International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 5 ǁ 2022 ǁ PP. 264-270

An Efficient Deep Learning Approach for Recognizing and Computing the Humans

S. Vasudeva
PG Scholar
Department of Computer Applications
Madanapalle Institute of Technology & Science, India

Mr. S. Balamurugan
Assistant Professor
School of Computers
Madanapalle Institute of Technology & Science, Madanapalle, Chittor Dist. AP, India

Mrs. S. Savitha
Assistant Professor
Department of Computer Science
Saradha Gangadharan College, Pondicherry

ABSTRACT:
This paper focuses to build the real time Human Recognition and counting system using a deep learning model. The input is given through a webcam by providing our own video or images and the output is the number of humans moving in and out in the video input. A number of surveillance applications require the detection and tracking of people to ensure security, safety, and support site management. Substantial progress has been made to detect people wherever it is predictable. Often it is for example assumed that the people in the scene are well separated and that it is possible to identify foreground objects using a statistical background model. Certain actions can simply be detected if the location of all individuals in the scene is known. However, in all of the scenarios just mentioned we have to anticipate that people can appear in groups. In addition it is often necessary to know how many people are present. Although this method is capable of segmenting a region of interest into individuals, it needs to be embedded into a comprehensive system which supports the detection, tracking, and detection of specific events of humans. One possible design of such a system is the focus of the presented work. Counting of people entering and leaving a site and the detection of events are presented as potential applications of this system.
Keywords: *SSD, CNN, Deep learning, Alexa, security and safety.*

Date of Submission: 15-05-2022 Date of acceptance: 30-05-2022

I. INTRODUCTION:

However with the increased amount of video data to be processed it is becoming more and more unmanageable for human beings to monitor continuously. So if we could develop a surveillance system which could detect and classify objects, take decisions and label events autonomously, then a complete revolution can be brought in the current surveillance system. Vision based Human detection and counting is currently one of the most challenging tasks in the field of computer vision. The general surveillance cameras are like machines that can only see, but cannot decide or identify things or events on its own. So, keeping in mind the present day scenarios, it is important that we make our surveillance system intelligent and smart. Therefore, we propose to design a new framework to robustly and efficiently detect and count human beings, for application in surveillance. For these, we first intend to subtract the background and extract the foreground of any real time video.

People detection is considered an essential task in various video surveillance applications including crowd analysis, behavior analysis, crime prevention, and monitoring and management of peoples in public environments like railway stations, airports and shopping malls. Due to such vast applications, it is considered as one of the active research areas. However, because of different factors, including occlusions, distortions of scenes, various crowd distributions, and person body appearance, it is a challenging task.

Previously developed methods of people counting are typically based on segmentation or head detection. A few researchers also used feature-based methods based on the sliding window approach to detect and classify the person in the scene. Other researchers used feature-based methods which were designed for crowded conditions . After the revolution of the deep learning models , accuracy of various tasks of computer vision, such as person detection and tracking, crowd behavior analysis, gesture recognition, pose estimation and many other surveillance applications have been remarkably observed which might vary in accuracy and high computational speed.

II. LITERATURE SURVEY:

Recently, Human detection methods based on deep learning techniques have exhibited state-of –the-art (SOTA) performance. Most existing human detectors employ either single-stage or two-stage strategy as their backbone architecture. Liu W et al. proposed SSD using a single deep CNN to detect objects of various scales. This method separates the output space of bounding boxes over different aspecct scales and ratios for every extracted feature map location. Szarvas M et al. implemented pedestrain detection using CNN.

List of References those we are referred to do this project are:

Afifi et al. implemented robust real-time pedestrian detection using YOLOv3 on collected aerial images from Embedded Real-Time Inference (ERTI) Challenge on Jetson TX2 and achieved more than 5 frames per second (fps).

Real-time pedestrian detection using a robust Enhanced Tiny-YOLOv3 network was proposed by C.B Murthy. This method introduces an anti-residual module to improve the network's feature extraction and bounding box loss error is minimized but fails to detect severely occluded and denser pedestrians in real-time.

Donahue et al. designed a recurrent convolutional architecture, which cascaded a CNN with a recurrent model into a unified model. CNN was used to extract features of each frame, and then, these features were fed into LSTM step by step for modeling dynamics of the feature sequence so that it could learn video level representation in both spatial and temporal dimensions

.

Ng et al. combined the temporal feature pooling architecture with LSTM to allow the model to accept arbitrary-length frames.

Weng et al. proposed the CNN-2D to process salient-aware clips and fed the features extracted from the fully connected layer of a 3-D-CNN into LSTM for action recognition.

According to the spatial–optical data organization, Yuan et al. synthesized motion trajectories, optical, and video segmentation into spatial–optical data and used a two-stream 3-D CNN to process synthetic data and RGB data separately. Then, the resulting spatiotemporal features were fed into LSTM to mine their patterns.

In addition to 2-D CNNs used for image processing, 3-D CNNs were proposed to process videos. They replaced 3 × 3 convolutional kernels with those of 3 × 3 × 3 to perform 3-D convolutions over stacked frames. However, these methods usually have abundant parameters and need to be pretrained on a large-scale video data set.

A brief literature survey shows that there has been a plenty of research in the area of video analysis and human action recognition. We have come a long way in the part 5-6 years after the advent of neural networks. Initially CNNs applied frame by frame helped in improving the accuracies as compared to the manual feature extraction techniques. Later 3D-CNNs further improved the accuracies of CNNs by processing multiple frames at a time. More recent architectures started focussing on RNNs and LSTMs to factor in the temporal component of the videos. Most recent architectures started incorporating attention mechanism to focus on the salient parts of the videos.

Human action recognition is still a very active research area and new approaches are still trying to solve the issues with the current approaches. Some of the existing issues are background clutter or fast irregular motion in videos, occlusion, view point changes, high computational complexity and responsiveness to illumination changes.

III. PROPOSED SYSTEM:

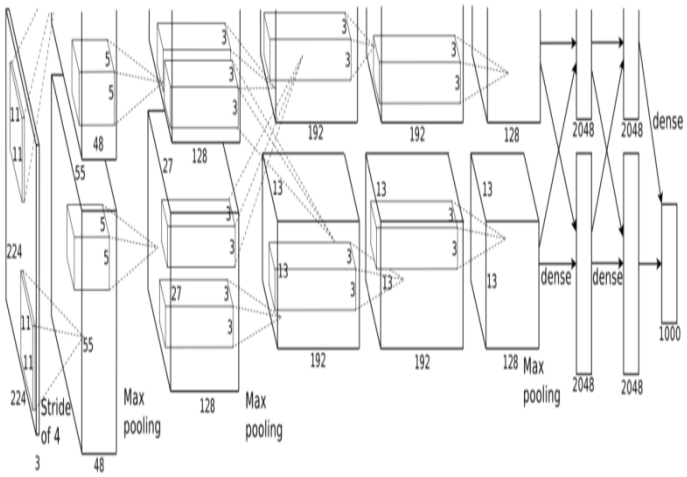
Neural Network, a trending and very useful technology while in the projects for future predictions. A neural network is a distributed processor that consists of artificial neurons as primary processing elements. Neural networks can be used for many applications including pattern classification, function approximation, clustering, prediction/forecasting, optimization, content addressable memory.

The main goal of the project is the implementation of a human recognition system with a camera or video stream. The application must be able to recognize people in real time. As an approach to solving this problem, deep neural networks were chosen as one of the popular methods for solving recognition problems.

compared with targets for determination of the magnitude of errors that are then used in the adjustment of the network weights .The SSD approach is based on a feed-forward convolutional network that produces a fixed size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. The MobileNet network architecture is a special class of convolutional neural models that are built using depth-wise separable convolutions and are therefore more lightweight in terms of their parameter count and computational complexity. These parameters can be used to directly influence the latency vs accuracy of the network depending on the end requirements of the user thereby detecting and counting the humans moving in the video

Advantages of proposed system:-
Efficient outcome as result with more accuracy.
Used to recognize humans in difficult abnormal situations.
Provide the count of humans moving in or out in a video.

SYSTEM DESIGN
OVERVIEW OF SYSTEM DESIGN



System Design
Figure. 1

In the above Process diagram, it shows a detailed explanation how our Project takes the Input and processes that and gives the Results. Here we explain it in the stepwise format,

In order to give Input to the System/Project we have to gather data from different repositories. One among the repositories that we used is "Pedestrian Detection", where it gives the human related Data like (human images in diff erent locations).

The Mobile Net model was proposed by Google and is a type of base architecture highly suitable for embedded-based vision applications with less computing power. The Mobile Net architecture uses depth wise separable convolutions instead of standard convolution. This reduces the number of parameters significantly as compared to the network with normal convolution with the same amount of depth in the network, which results in lightweight deep neural networks. The activation function "Re LU " is replaced by "ReLU6", and the "Batch Normalization" layer.

In the above Process diagram, it shows a detailed explanation how our Project takes the Input and processes that and gives the Results. Here we explain it in the stepwise format.

In order to give Input to the System/Project we have to gather data from diff erent repositories. One among the repositories that we used is "Pedestrian Detection", where it gives the human related Data like (human images in diff erent locations).

ADVANTAGES:

- Reduce ubnormal acitivities
- To reduce cases of various viruses.
- To maintain safety regulations.

IV. IMPLEMENTATION

Alexnet has eight unreadable layers. In addition to the output layer, the model consists of five layers with a combination of large integration followed by three fully integrated layers, all of which are used to activate Relu. It has been found that using a relaxed exercise increases the speed of the training process by about six times. They also use stop layers to prevent overlap in their models. The model is also trained using the Imagenet database. The Imagenet database contains approximately 14 million images divided into 1,000 categories.

Alexnet is a deep structure; the authors added padding to keep feature maps at a minimum. Images with size 227X227X3 are used as inputs for this model.

Layer	# filters / neurons	Filter size	Stride	Padding	Size of feature map	Activation function
Input	-	-	-	-	227 x 227 x 3	-
Conv 1	96	11 x 11	4	-	55 x 55 x 96	ReLU
Max Pool 1	-	3 x 3	2	-	27 x 27 x 96	-
Conv 2	256	5 x 5	1	2	27 x 27 x 256	ReLU
Max Pool 2	-	3 x 3	2	-	13 x 13 x 256	-
Conv 3	384	3 x 3	1	1	13 x 13 x 384	ReLU
Conv 4	384	3 x 3	1	1	13 x 13 x 384	ReLU
Conv 5	256	3 x 3	1	1	13 x 13 x 256	ReLU
Max Pool 3	-	3 x 3	2	-	6 x 6 x 256	-
Dropout 1	rate = 0.5	-	-	-	6 x 6 x 256	-

Figure: 2

We then apply the first layer of convolution with 96 size 11X11 filters per stride 4. The activation function used in this layer is the rest. Output feature map is 55X55X96. In that case, you do not know how to calculate the output size of the convolution layer.

Output = ((Input filter size) / step) +1

In addition to that opening function used is a relax. Now the output size remains unchanged i.e. 13X13X384. After that, we have a final 3X3 size conversion layer with 256 such filters. The stride and padding are set to one and the opening function is relax. The result map is 13X13X256.

Layer	# filters / neurons	Filter size	Stride	Padding	Size of feature map	Activation function
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-
Dropout 1	rate = 0.5	-	-	-	6 x 6 x 256	-
Fully Connected 1	-	-	-	-	4096	ReLU
Dropout 2	rate = 0.5	-	-	-	4096	-
Fully Connected 2	-	-	-	-	4096	ReLU
Fully Connected 3	-	-	-	-	1000	Softmax

Figure : 3

After this, we have our first dropout layer. The drop-out rate is set to be 0.5.

Then we have the first fully connected layer with a relu activation function. The size of the output is 4096. Next comes another dropout layer with the dropout rate fixed at 0.5. This followed by a second fully connected layer with 4096 neurons and relu activation.

Finally, we have the last fully connected layer or output layer with 1000 neurons as we have 10000 classes in the data set. The activation function used at this layer is Softmax. This is the architecture of the Alexnet model. It has a total of 62.3 million learnable parameters.

V. RESULTS:

Extensive surveys are conducted with CNN-based material finders. In tests, it was found that CNN-based acquisition models were better with accuracy than others. In some cases, it produces false positive effects when working with real-time video sequences. In the future, acquisition of different modern materials such as RCNN, Faster RCNN, SSD, RFCN, YOLO, etc. they can also be used with the data they have created to increase detection accuracy and reduce favorable false positives. Additionally, a single view obtained from a single camera cannot show a very effective result. Therefore, the proposed algorithm may be set to different views by multiple cameras in the future to get more accurate results

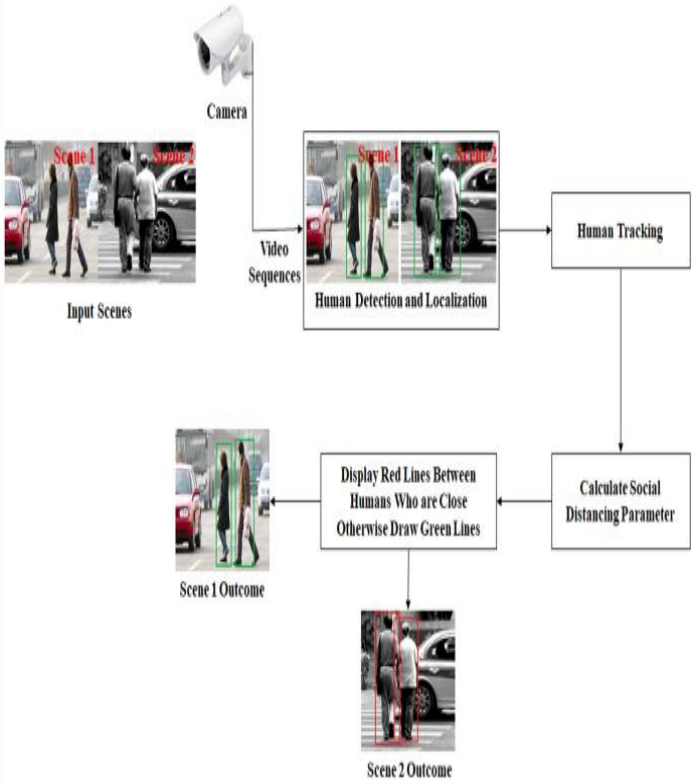
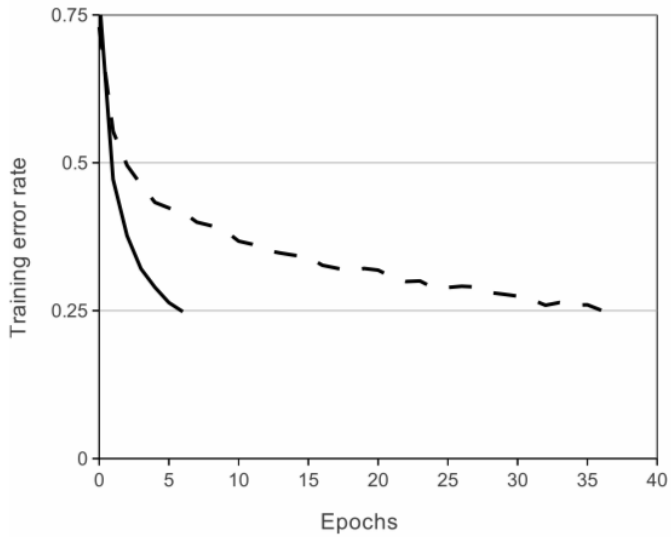


Figure: 3

BLOCK DIAGRAM OF PROPOSED SYSTEM

A dataset made of more than 15 million high-resolution images labeled with 22 thousand classes. The key: web-scraping images and crowd-sourcing human labelers. ImageNet even has its own competition: the Large-Scale Visual Recognition Challenge (ILSVRC). This competition uses a subset of ImageNet's images and challenges researchers to achieve the lowest top-1 and top-5 error rates (top-5 error rate would be the percent of images where the correct label is not one of the model's five most likely labels). In this competition, data is not a problem; there are about 1.2 million training images, 50 thousand validation images, and 150 thousand testing images. The authors enforced a fixed resolution of 256x256 pixels for their images by cropping out the center 256x256 patch of each image.

The below graph shows between the training error rate and Epochs.



VI. CONCLUSION:

In this deep learning project we have proposed a system which is used to recognize humans and also gives the number of humans moving in or out in a captured video which is the detection object counting and tracking in a frame .The bulk amount of data which is collected is trained accordingly by using convolutional neural networks for the prediction. At present we are having many applications which can capture humans as well as other objects in the surroundings using a camera or CCTV in different situations, not only present but also in the future , but they are just for video surveillance . In our system the neural network is trained with data which can predict the person with great accuracy. We choose Convolutional Neural Networks because it is a self trained network. The Mobilenet-SSD architecture uses a series of CNN layers for detection. And further a tracking algorithm is used for tracking the humans moving who are detected. We are able to extend our project by adding a few features for the future also.

REFERENCES:

[1]. Balamurugan, S., Ayyasamy, A. & Joseph, K.S. Enhanced petri nets for traceability of food management using internet of things. Peer-to-Peer Netw. Appl. 14, 30–43 (2021). https://doi.org/10.1007/s12083-020-00943-0
[2]. Balamurugan, S., Ayyasamy, A. & Joseph, K.S. IoT-Blockchain driven traceability techniques for improved safety measures in food supply chain. Int. j. inf. tecnol. 14, 1087–1098 (2022). https://doi.org/10.1007/s41870-020-00581-y
[3]. Worldometer. COVID-19 coronavirus pandemic. In: https://www.worldometers.info/corona-virus/, 2020, [Accessed 10 June 2020].
[4]. World Health Organization (2020) Coronavirus disease (COVID-19) advice for the public. In: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public. [Accessed 10 June 2020].
[5]. Adlhoch C, et al. (2020) Considerations relating to social distancing measures in response to the COVID-19 epidemic. European Centre for Disease Prevention and Control.
[6]. Singhal TA (2020) Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr 87:281–286. https://doi.org/10.1007/s12098-020-03263-6
[7]. Singh DK, Kushwaha DS (2016) Tracking Movements of Humans in a Real-Time Surveillance Scene. In: Pant M, Deep K, Bansal J, Nagar A, Das K (eds) Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing. Springer, Singapore
[8]. Abughalieh, Karam, Shadi Alawneh (2020) Pedestrian orientation estimation using CNN and depth camera. No. 2020–01–0700. SAE technical paper.
[9]. Jiao L et al (2019) A survey of deep learning-based object detection. IEEE Access 7:128837–128868. https://doi.org/10.1109/ACCESS.2019.2939201
[10]. Hsu FC, Gubbi J, Palaniswami M (2013) Human head detection using histograms of oriented optical low in low quality videos with occlusion. In: 2013, 7th International conference on signal processing and communication systems (ICSPCS), IEEE, pp 1–6.
[11]. Bahri H, Chouchene M, Sayadi FE et al (2019) Real-time moving human detection using HOG and Fourier descriptor based on CUDA implementation. J Real-Time Image Proc. https://doi.org/10.1007/s11554-019-00935-1.6
[12]. Gaikwad V, Lokhande S (2015) Vision based pedestrian detection for advanced driver assistance. Proc Comput Sci 46:321–328
[13]. Choudhury SK, Sa PK, Padhy RP, Sharma S, Bakshi S (2018) Improved pedestrian detection using motion segmentation and silhouette orientation. Multimedia Tools Appl 77(11):13075–13114
[14]. Seemanthini K, Manjunath S (2018) Human detection and tracking using hog for action recognition. Proc Comput Sci 132:1317–1326
[15]. Zhu A, Wang T, Qiao T (2019) Multiple human upper bodies detection via candidate-region convolutional neural network. Multimedia Tools Appl 78(12):16077–16096
[16]. Singh DK, Paroothi S, Rusia MK, Ansari MA (2020) Human crowd detection for city wide surveillance. Proc Comput Sci 171:350–359. https://doi.org/10.1016/j.procs.2020.04.036

[17]. Gajjar, Vandit, Ayesha Gurnani, Yash Khandhediya (2017) Human detection and tracking for video surveillance: A cognitive science approach. In: Proceedings of the IEEE International Conference on Computer Vision Workshops.
[18]. Najva N, Edet Bijoy K (2016) SIFT and tensor based object detection and classification in videos using deep neural networks. Procedia Comput Sci 93:351–358.
[19]. Agnes SA, Anitha J, Pandian SIA et al (2020) Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J Med Syst 44:
[20]. Zhang H, Hong X (2019) Recent progresses on object detection: a brief review. Multimedia Tools Appl 78(19):27809–27847
[21]. Chahyati D, Fanany MI, Arymurthy AM (2017) Tracking people by detection using CNN features. Procedia Comput Sci 124:167–172. https://doi.org/10.1016/j.procs.2017.12.143
[22]. Ansari, Mohd Ali, Dushyant Kumar Singh (2018) Review of deep learning techniques for object detection and classification. In: international conference on communication, networks and computing. Springer, Singapore.
[23]. U. Ojha, U. Adhikari and D. K. Singh (2017) Image annotation using deep learning: A review. International Conference on Intelligent Computing and Control (I2C2). Coimbatore, 2017, pp. 1–5.
[24]. Marquez ES, Hare JS, Niranjan M (2018) Deep cascade learning. IEEE Trans Neural Netw Learn Syst 29(11):5475–5485
[25]. Wang D et al (2019) Daedalus: breaking non-maximum suppression in object detection via adversarial examples. arXiv 6:1945–1902
[26]. Ansari MA, Dixit M (2017) An enhanced CBIR using HSV quantization, discrete wavelet transform and edge histogram descriptor. International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, pp. 1136–1141.
[27]. INRIA Person Dataset (2018) Available:http://pascal.inrialpes.fr/data/human/.
[28]. R. Jayaswal, J Jha (2017) A hybrid approach for image retrieval using visual descriptors, 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 1125–1130, doi: https://doi.org/10.1109/CCAA.2017.8229965.
[29]. Zhang H, Parker LE (2016) CoDe4D: color-depth local spatio-temporal features for human activity recognition from RGB-D videos. IEEE Trans Circuits Syst Video Technol 26(3):541–555. https://doi.org/10.1109/TCSVT.2014.2376139
[30]. Zhang D, Zhou J, Guo M, Cao J, Li T (2011) TASA: tag-free activity sensing using RFID tag arrays. IEEE Trans Parallel Distrib Syst 22(4):558–570. https://doi.org/10.1109/TPDS.2010.118
[31]. Zhang M, Sawchuk AA (2012) USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: UbiComp'12—proceedings of 2012 ACM conference on ubiquitous computing, pp 1036–1043
[32]. Zhou X, Liang W, Wang KIK, Wang H, Yang LT, Jin Q (2020) Deep-learning-enhanced human activity recognition for internet of healthcare things. IEEE Internet Things J 7(7):6429–6438. https://doi.org/10.1109/JIOT.2020.2985082