A Review of Foreground Segmentation based on Convolutional Neural Networks

Pavan Kumar Tadiparthi[1], Sagarika Bugatha[2], Pradeep Kumar Bheemavarapu[3]
Associate Professor, Department of Information Technology, MVGR College of Engineering (A), Vizianagaram, A.P, India[1]
Student, Department of Information Technology, MVGR College of Engineering (A), Vizianagaram, A.P, India[2, 3]

***Abstract*—Foreground segmentation in dynamic videos is a challenging task for many researchers. Many researchers worked on various methods that were traditionally developed; however, the performance of those state-of-art procedures has not yielded encouraging results. Hence, to obtain efficient results, a deep learning-based neural network model is proposed in this paper. The proposed methodology is based on Convolutional Neural Network (CNN) model incorporated with Visual Geometry Group (VGG) 16 architecture, which is further divided into two sections, namely, Convolutional Neural Network section for feature extraction and Transposed Convolutional Neural Network (TCNN) section for un-sampling feature maps. Then the thresholding technique is employed for effective segmentation of foreground from background in images. The Change Detection (CDNET) 2014 benchmark dataset is used for the experimentation. It consists of 11 categories, and each category contains four to six videos. The baseline, camera jitter, dynamic background, and bad weather are the categories considered for the experimentation. The performance of the proposed model is compared with the state-of-the-art techniques, such as Gaussian Mixture Model (GMM) and Visual Background Extractor (VIBE) for its efficiency in segmenting foreground images.**

Keywords*—Foreground segmentation; deep learning; Convolutional Neural Network (CNN); Visual Geometry Group (VGG) 16 architecture; Transposed Convolutional Neural Network (TCNN); Gaussian Mixture Model (GMM); Visual Background Extractor (VIBE)

I. Introduction

Foreground segmentation [1] is a major part of various applications of computer vision. Foreground segmentation means that segmenting moving information from static information (background). Foreground segmentation is also called as Background Subtraction or Change Detection. Foreground segmentation is widely used in several applications like video surveillance [2], traffic monitoring, shopping malls, airports, etc. It analyzes a video sequence by using a set of techniques and those video sequences are recorded by a stationary camera.

Gaussian Mixture Model introduced by Chris Stauffer et al., [3] works on pixel-based classification. GMM models each pixel with K-Gaussians. It easily copes up with illumination changes. Even though a single Gaussian function is not able to deal with a dynamic background by providing a low updating rate of the background model. It is failed by the camouflage effect. The number of Gaussians here is predetermined as either 3, 4 or 5.

Visual Background Extractor introduced by O. Barnich et al., [4] is a non-parametric method and it works on pixels. This method utilizes the spatial information around the pixel for the background model. First of all, a set of values should be taken for each pixel at the same location in the neighborhood. Later, it compares this set to the current pixel value to determine whether it is background or foreground and adapts the model by choosing randomly among values to substitute from the background model. This approach differs from classical approach and belief that the oldest values should be replaced first. Finally, when the pixel is found to be part of the background, its value is generated into the background model of a neighboring pixel. Visual Background Extractor (VIBE) applied to color values of pixels of background training sequences as samples of observed backgrounds. Visual Background Extractor (VIBE) shows the best performance because it using samples as background models to represent the background changes. However, Visual Background Extractor (VIBE) has a major disadvantage that it uses color values of pixels to build the background model but color values are found to be sensitive to noise and illumination changes.

This article concentrated on developing a foreground segmentation model based on Deep Learning technique called Convolutional Neural Networks. The Convolutional Neural Network (CNN) is associated with Transposed Convolutional Neural Network (TCNN) for extracting the feature maps to identify foreground image. A thresholding technique is utilized to filter out the feature maps which distinguishes foreground object from a background object in an image. The main contributions of this article are:

- Detecting foreground objects in an image with improved accuracy.
- Design and implementation of a model using Deep Learning technique for effective segmentation of foreground objects.
- Evaluation of the proposed model using Gaussian Mixture Model (GMM) & Visual Background Extractor (VIBE) techniques.

The reminder of the paper is further organized as follows: Section II discusses related work, Section III presents the proposed methodology, Section IV of the article illustrated the experimental setup, and Section V illustrates the performance evaluation and experimentation results. Section VI gives out the conclusion and future work.

II. Related Work

In the past few years, eminent researchers experimented on foreground object segmentation techniques such as, Midhula Vijayan et al., [5] proposed a deep-neural network architecture using temporal and spatial information from background images and current processing images. Their experimentation obtained a better performance model when compared with existing background subtraction methods both qualitatively and quantitatively. Tsung-Han Tsai et al., [6] elucidated an unsupervised segmentation technique using distinct thresholding techniques by sorting of changed pixels ratio for precise decision. They obtained satisfactory results in generic images. Patrick Dickinson et al., [7] addressed the problem of foreground segmentation, when there is a varying background over time using Adaptive Gaussian Mixtures model (AGMM). Their experimentation resulted in better performance than the per-pixel and Markov Random Field-based Models and achieved a Jaccard coefficient of 0.59.

Jaime Gallego et al., [8] experimented by combining pixel-wise and region-based model for foreground segmentation using one Gaussian per pixel. This improved the performance of the system having similar colors to those of the background. Jaime Gallego et al., [9] proposed a method for monocular static camera sequences and indoor scenarios using Gaussian pixel color, Modified Mean Shift algorithm, and by Bayesian framework. Their methodology yielded robust segmentation and tracking of objects than the state-of-art methodologies. Jaime Gallego et al., [10] illustrated the reduction of false positive and false negative by using region-based models for modeling foreground and background region. Their model surrounds the foreground by Maximum A Posteriori and Markov Random Field model which uses pixel-wise color GMM for background sequence classification.

Jiayu Liang et al., [11] constructed a new method using Genetic Programming for feature construction by incorporating subtree technique. The simultaneous construction of multiple features and parsimony pressure techniques are introduced to improve the proposed techniques bloat control. Xuchao Gong et al., [12] developed a method using the GMM for modeling static background regions and inter-frame change detection and Scale-Invariant Feature Transform (SIFT) feature analysis is used for boundary identification of foreground regions. The Grab cut methods are used for segmentation of foreground moving objects. Yizheng Guo et al., [13] illustrated the segmentation of pigs in group-housed environments by combining a mixture of Gaussians using prediction mechanism and threshold segmentation algorithm.

Nikolaos Katsarakis et al., [14] considered Stauffer and Grimson’s algorithm as a baseline algorithm and enhanced their algorithm by changing the learning rate and combining the Gaussian mixture if they are similar. They yielded good results than the baseline algorithm.

As per the literature review, many models proposed by eminent researchers have failed to obtain efficient results for foreground segmentation. This paper focuses on the implementation of foreground segmentation methodology using deep learning techniques for enhanced segmentation results.

III. Proposed Methodology

The proposed methodology uses the VGG-16 model for foreground segmentation. It comprises two sections namely Convolutional Neural Network (CNN) and a Transposed Convolutional Neural Network (TCNN).

Each frame is extracted from video data and moved onto the following Convolutional Neural Network (CNN) and Transposed Convolutional Neural Network (TCNN).TCNN networks, where, the Convolutional Neural Network (CNN) network extracts the features from the input frame which are related to the foreground object by moving the frame onto different layers which are described in section A. The extracted features from the Convolutional Neural Network (CNN) network are fed into Transposed Convolutional Neural Network (TCNN), where the feature maps are unsampled to obtain original input size. By applying the thresholding technique based on probability to the unsampled features, the foreground objects are segmented from the background objects in a frame and hence for all the frames.

The proposed model is highlighted in Fig. 1. The Convolutional Neural Network and Transposed Convolutional Neural Network are described as follows:

A. Convolutional Neural Network (CNN)

The actual VGG-16 model contains 5 blocks with 16 layers. The feature extractor section uses 5 blocks and each block contains a set of 3×3 kernels like a stack with a max-pooling layer. The filters used in each convolutional layer are 64, 128, 256, 512, and 1024.

Our proposed model contains only 4 blocks as a feature extractor section and each block having 3×3 kernels as a smaller kernel size. Two receptive fields of 3×3 kernels are equal to a 5×5 receptive field. The three 3×3 kernels are equal to the 7×7 receptive field. Due to this, the parameters are reduced to 30%. The input is a W×H RGB image. The four convolutional layers of 3×3 kernels are followed by a max-pooling layer with stride 2. This first section produces the feature maps of size W/8×H/8 with 512. Doing this input by half, we get fine features from each frame, so it will have minute information from frame. The feature extraction section is done well because the Visual Geometry Group (VGG) 16 architecture, network is pre-trained. It is already trained on millions of images. It uses the CDNET-2014 dataset as input. The obtained feature maps are fed as input for the Transposed Convolution Neural Network (TCNN) part.

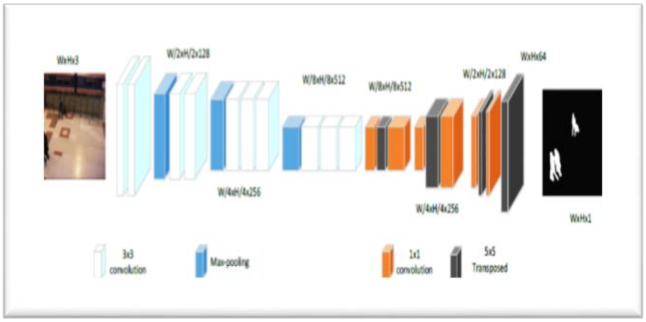

Fig. 1. Architecture of the Experimented CNN.

B. Transposed Convolutional Neural Network (TCNN)

The feature maps in the CNN section will be unsampled by Transposed Convolution Neural network (TCNN). It un-samples the features into original input by performing the multiplication of output by transposing the kernel or padding output to reconstruct the input. It down samples the feature maps by decreasing them from 512 to 64.

In this TCNN section, we have four blocks each block contains two 1×1 convolutions followed by 5×5 transposed convolutions with stride 2. These 1×1 convolutions are used to shrink the feature maps to get the original input size. After that thresholding technique is applied for the segment the foreground object.

IV. Experimental Setup

The overall experimentation was carried out on a 64-bit Windows 10 operating system having Inter® core™ i5 processor clocked at 2.24 GHz, 8 GB RAM, and 1 TB hard drive installed with Anaconda, Python platform with Tensor flow as backend and supporting image processing packages.

The CDNET-2014 benchmark dataset is used for experimentation. The dataset comprises of 11 categories and each category has 4 or 5 videos. For the experimentation, only four categories are namely: baseline, camera jitter, bad weather, and dynamic background to validate the performance of the model.

V. Performance Evaluation and Experimentation Results

To evaluate the performance of the proposed model different quality metrics are considered such as: Precision, Recall, Accuracy, F-Score, mean Squared Error (MSE), Root Mean Squared Error (RMSE), False Negative Rate (FNR), False Positive Rate (FPR), Peak Signal to Noise Ratio (PSNR), and Pair Wise Correlation (PWC) coefficient which are defined as follows:

$$Precision = \frac{True_{pos}}{False_{pos} + True_{pos}} \quad (1)$$

$$Recall = \frac{True_{pos}}{False_{neg} + True_{pos}} \quad (2)$$

$$Accuracy = \frac{True_{pos} + True_{neg}}{True_{pos} + False_{pos} + True_{neg} + False_{neg}} \quad (3)$$

$$F - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Actual\ Values - Predicted\ Values)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual\ Values - Predicted\ Values)^2} \quad (6)$$

$$FNR = \frac{False_{neg}}{True_{pos} + False_{neg}} \quad (7)$$

$$FPR = \frac{False_{pos}}{True_{neg} + False_{pos}} \quad (8)$$

$$PSNR = 10log10\,\frac{R2}{MSE} \quad (9)$$

$$PWC = 100 * \frac{False_{neg} + False_{pos}}{True_{pos} + False_{neg} + True_{neg} + False_{pos}} \quad (10)$$

To validate the performance of the proposed model, GMM and VIBE are considered as reference model. The graphical illustration of experimental results is shown in Fig. 2 and their corresponding results are given in Tables I to IV.

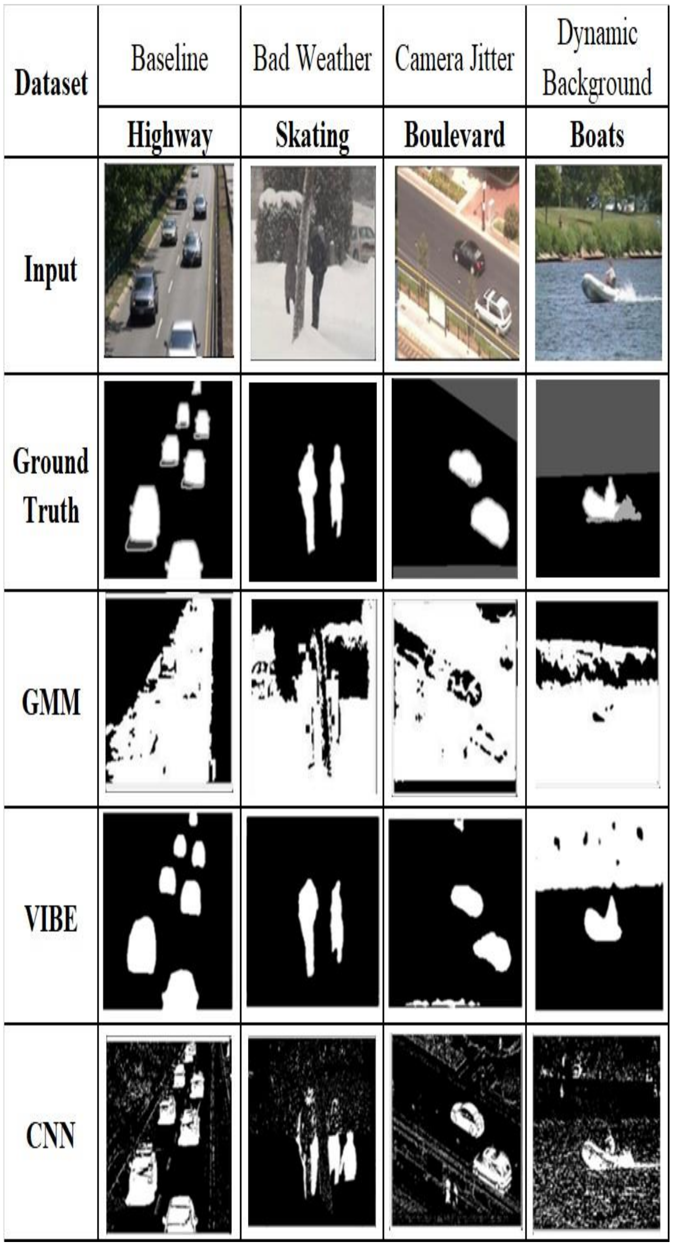

Fig. 2. Foreground Detection of Baseline, Camera Jitter, Dynamic Background and Bad Weather from CDNET-2014.

TABLE I. EVALUATION METRICS OF DIFFERENT METHODS ON CAMERA JITTER FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0127	0.0168	0.0197
Recall	0.124	0.0427	0.0314
Accuracy	0.9155	0.9853	0.9998
F-Score	0.0152	0.0372	0.0584
MSE	0.029	0.025	0.021
RMSE	0.0541	0.0502	0.0122
FPR	0.0845	0.046	0.025
FNR	0.743	0.852	0.975
PSNR	73.5035	74.1422	86.4039
PWC	6.4477	4.4724	2.0219

TABLE II. EVALUATION METRICS OF DIFFERENT METHODS ON DYNAMIC BACKGROUND FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0138	0.0154	0.0185
Recall	0.132	0.0316	0.0213
Accuracy	0.9725	0.9836	0.9999
F-Score	0.0131	0.0281	0.0673
MSE	0.039	0.036	0.034
RMSE	0.0430	0.0325	0.0132
FPR	0.0032	0.023	0.045
FNR	0.621	0.743	0.864
PSNR	71.824	72.9144	91.1751
PWC	5.3367	4.6399	3.0136

TABLE III. EVALUATION METRICS OF DIFFERENT METHODS ON BASELINE FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0154	0.0162	0.0174
Recall	0.135	0.0538	0.0125
Accuracy	0.9412	0.9861	0.9998
F-Score	0.0125	0.0473	0.0762
MSE	0.0204	0.047	0.053
RMSE	0.0571	0.0492	0.0100
FPR	0.0588	0.0137	0.077
FNR	0.632	0.835	0.952
PSNR	65.0579	74.3177	88.1648
PWC	5.8794	0.3922	0.0154

TABLE IV. EVALUATION METRICS OF DIFFERENT METHODS ON BAD WEATHER FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0134	0.0184	0.0195
Recall	0.128	0.0724	0.0512
Accuracy	0.9202	0.9674	0.9999
F-Score	0.0236	0.0584	0.0873
MSE	0.0408	0.0219	0.0132
RMSE	0.1442	0.0439	0.0133
FPR	0.0798	0.0526	0.0266
FNR	0.543	0.721	0.843
PSNR	64.9842	75.3205	93.4217
PWC	7.9846	4.2576	0.0143

VI. CONCLUSION AND FUTURE WORK

This paper focused on the enhancement of foreground object segmentation using deep neural network model viz., VGG-16 which comprises of CNN and TCNN. To evaluate the performance of the proposed model, comparison with the traditional methods such as GMM and VIBE is done. The results showcased that the proposed VGG-16 model has proven its supremacy in obtaining the highest accuracy of 99.99% in -comparison to the state-of-the-art techniques such as GMM and VIBE. Hence, the proposed model performed better in segmenting the foreground image with improved accuracy. Exploring other advanced deep learning techniques for improved segmentation of foreground objects in images is considered as the future work.

REFERENCES

[1] Shahbaz, L. Kurnianggoro, Wahyono, and K.-H. Jo, “Recent Advances in the Field of Foreground Detection: An Overview,” Advanced Topics in Intelligent Information and Database Systems. Springer International Publishing, (2017), pp. 261–269.

[2] Shahbaz, J. Hariyono, and K. H. Jo, “Evaluation of background subtraction algorithms for video surveillance,” 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), Jan (2015), pp. 1–4.

[3] Chris Stauffer, W.E.L Grimson, “Adaptive background mixture models for real-time tracking,” Proceedings of IEEE Conference Computer vision Pattern Recognition, Vol. 2, Jan, (2007).

[4] O. Barnich, M. Van Droogenbroeck, “ViBe: A universal background subtraction algprithm for video sequences,” IEEE Transactions on Image Processing, Vol. 20: Issue. 6, Jun (2011).

[5] Midhula Vijayan, R. Mohan, “A Universal Foreground Segmentation Technique using Deep-Neural Network”, Multimedia Tools and Applications, May, (2020).

[6] Tsung-Han Tsai, Guan-Jun Chen, Wen-Liang Tzeng, “A Novel Foreground/Background Decision using in Unsupervised Segmentation of Moving Objects in Video Sequences”, 46th Midwest Symposium on Circuits and Systems, Cairo, Vol. 3, Dec, (2003).

[7] Patrick Dickinson, Andrew Hunter, Kofi Appiah, “A spatially distributed model for foreground segmentation”, Image and Vision Computing, Vol. 27: Issue. 9, Aug, (2009).

[8] Jaime Gallego, Montse Pardas, Gloria Haro, “Bayesian Foreground Segmentation and Tracking using Pixel-wise Background Model and Region based Foreground model”, 16th IEEE International Conference on Image Processing (ICIP), Cairo, Nov, (2009).

[9] Jaime Gallego, Montse Pardas, Gloria Haro, “Enhanced foreground segmentation and tracking combining Bayesian background, shadow and foreground modeling”, Pattern Recognition Letters, Vol. 33: Issue 12, Sep, (2012).

[10] Jaime Gallego, Pascal Bertolino, “Foreground Object Segmentation for moving camera sequences based on foreground probabilistic models and prior probability maps,” IEEE International Conference on Image Processing (ICIP), Paris, Oct, (2014).

[11] Jiayu Liang, Yu Xue, Jianming Wang, “Genetic programming-based feature construction methods for foreground object segmentation,” Engineering Applications of Artificial Intelligence, Vol. 89, Mar, (2020).

[12] Xuchao Gong, Zongmin Li, “Efficient Foreground Seg mentatin Using an Image Matting Technology,” 2013 International Conference on Computational on Computational and Information Sciences, Shiyang, Jun, (2013).

[13] Yizheng Guo, Weixing Zhu, Pengpeng Jiao, Jiali Chen, “Foreground detection of group-housed pigs based on the combination of Mixture of Gaussians using prediction mechanisms and threshold segmentation,” Biosystems Engineering, Vol. 125, Sep, (2014).

[14] Nikolaos Katsarakis, Zheng-Hua Tan, Ramjee Prasad, Aristodemos Pnevmatikakis, “Improved Gaussian Mixture Models for Adaptive Foreground Segmentation,” Wireless Pers Commun 87, Apr, (2016).