100*100, then we need 100*100*3=30000 weights in the hidden layer. CNN input can solve scalability problem. The Layers of CNN have neurons arranged in three dimensions (Height, Width, and Depth) as shown in Fig. 2.

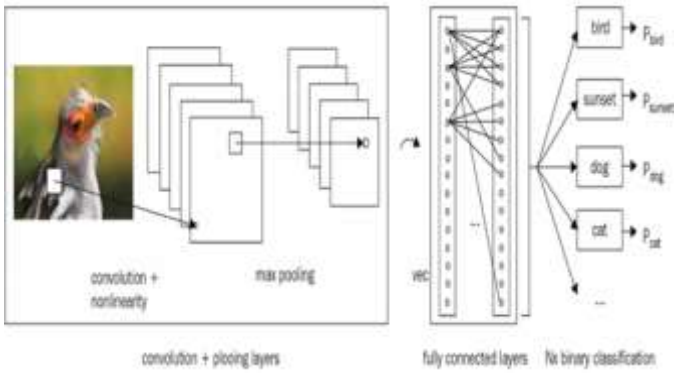

Fig. 2. Operation of CNN.

VI. Experiments and Results

To validate the proposed system, we downloaded 8 different videos from Google, the total video time is 15.4 minutes, each consists of 23100 frames, contains animals and humans. Two basic steps are applied, the first step is for the detection and tracking of the object then removing background by Kalman filtering, the second step is applying the CNN to recognize the object.

A. Detection and Tracking of an Object by Kalman Filter

We apply Kalman filtering for detecting the moving objects. Kalman filter help in three functions: guessing the future location of object, decreasing of noise coming from faulty detections, and facilitating the operation of associating multiple objects to their tracks. The first step is to load the video by using vision.Videoplayer function of MATLAB. After extracting frames of a video, we find the prediction (future position) and correction (error) by the Kalman filtering, then we compute the variation of intensity between current frame and next frame, and select a threshold (the threshold should be less than variation between frames). After that we compute the centroid of object and create a mask of detection and position of matching, move to video components and define object movement, where the object moves vertically or horizontally based on center dimension, where movements are in anticlockwise or clockwise.

B. Recognition of Moving Object by CNN

After detecting the location of object and removing the background of frame, then CNN is introduced for a training model. The size of object is 100*100*3 and the training setup of CNN is as follows. Learning rate is 0.001 and momentum is 0.9. The network architecture consists of four convolutional layers and four pooling layers, followed by two fully connected layers. Each convolutional layer is followed by a ReLU layer, which is an effective activation function to improve the performance of the CNNs. Regularization with the weight decay $5\times10-4$ is used in the network training. The dropout ratio is set as 0.5. The learning rate is initially set as 0.001 and the training is stopped after 1000 epochs. After building the network architecture, then we train CNN model. After building the CNN training model, the video will be processed for detection and background removal. The last process is to classify the moving object using the hybrid technique of Kalman filtering followed by CNN, which achieved (in this experiment) accuracy of 100%. The proposed system has been shown to outperform six other works, as in the Table II.

TABLE II. Comparisons with Existing Methods

Authors	tools	accuracy
This work	CNN + Kalman filter	100%
Muhammad [22]	SVM classifier	97.95%
Bruno et al. [23]	ANN and k-NN	97%
Tian et al. [24]	Hierarchical Filtered Motion	94%KTH Human Action Dataset
Modarres et al. [25]	Body Posture Graph	94%KTH Human Action Dataset
Sheng et al. [26]	HOG Feature Directional Pairs	94.99 % KTH Human Action Dataset
Kuma et al. [27]	Gabor-Ridgelet Transform	96% KTH Human Action Dataset

VII. CONCLUSIONS

This paper proposed a hybrid system of Kalman filtering and CNN for detection (with background removal) and recognition of moving objects in videos. The algorithm shows an increase in accuracy of guessing new cases when tested using 8 different videos with total video time of 15.4 minutes and 23100 frames. In that test, the rate of accuracy 100% has been reached for identification and recognition of moving objects. Experimental results show the superiority of the proposed detection and recognition approach as compared to existing algorithms, especially in presence of occlusions, making it more appropriate for many applications such as airport control.

REFERENCES

[1] M. Everingham, et al, The PASCAL Visual Object Classes Challenge: A Retrospective, International Journal of Computer Vision, 2015.

[2] O. Russakovsky et al, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision volume, 2015.

[3] C. Crispim-Junior, et al, Semantic event fusion of different visual modality concepts for activity recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2016).

[4] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE, 2012.

[5] S. Chaabouni, J. Benois-Pineau, O. Hadar, C. B. Amar, Deep learning for saliency prediction in natural video.

[6] P. Parker, K. Englehart, B. Hudgins, Myoelectric signal processing for control of powered limb prostheses, J Electromyogr Kinesiol 16 (6) (2006) 541–548.

[7] L. H. Smith, L. J. Hargrove, B. A. Lock, T. A. Kuiken, Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay, IEEE Trans Neural Syst Rehabil Eng 19 (2) (2011) 186–192.

[8] Jatuporn ToyLeksut, et al, "Learning visual variation for object recognition", Image and Vision Computing, 2020.

[9] M. Gomathy Nayagam, "Reliable object recognition system for cloud video data based on LDP features", Computer Communications Volume 149, January 2020, Pages 343-349.

[10] Rebecca K. Hendel, "The good, the bad, and the average: Characterizing the relationship between face and object processing across the face recognition spectrum", Neuropsychologia, 2019.

[11] Chengpeng Chen, "Hybrid incremental learning of new data and new classes for hand-held object recognition", J. of Visual Communication and Image Representation, 2019.

[12] Fang Wei," DOG: A new background removal for object recognition from images", Engineering Applications of Artificial Intelligence, Volume 91, 2020.

[13] Boureau, Y.-L, 'A theoretical analysis of feature pooling in visual recognition', Proceedings of the 27th International Conference on Machine Learning, 2010.

[14] Nair, V., "Rectified linear units improve restricted boltzmann machines", in 'Book Rectified linear units improve restricted boltzmann machines' (2010, edn.), pp. 807-814.

[15] Krizhevsky, A., "Imagenet classification with deep convolutional neural networks', 'Book Imagenet classification with deep convolutional neural networks' (2012, edn.), pp. 1098-1106.

[16] Bottou, L., 'Stochastic gradient descent tricks', Neural networks: Tricks of the trade, 2012.

[17] Christos-Christodoulos,"Hydrophobicity classification of composite insulators based on convolutional neural networks", Engineering Applications of Artificial Intelligence, 2020.

[18] T.Shanthi, "Automatic diagnosis of skin diseases using convolution neural network", Microprocessors and Microsystems Volume 76, July 2020.

[19] Amin Khatamia,"A weight perturbation-based regularisation technique for convolutional neural networks and the application in medical imaging", Expert Systems with Applications, 2020.

[20] Ding-XuanZhou, "Theory of deep convolutional neural networks: Downsampling", Neural Networks Volume 124, April 2020, Pages 319-327.

[21] The CIFAR-10 dataset.

[22] Muhammad Ehatisham-ul-Haq, "Continuous authentication of smartphone users based on activity pattern recognition using passive mobile sensing", J. Network and Computer Applications, 2018.

[23] Bruno Urbano Rodrigues, " Cachaça Classification Using Chemical Features and Computer Vision", Procedia Computer Science, Volume 29, 2014, Pages 2024-2033.

[24] D. K. Vishwakarma, et al, "Unified framework for human activity recognition: An approach using spatial edge distribution and R-transform," AEU-Int. J. of Electronics and Communications, 2016.

[25] T. Cover and P. Hart, "Nearest neighbour pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, p. 1967, 21-27.

[26] Y. L. Tian, L. Cao, Z. Liu and Z. Zang, "Hierarchical Filtered Motion for Action Recognition in Crowded Videos," IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2012.

[27] Kumar Vishwakarma, D., "A Two-fold Transformation Model for Human Action Recognition using Decisive Pose," Cognitive Systems Research, 2020.