How to Cite:
Sowmeya , V., & Karthik, R. J. (2022). Internet of things module accelerated dense deep learning for crime detection in surveillance systems. *International Journal of Health Sciences*, *6*(S1), 6364–6379. https://doi.org/10.53730/ijhs.v6nS1.6353

Internet of things module accelerated dense deep learning for crime detection in surveillance systems

Sowmeya V
Research Scholar, Department of Computer Science, Vels Institute of Science Technology and Advanced Studies, Chennai

DR. R. Jaya Karthik
Assistant Professor, Department of Computer Science, Vels Institute of Science Technology and Advanced Studies, Chennai

***Abstract*---**The smart surveillance system is becoming a vital application in each streets or houses. Most of the streets are prone to several misbehavior conducts for instance theft in atm, robbery, fights etc. and hence it is necessary to detect and analyse the crime scenes for finding the suspects. However, most of the surveillance system suffers from poor detection of objects due to poor camera resolution, absence of light and other factors. In order to improve the detection of faces after detecting the objects using ResNet, it is necessary to adapt some advanced devices for image capturing and analyzing. In this paper, Internet of Things (IoT) based ESP32 CAM WiFi Module Bluetooth with OV2640 Camera Module 2MP is used for image acquisition that capture better images from the scenes. The study uses dense convolutional network namely DenseNet to detect the faces present in the crime scenes after the object detection. The deep learning module is trained with selected crime scenes for training the classifier. The simulation is conducted further to validate the model with other variants of deep learning.

***Keywords*---**Crime Scene, Internet of Things, DenseNet, Surveillance System.

Introduction

Video surveillance and face recognition systems have garnered considerable consideration in recent years. Building strong video-based face recognition systems is vital, as the number of surveillance cameras in public spaces increases. [1]. In video surveillance, capture conditions normally range from semi-controlled free flow in a congested scene to one-man in the scene. The video

surveillance still-to-video face recognition and the video-to-video face recognition [2] are two typical uses. Reference faces of individuals of interest in the first application are used to construct facial models, whereas facial models in the latter application are designed with faces captured in films. The main topic of this chapter is still video FR with a single sample per person in semi- and uncontrolled video surveillance situations.

In still-to-video face recognition applications, the number of target references is one or very small and the characteristics of the still camera (s) used for design vary considerably from the video cameras used in operations [3]. Thus, according to different variations in ambient lighting, posture, blur and occlusion, there are considerable variances between the looks of still ROI (s) and ROI taken with security cameras [4]. While the target individual registration is done with facial ROIs that are isolated in still images, the ROIs on facial models are matched to these facial models throughout operations. In video surveillance, a person in a scene can be followed for a range of frames, and matching scores can be collected for strong spatiotemporal face recognition on a facial trajectory [5].

The conventional approaches are supported by hand-made procedures for extracting functions and a prepared classifier together with fusion, while profound learning methods automatically learn characteristics and classify them with a mass of data. Despite the advances made using conventional methodologies, the face recognition scenario is less resistant to the real world. On the other hand, no feature extraction technology is available that can individually solve all the obstacles facing video surveillance [6].

Typically, standard approaches proposed to use face recognition still-to-video are modelled on one or two-class face detector systems to allow the system to add or remove other individuals and to quickly adapt over time [7]. The video surveillance [2] has successfully implemented modular systems developed for individual ensembles. Ensemble based methods were therefore demonstrated as a reliable solution to handle imbalanced data.

In order to improve the robustness of face recognition images, numerous representations of the face were encoded into groups of classifiers [8]. While the development of robust facial models by means of a single sample of training is tough, numerous ways have solved this problem, such as several figures for the face, synthetic virtual face generation and the use of auxiliary data by others to expand the training set [9].

These technologies aim to increase intra-class variations and the resilience of facial models. In some displays, different patches and facial descriptors are used [8] and artificial facial images are synthesised using 2D morphing or 3D reconstructions [10]. [10]. A generic auxiliary dataset comprising other people faces can be used to modify domains [11] and to classify displays through dictionary training [12]. However, in terms of the prior knowledge required to find facial components reliably, and large differences between the quality of still and video ROIs, techniques based on synthetic facial generation and auxiliary data are more complex and costly in computational terms for real-time applications.

Several strategies have been recently suggested to provide effective face representatives utilising neural networks and non-linear mappings straight from training data [13] [14]. In such ways, the training process might take into account distinct loss functions to improve interpersonal variations whilst at the same time reducing intrapersonal variations. They may learn nonlinear and discriminatory representations of features to cover the existing gaps compared with the human visual system [14], albeit computationally expensive and generally need to train huge numbers of marked data.

A three-fold loss function has been established in [15] to address the problem of a single sample per individual in face recognition to classify between two ROIs and one pair of ROIs that do not match. The CNN assembly, like CNN [16] and HaarNet [17], has been shown to extract features from the global face appearance as well as incorporate asymmetric properties for the treatment of partial occupation. In order to develop robust representations of the function, supervised autoencoders were further proposed to enforce faces with variations mapped to the canonical surface in a single sample per individual scene [18].

In this paper, Internet of Things (IoT) based ESP32 CAM WiFi Module Bluetooth with OV2640 Camera Module 2MP is used for image acquisition that capture better images from the scenes. The study uses dense convolutional network namely DenseNet to analyse the crime scenes.

Related works

In various computer vision applications, such as object detection, object reconnaissance, and so on, deep CNNs have recently shown significant success. Such profound CNN models have shown that distinct variations are adequately characterised within a wide range of data and the nonlinear representation in a discriminatory way. In addition, pre-trained models by transmission learning can easily be generalised to various vision tasks. They are thus a successful instrument for various applications of FR through the learning directly from face-size images of efficient feature representations [15].

For example, an ensemble of CNN models were trained [20] to manage position and partial occlusion fluctuation using holistic face image and various overlapping/non-overlapping face patches. Typically, the merging of these models occurs through a combination of features to create overall and compact representations. In [21], the last hidden layer has become more prominent, and supervision has been used for convolutionary layers to acquire hierarchical and non-linear representations. These representations aim at enhancing interpersonal differences by separately extracting elements from distinct identities and, at the same time, reducing intrapersonal changes. In DeepFace [22], a precise facial alignment was included to provide a sturdy face image through a 9-layer deep CNN, as opposed to the DeepID series. In [23], the high level of face-like features were recovered jointly by multiple deep CNNs for facial inspection applications from a pair of faces instead of a single face. Because fluctuations like blurredness or size shifts are no longer considered, these techniques are not fully suitable for FR applications based on video.

Likewise, a threefold loss function was recently used to learn robust facial embedding for the SSPP problems in [16], where this sort of loss tries to differentiate the positive pair of matching facial ROIs from the negative facial ROI. In [24], a compact and quick cross-correlation CNN was presented to achieve a robust facial representation that was learned using triple-loss optimization.

But the robustness of facial appearance changes can be further increased through increasing computational complexity, by CNN models like the CNN [8] trunk-branch ensemble and HaarNet [17]. The trunk network extracts elements from the global appearance of the faces in these models, and the branch networks incorporate asymmetrical and complicated facial characteristics. For example, HaarNet uses three hair-like branch networks, whereas the TBE-CNN considers landmarks. These specialised CNNs, however, are sophisticated systems not well suited for FR applications in real-time [25].

In addition, autoencoders may often be used to extract non-linear deterministic feature mappings that are robust to images that are contaminated by different noises, such as lighting, appearance, and poses. An autoencoder network has encoder and decoder modules, where the former incorporates input data into hidden nodes, and the latter returns the hidden nodes to the original input area, which reduces errors in reconstruction (s). Several networks with autoencoding inspired by [26] were suggested to remove the above variances in face pictures [27]. These networks address faces with many variations, such as noisy pictures.
For example, a facial CNN was trained in [28] to morph facial faces, where changes in posture and lighting are used to make facial representations of the last concealed layer. Similarly, multifunctional learning was proposed for the rotation of faces with arbitrary poses and lightings to the target face [29]. In [30], a deep general architecture was also developed to encode and integrate the desired attribute with the image input to generate target images as comparable to the visual attribute input photos, but other characteristics of a face have been not changed.

Proposed Method

The study considers a deep learning model to extract the features from the input video frames, where a convolutional neural network (CNN) offers improved extraction of features. It enables the production of feature map from an input video frame. After obtaining the feature map, the objects are detected using ResNets based on its objectness score. These detected objects are resized and fitted to a specific dimension and then it is sent to DenseNet for the purpose of facial detection. The process of facial detection in a video frame is given in Figure 1.

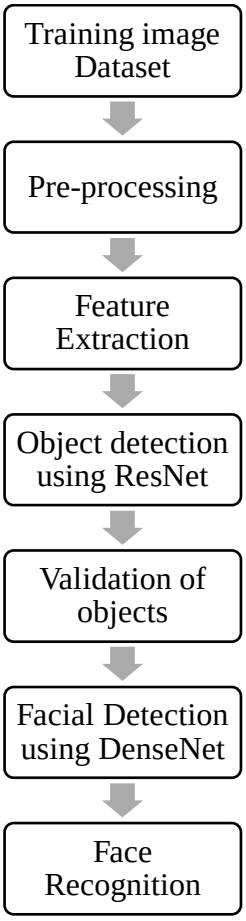

Figure 1: Facial Detection model

We initially apply face detection in an image via DenseNet in order to do facial recognition. Detected and aligned faces are then sent into a trained DenseNet model that builds facial embedding. Using k-NN approach, face recognition is carried out if the embeddings are produced. This is because DenseNet converts face images into Euclidean space so as to directly match distances. Therefore, it is possible to evaluate the similarity between two face photos simply by determining the distance between the points that represent the face, described by:

$$d(a,b) = d(b,a)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

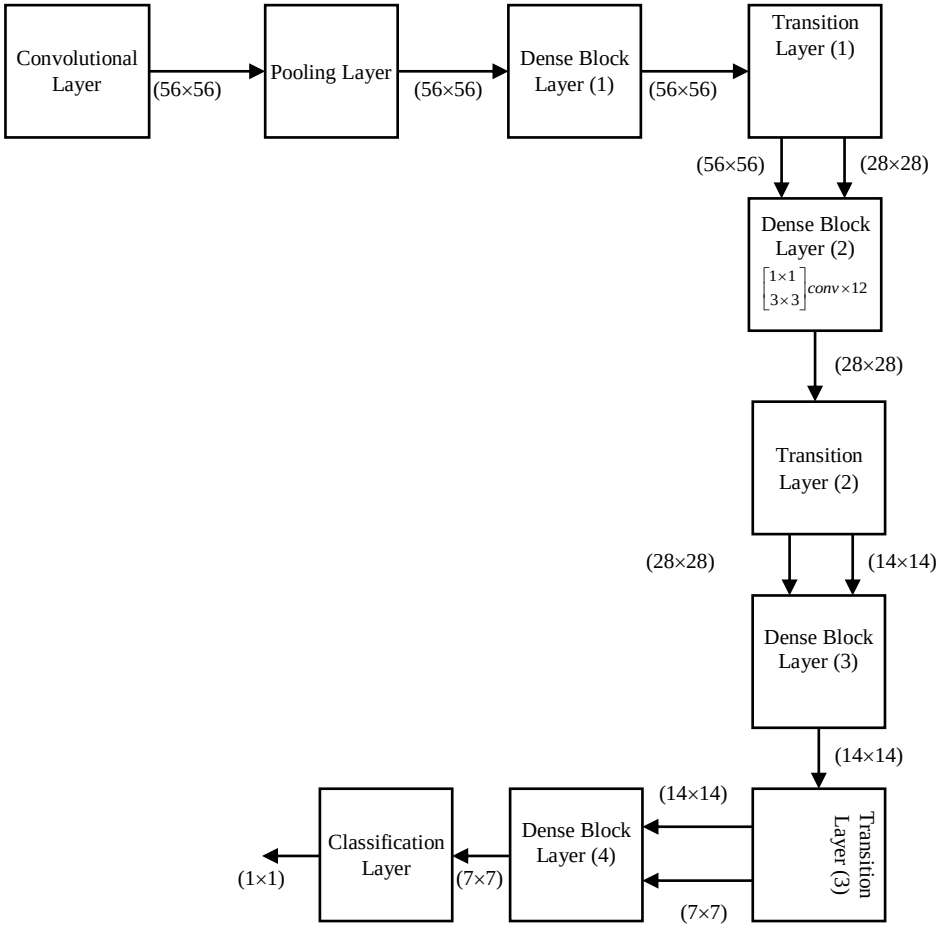

Figure 2: Proposed DenseNet-169 Architecture

DenseNet-169+ Detection

DenseNet-169+ is a sort of neural network which uses thick layer links via thick blocks, where the test links the whole layer directly. Each layer receives additional inputs from all previous layers and sends them through its own maps with their subsequent layers to maintain the classifier.

Take a single video frame x_0 through the network initial layer, i.e., convolutional network. The DenseNet-169+ network is built with L levels, where the non-linear $H_l(\cdot)$ layers are used in each layer, and the l layers are regarded as being indexed. The $H_l(\cdot)$ composite function includes convolution, bundling, linear remedial units and standardisation batch operations. The 'l^{th} layer' output is indicated as x_l.

ResNets

The output of the l^{th} layer acts as a result of the transition to $(l+1)^{th}$ consecutive layer, which connects it via a typical convolutional feed network:

$$x_l = H_l(\mathbf{x}_{l-1}). \quad (1)$$

ResNets adds a skip connection to sidestep non-linear transformation operations with the following identity function:

$x_l = H_l(x_{l-1}) + x_{l-1}$. (2)

ResNets have the benefit that the gradient can travel through an identity function directly from conventional layers to past layers. But H_l output obstructs the DenseNet data flow.

Dense connectivity

This work proposes a different structure for the connectivity to further increase the data flow between layers, e.g. from any layer, and to enhance dense learning. The layout of the DenseNet is shown in Figure 2. This means that all previous layer mappings $x_0,x_1,\ldots,x_{l-1}$ are provided as input for the last layers l:

$x_l = H_l(|x_0,x_1,\ldots,x_{l-1}|)$ (3)

where

$[x_0,x_1,\ldots,x_{l-1}]$ refers to the feature-map concatenation at $0,\ldots,l-1$.

Due to its strong interconnectedness, this network architecture is called the DenseNet. For easy application in a tensor, the study combines various $H_l(\cdot)$ inputs into Eq.(3).

Composite function.

$H_l(\cdot)$ is defined by the study as a composing three-fold function including batch normalisation, a linear corrected unit and 3×3 Conv.

Pooling layers

The map size of the feature alters the Eq. (3), which is not regarded as viable. The down sample layers, however, vary the map size, which acts in dense networks as an essential component. The research splits the network into layers, where the layers comprise highly connected blocks to assist the process of architectural downsampling. The study refers to block layers as converging and pooling transitional layers. The layers used for the studies are a standardisation batch layer and a convolutional 1×1layer followed by a mean pooling layer of 2×2, respectively.

Growth rate

If k character maps are provided by each H_l function, then the DenseNet-169+ function will be followed by $k_0 + k\times(l-1)$ input-level feature maps with k_0 channels. K is called the growth rate of the network. The study reveals that a decreased rate of increase is sufficient for the most advanced sets of data to be tested.

Bottleneck layers

Although each layer has k output properties, there are usually many more inputs. The 1×1 convolution can be implemented as a bottleneck layer before each three to three convolutions to decrease the number of convolutions, hence improving the computational efficiency of inputs.

Compression

In order to further improve the model, this work minimises the number of functional mappings on transition layers. If a dense block contains m function maps, the study creates output function maps where 0 is the current compression factor with the next transition layer. If θ =1, the number of characteristic maps throughout the levels of transition is not changed. The value in the experiment DenseNet-169+ is 1 and the value of the study θ=0.5 is set. When both bottleneck and transition layer 1 are implemented, the study refers to the model as DenseNet-169+.

Results and Discussions

In this section, we discuss the simulation of proposed face detection model using DenseNet module. The simulation is conducted in python on a high end computing machine that consists of a 16 GB RAM operates on an i5 core processor and a GPU of GTX 1060 4 GB.

In order to train and test the DenseNet model, the study uses Open Images Dataset (OID), which is the most relevant datasets for the forensic detection. The training is conducted in order to build a model using DenseNet model that detects the objects during the testing process. Such a training process will reduce the likelihood of DenseNet model to miss out the faces while the detection of faces from a video frame for possible forensics. The fine-tuning is conducted on each iterations in order to detect the objects with precise tuning.

The validation is conducted for the purpose of facial detection and the performance is evaluated in terms of different performance metrics that includes accuracy, precision, recall, F-measure and mean average percentage error (MAPE). All algorithms are compared with the following metrics:

Accuracy*:* It defines closeness of predicted value to real value. For accuracy measurement, the confusion matrix is used to obtain the accurate measure of the accuracy in each class with result distributions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is defined as the ratio of actual positive results and predicted positive results.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Recall is defined as the actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-Measure: It is defined as the metric to estimate the accuracy of classifier and it is the average of recall and precision metrics.

$$\text{F-Measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Here,
TP - True Positive, where the actual and predicted results are positive
TN - True Negative, where the actual and predicted results are negative
FP – False Positive, where the actual is negative and predicted results is positive and
FN – False Negative, where the actual is positive and predicted results is negative.

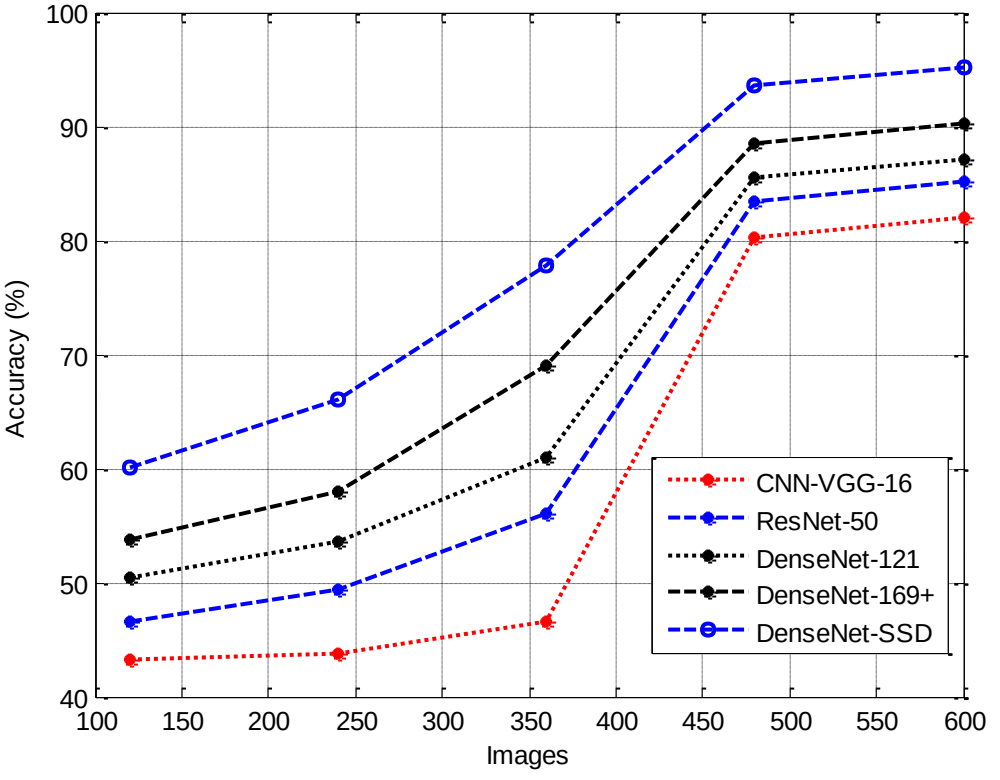

Figure 3: Accuracy

The Figure 3 shows the Results of accuracy between the proposed DenseNet SSD models with existing CNN-VGG-16. ResNet-50, DenseNet-121 and DenseNet-169+ model. The Results of simulation shows that the proposed model obtains an increased object detection accuracy than the other methods.

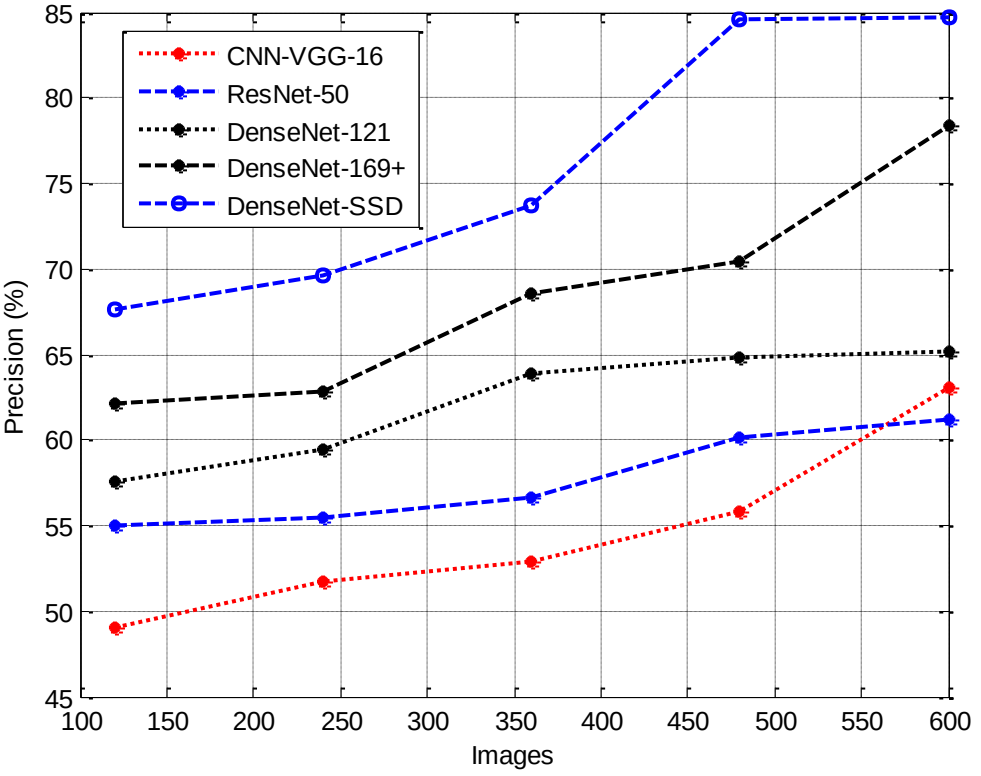

Figure 4: Precision

The Figure 4 shows the Results of precision between the proposed DenseNet SSD model with existing CNN-VGG-16. ResNet-50, DenseNet-121 and DenseNet-169+ model. The Results of simulation shows that the proposed model obtains an increased precision of object detection than the other methods.

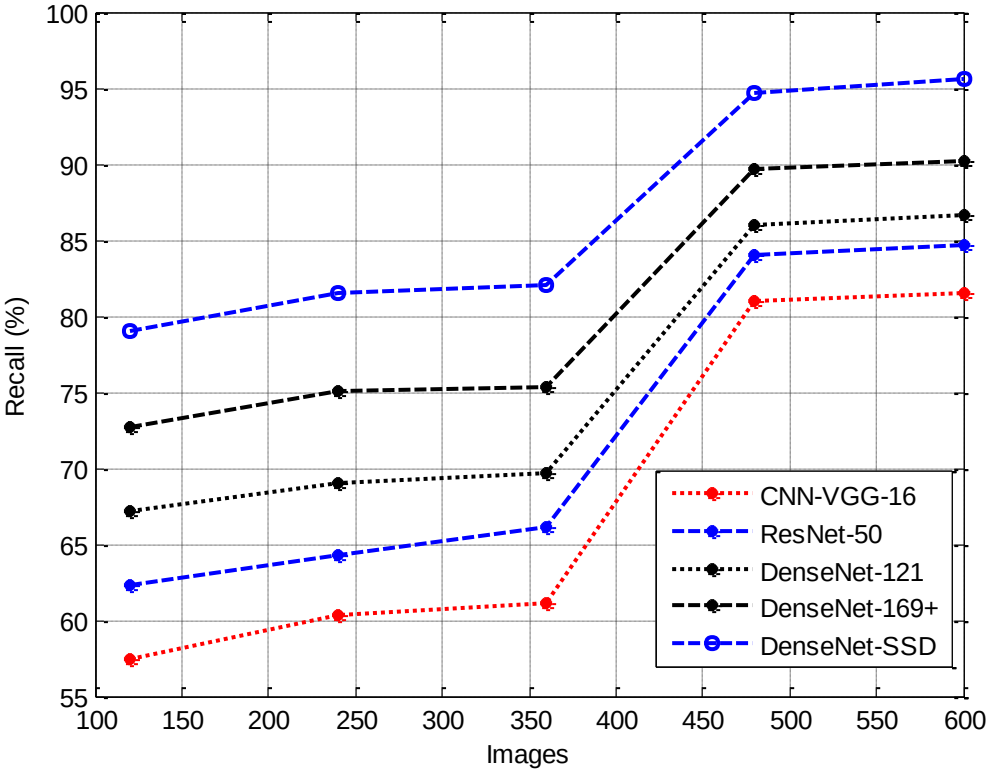

Figure 5: Recall

The Figure 5 shows the Results of recall between the proposed DenseNet SSD model with existing Faster CNN-VGG-16. ResNet-50, DenseNet-121 and DenseNet-169+ model. The Results of simulation shows that the proposed model obtains an increased recall on object detection than the other methods.

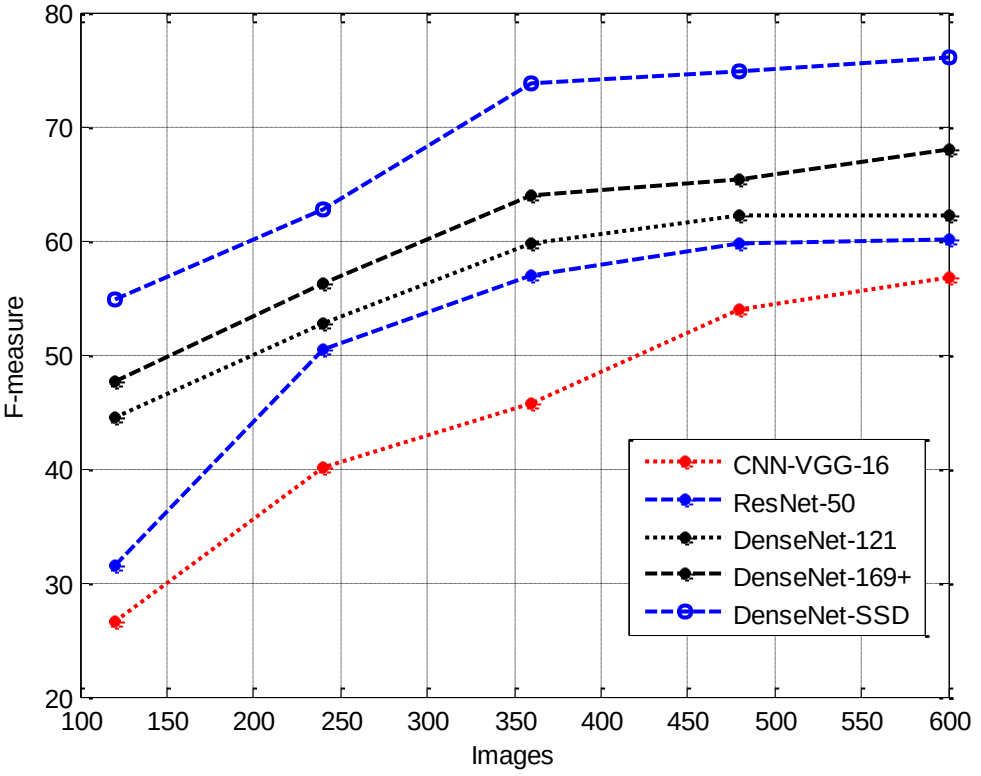

Figure 6: F-measure

The Figure 6 shows the Results of F-measure between the proposed DenseNet SSD model with existing CNN-VGG-16. ResNet-50, DenseNet-121 and DenseNet-169+ model. The Results of simulation shows that the proposed model obtains an increased F-measure on object detection in video surveillance than the other methods.

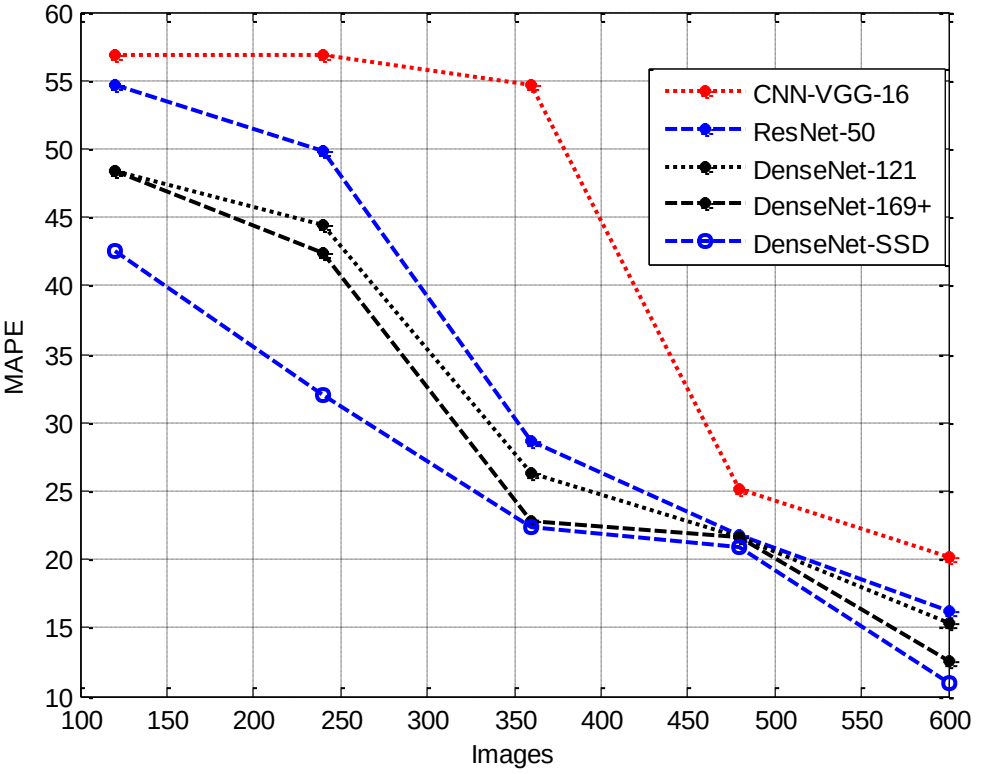

Figure 7: MAPE

The Figure 7 shows the Results of MAPE between the proposed DenseNet SSD model with existing CNN-VGG-16. ResNet-50, DenseNet-121 and DenseNet-169+ model. The Results of simulation shows that the proposed model obtains a reduced MAPE during the process of object detection accuracy than the other methods.

Conclusions

In this paper, we use DenseNet model to detect the faces after the detection of objects using ResNets. The data from the input acquisition device: ESP32, an IoT device from various regions are captured and processed by the deep learning model to forensic the crime scenes in smart cities. The videos are converted into frames and then it is pre-processed, extraction of features is conducted further to detect the objects by ResNets and faces by DenseNets. The simulation is conducted to validate the performance of DenseNet with existing deep learning models. The results of simulation shows that the proposed DenseNet is effective in detecting the faces from the input videos than the conventional deep learning mechanism.

References

[1] Zheng, J., Patel, V. M., & Chellappa, R. (2017). Recent developments in video-based face recognition. In *Handbook of Biometrics for Forensic Science* (pp. 149-175). Springer, Cham.

[2] Pagano, C., Granger, E., Sabourin, R., Marcialis, G. L., & Roli, F. (2014). Adaptive ensembles for face recognition in changing video surveillance environments. *Information sciences*, *286*, 75-101.

[3] Bashbaghi, S., Granger, E., Sabourin, R., & Bilodeau, G. A. (2017). Dynamic ensembles of exemplar-svms for still-to-video face recognition. *Pattern recognition*, *69*, 61-81.

[4] Matta, F., & Dugelay, J. L. (2009). Person recognition using facial video information: A state of the art. *Journal of Visual Languages & Computing*, *20*(3), 180-187.

[5] Dewan, M. A. A., Granger, E., Marcialis, G. L., Sabourin, R., & Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern recognition*, *49*, 129-151.

[6] Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., & Chen, X. (2015). A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, *24*(12), 5967-5981.

[7] Bashbaghi, S., Granger, E., Sabourin, R., & Bilodeau, G. A. (2014, August). Watch-list screening using ensembles based on multiple face representations. In *2014 22nd International Conference on Pattern Recognition* (pp. 4489-4494). IEEE.

[8] Bashbaghi, S., Granger, E., Sabourin, R., & Bilodeau, G. A. (2017). Robust watch-list screening using dynamic ensembles of svms based on multiple face representations. *Machine vision and applications*, *28*(1-2), 219-241.

[9] Kamgar-Parsi, B., Lawson, W., & Kamgar-Parsi, B. (2011). Toward development of a face recognition system for watchlist surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(10), 1925-1937.

[10] Mokhayeri, F., Granger, E., & Bilodeau, G. A. (2015, September). Synthetic face generation under various operational conditions in video surveillance. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 4052-4056). IEEE.

[11] Ma, A. J., Li, J., Yuen, P. C., & Li, P. (2015). Cross-domain person reidentification using domain adaptation ranking svms. *IEEE transactions on image processing*, *24*(5), 1599-1613.

[12] Yang, M., Van Gool, L., & Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. In *Proceedings of the IEEE international conference on computer vision* (pp. 689-696).

[13] Huang, G. B., Lee, H., & Learned-Miller, E. (2012, June). Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2518-2525). IEEE.

[14] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition (pp. 1701-1708).

[15] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

[16] Ding, C., & Tao, D. (2017). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 1002-1014.

[17] Parchami, M., Bashbaghi, S., & Granger, E. (2017, May). Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 4625-4632). IEEE.

[18] Parchami, M., Bashbaghi, S., Granger, E., & Sayed, S. (2017, August). Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

[19] Chellappa, R., Chen, J. C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Patel, V. M., & Castillo, C. D. (2016, October). Towards the design of an end-to-end automated system for image and video-based recognition. In *2016 Information Theory and Applications Workshop (ITA)* (pp. 1-7). IEEE.

[20] Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1891-1898).

[21] Sun, Y., Wang, X., & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2892-2900).

[22] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1701-1708).

[23] Sun, Y., Wang, X., & Tang, X. (2013). Hybrid deep learning for face verification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1489-1496).

[24] Parchami, M., Bashbaghi, S., & Granger, E. (2017, August). Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.

[25] Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*.

[26] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

[27] Gao, S., Zhang, Y., Jia, K., Lu, J., & Zhang, Y. (2015). Single sample face recognition via learning deep supervised autoencoders. *IEEE transactions on information forensics and security*, *10*(10), 2108-2118.

[28] Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014). Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*.

[29] Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., & Kim, J. (2015). Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 676-684).

[30] Ghodrati, A., Jia, X., Pedersoli, M., & Tuytelaars, T. (2015). Towards automatic image editing: Learning to see another you. *arXiv preprint arXiv:1511.08446*.