

# DEEP LEARNING BASED VIDEO ANALYTICS FOR PERSON TRACKING

Vignesh Kanna, J.S.

*Department of Electronics &  
Communication Engg.,  
Thiagarajar College of Engg.,  
Madurai Tamilnadu India  
vigneshkanna@student.tce.edu*

Ebenezer Raj, S M

*Department of Electronics &  
Communication Engg.,  
Thiagarajar College of Engg.,  
Madurai Tamilnadu India  
ebenezer@student.tce.edu*

Meena, M

*Department of Electronics &  
Communication Engg.,  
Thiagarajar College of Engg.,  
Madurai Tamilnadu India  
meena@student.tce.edu*

Meghana, S

*Department of Electronics &  
Communication Engg.,  
Thiagarajar College of Engg.,  
Madurai Tamilnadu India  
meghana@student.tce.edu*

Md. Mansoor Roomi, S

*Associate Professor, Department of  
Electronics & Communication Engg.,  
Thiagarajar College of Engg.,  
Madurai Tamilnadu India  
smmroomi@tce.edu*

**Abstract—** As the assets of people are growing, security and surveillance have become a matter of great concern today. When a criminal activity takes place, the role of the witness plays a major role in nabbing the criminal. The witness usually states the gender of the criminal, the pattern of the criminal's dress, facial features of the criminal, etc. Based on the identification marks provided by the witness, the criminal is searched for in the surveillance cameras. Surveillance cameras are ubiquitous and finding criminals from a huge volume of surveillance video frames is a tedious process. In order to automate the search process, proposed a novel smart methodology using deep learning. This method takes gender, shirt pattern, and spectacle status as input to find out the object as person from the video log. The performance of this method achieves an accuracy of 87% in identifying the person in the video frame.

**Keywords—** Image Classification, Deep Learning, Convolutional Neural Network, Pattern Recognition, Video Analytics

## I. INTRODUCTION

Recent growth in the assets of people has increased the malicious activities amongst them in real world. Crime is an illegal act that is to be punished. Many ideas are flooding from every nook and corner of the world to control and solve these acts. Research teams are striving to find out methods to quickly identify the criminal. As technology has reached far and wide across the continents, even laypersons have started to use technology for surveillance. The most common surveillance device is a video camera. Due to drastic improvements in technology, we can find the surveillance cameras have become ubiquitous. In case of any criminal activity in an area, the law protectors first pick out the surveillance camera footage from the surveillance cameras in and around the crime scene. To do criminal forensics, the law protectors not only see the footages at the time of the crime but also the previous days' footage to find out if the criminal has come to plan out his crime. When a criminal activity takes place, the role of the witness plays a major role in nabbing the criminal. The witness usually states the gender of the criminal, the pattern of the criminal's dress, facial features

of the criminal, etc. Based on the identification marks provided by the witness, the criminal is searched for in the surveillance camera footages. The footages are huge in size and require time and effort to find out the criminal based on the clues provided by the witness [1]. The manual way of surveillance footage checking is prone to error and distractions [2].

Various types of witnesses will give their statements from alternative points of view, and these views should be evaluated by the criminologist to set up the unwavering quality of the proof given. The testimony of a witness can lose validity due to too many external stimuli that may influence what was seen during the wrongdoing, and consequently impede memory. It is of particular interest that the memory of a witness can progress toward becoming undermined by other information, with the end goal that a person's memory winds up one-sided. During these cases and during a time when witnesses can't retrieve features of criminal's face, other basic descriptions of a human-like his/her gender, attire and wearing may be helpful. Hence, we propose the usage of some of the basic information to track the malefactors.

To automate the search process and avoid the human errors, we have proposed a novel smart methodology using deep learning. Our methodology also takes up the basic first-hand information like gender, shirt pattern, and spectacle status to find out a person in the video log.

The rest of our paper is organized as follows: section II gives a gist of the previous works done in the realm of image identification in the video; section III provides a detailed description of our proposed methodology; section IV gives the results we obtained during the course of experimentation; section V provides the conclusion and the future planned work.

## II. RELATED WORK

Jing et al. [3] have used a constrained optimization process named deformation contour to humans wearing spectacles in the given set of images. The authors have used a three-step procedure namely edge map producer,

filtering unnecessary edge points, and optimization of the shape and position of the spectacles.

Jia et al. [4] have used the phase congruency method to detect the spectacles in the face of human beings. The authors have first chosen only the upper part of the face and used the elliptical model to choose the human eye part of the face. Once the authors have retrieved the eye part, the authors have used the phase congruency method to detect the spectacles pattern of the human.

Lorenzo et al. [5] have used the Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) to classify different attributes of clothes such as length of the sleeve, the color of the fabric, pattern in the fabric and existence of collar in the cloth. The authors have used two LBP variants namely uniform LBP and original LBP.

Bregler et al. [6] have proposed a method to identify human in the video sequence. The authors have used Markov estimation, recursive Kalman, EM algorithm, and mixture models to track humans even during their complex movements in the video sequence.

Chen et al. [7] have used five different Convolutional Neural Network (CNN) models namely conventional CNN, inception module-based CNN, residual block and inception module-based CNN, and two transfer learning-based CNNs for classification of clothing images.

Bhatnagar et al. [8] have used CNN to classify fashion article images in the Fashion-MNIST dataset. The authors have proposed three CNN architectures that use residual skip connections and batch normalization to speed up and ease the process of learning.

Meghdadi et al. [9] have proposed a method to analyze the video data. The authors have dealt with the method to find out specific events in video frames. The authors' prime motive is to reduce the browsing time to search for a specific image frame. The authors have done spatial and temporal filtering by plotting the trajectory of a still image from the video in space-time cube.

Seo et al. [10] have used CNN based architecture to classify the apparel images. The authors have used the transfer learning method to tackle the issues of a small dataset. The authors have pre-trained using the GoogleNet architecture and fine-tuned the last layer of connected neurons with their fashion images dataset.

Gavai et al. [11] have proposed the usage of the MobileNets model on the TensorFlow platform to minimize the storage space and time to train the model for the classification of images.

Krizhevsky et al. [12] have proposed the usage of deep CNN with 5 layers and 650,000 neurons to place the 1.2 million images of high-resolution from the ImageNet LSVRC-2010 contest in one of the 1000 classes. The authors have accelerated the training process by using non-saturating neurons, max-pooling, and soft-max functions.

Most of the existing methods have trained the system to detect a person with a single entity such as either apparel pattern alone or spectacles alone. Searching for a person with only apparel information or spectacle status is not very efficient. When training is done with more entities, it is rather easy to identify a person with improved precision. Thus, to improve the efficacy of searching a

person in the video data, we have trained the system with a combination of three entities of a person.

### III. PROPOSED METHODOLOGY

An overview of our proposed methodology is given in figure 1. As seen in the figure 1, we have used a combination of annotation and MobileNet to train the system to detect images in the video data. We are having 8 different classes to detect a person in video data.

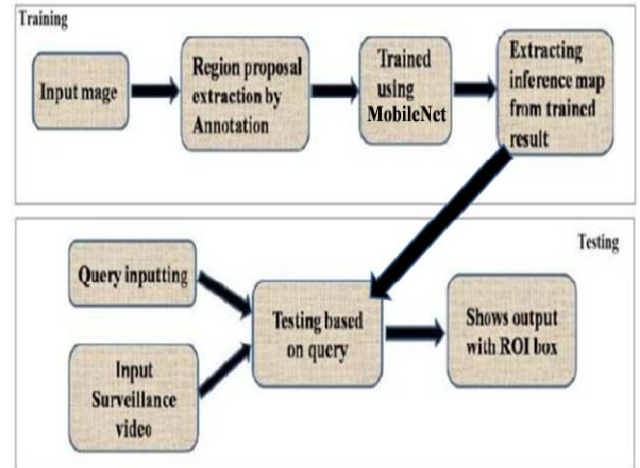


Fig. 1. Proposed Methodology

#### A. Pre-processing

Collected dataset is annotated and the result is converted to the binary holding file as this binary record viably diminishes the training time and testing dataset size.

#### B. Training

For training the MobileNet has been used with R-CNN as a platform. R-CNN creates these bounding boxes, or region proposals, utilizing a procedure called Selective Search. Selective Search looks at the image through windows of various sizes, and for each size attempts to bunch contiguous pixels by texture, color, or intensity to identify objects and then for purpose of feature extraction and training a pre-built model of mobilenet is used.

#### R-CNN:

The training is based on the Selective search algorithm of RCNN that identifies the patterns in the image based on the varying scales, colors, textures, and enclosure which are the four basic regions of an object. A pre-trained convolutional neural network is used and then, this model is retrained by modifying the last layer of the network based on the eight different classes that need to be detected. Thus this process extracts 2,000 regions from each image based on selective search. The entire process of object detection using RCNN has three models:

1. CNN for feature extraction
2. Linear SVM classifier for identifying objects
3. Regression model for tightening the bounding boxes.

### MobileNet

This architecture uses depthwise separable convolutions which exiguently diminishes the number of parameters when compared to the network with normal convolutions with the same depth in the network resulting in the light weight deep neural networks. This depthwise separable convolution means replacement of depthwise convolution followed by pointwise convolution.

- In the normal convolution, if the input feature map is of  $H_i, W_i, C_i$  dimension and we want  $C_o$  feature maps with convolution kernel size  $K \times K$  then there are  $C_o$  convolution kernels each with dimension  $K, K, C_i$ . This results in a feature map of  $H_o, W_o, C_o$  dimension after convolution operation.
- In the depthwise separable convolution, if the input feature map is of  $H_i, W_i, C_i$  dimension and we want  $C_o$  feature maps in the resulting feature map and the convolution kernel size is  $K \times K$  then there are  $C_i$  convolution kernels, one for each input channel, with dimension  $K, K, 1$ . This results in a feature map of  $H_o, W_o, C_i$  dimension after depthwise convolution. This is followed by pointwise convolution [ $1 \times 1$  convolution]. This convolution kernel is of dimension  $1, 1, C_i$  and there are  $C_o$  different kernels which results in the feature map of  $H_o, W_o, C_o$  dimension. (from [13]).

So, here the mobilenet version2 is used as the base network for feature extraction. It has three convolutional layers, the last two are depthwise convolution that filters the inputs, followed by a  $1 \times 1$  pointwise convolution layer.

This does the opposite of its former versions by making the number of channels smaller instead of doubling or increasing them and so this is named as projection layer. This projects data with a high number of dimensions (channels) into a tensor with a much lower number of dimensions. Thus this bottlenecks the amount of data that flows through the network resulting in a fewer multiplications in the convolutional layers. And here the first layer expands the number of channels in the data before it goes into the depthwise convolution. Hence, this **expansion layer** always has more output channels than input channels pretty much doing the opposite process of the projection layer. MobileNet version2 also provides with a **residual connection** to help with the flow of gradients through the network. Each layer has batch normalization and the activation function is ReLU6. Thus the full MobileNet version2 architecture, consists of 17 of these 3 layers in a row which are followed by a regular  $1 \times 1$  convolution, a global average pooling layer, and a classification layer.

Thus this human detection code is an aggregate of smaller code fragments that individually does the steps of our proposed work i.e. detecting whether the person in the video frame is wearing a plain or checked shirt, detecting whether the person in the video frame is wearing spectacles or not and detecting whether the

person in the video frame is male or female.

Figure 2 shows an illustration of the process that takes place in our methodology to detect human using the shirt pattern.

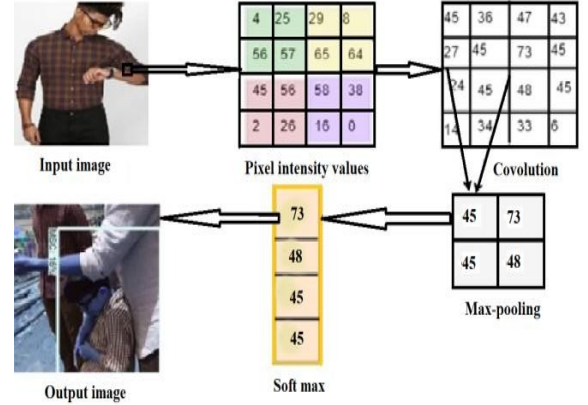


Fig. 2. Illustration of our methodology

## IV EXPERIMENTAL RESULTS AND DISCUSSION

### A. Dataset

The pre-trained database is the backbone for image classification or detection employing deep learning. The number of images in the dataset plays a vital role in determining the accuracy of the model. As image detection and classification in live video frames requires fine-tune pattern extraction and more convolution layers in CNN, a large number of pre-labeled data are required. Around 2,400 images were collected from the internet by various sources like Google Image, Bing, Yahoo, etc. The various images were further expanded by augmenting them. Finally, a dataset was created with 8 classes having 400 images each. 70% of images were used for training, 10% for validation and 20% for testing. The sample image dataset is shown in Figure 3.



Fig. 3. Sample images of the dataset



### B. Test framework

We used TensorFlow [13], OpenCV [14] and MobileNet Single Shot Detection (SSD) to implement our methodology. TensorFlow is an open-free source available for deep-learning, data-science and differential programming supported across various programming languages. It is predominately used with a python programming language as it is easy to understand and time effective. It also supports machine learning applications and CNN. We have used the GPU version of TensorFlow because it takes less time to train large datasets of the images than the CPU version of tensor flow. In addition to TensorFlow, we ran our proposed work in the workstation having a CUDA toolkit and Intel graphic card of 5GB. OpenCV is one of the main libraries for Image processing and is supported by many programming languages. OpenCV aims at real-time environmental computer vision. This is free open source software that supports python deep learning frameworks like TensorFlow, Torch, and Caffe. It has pre-built methods that make it easier for programmers to enhance image processing technology. Facial recognition, gesture recognition, object identification were the main three applications that we used from OpenCV. In our proposed work we use SSD MobileNet pre-built configuration file to increase the training speed. The neural network process is accompanied by the MobileNet library to help in object detection and recognition of objects. SSD helps in the multi-box detector in testing which draws a bounding box around the detected persons based on queries. The whole testing was done in HP xw8600 workstation having 16Gb DDR RAM, 1TB hard-disk and graphics card corresponding to Nvidia P2000 5Gb RAM Graphical Processing Unit. The 20,000 steps training of the dataset took over 2 hours using TensorFlow-GPU API. A gist of code development flow is shown in Figure 4

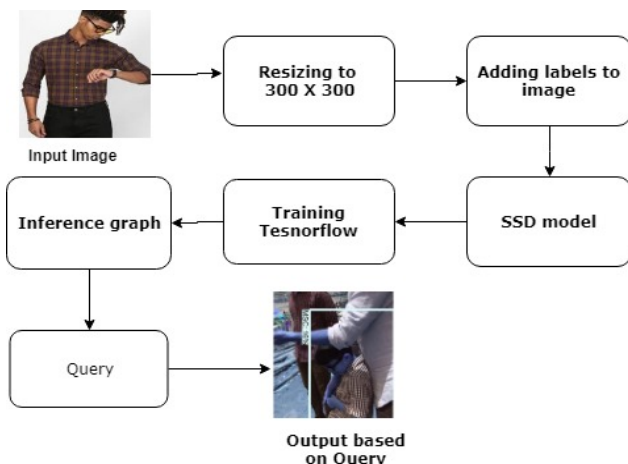


Fig. 4. Proposed methodology code development flow

### C.Results

The SSD MobileNet was compared with other configurations namely SSD inception and Faster rcnn. The results are shown in figure 6. As seen in Figure 5, SSD MobileNet was better and it processed a higher number of frames per second than the other two configurations

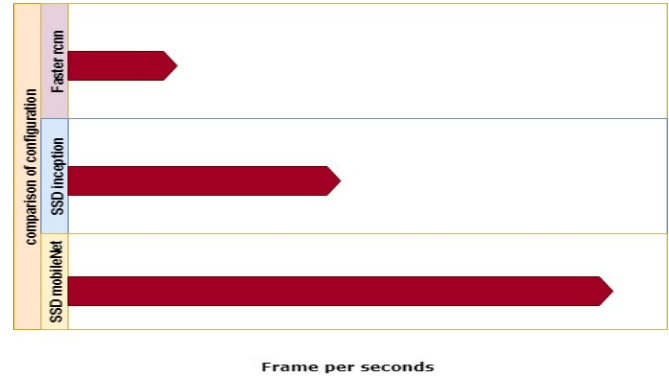


Fig. 5. Comparison between the various configuration file

Finally, with the help of the trained model and configuration file, the detection code was tested in large size video frames. Further, a ptxt file of label-map was enclosed in the testing program. On running the testing software, the user interface asks the user to choose the class to which the output bounding box must be shown on the display. Based on the user query the output video frames were displayed with a person detected on chosen class. Figure 6 shows some results that we obtained during testing. The results show that, the practice of large size dataset and properly built SSD MobileNet configuration file gave better performance on human detection on different lighting, environment and different events. Due to nearly the same size of weight employed in each class, it becomes easy to detect the human in video frames with low loss.

The confusion matrix of proposed model is shown in Figure 7. Our model showed an accuracy of 87% and an average loss of 1.4. The output frames were displayed with the bounding box and frame numbers corresponding to the frame having the detected object were also displayed. The model was tested in different environmental videos having different class of humans. The output frame is shown as Figure 8.







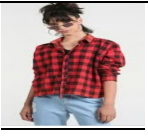
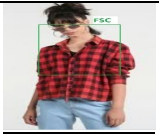
ORIGINAL IMAGE	LABEL	OUTPUT BASED ON QUERY
	Male-Spec-Check	
	Male-Spec-Plain	
	Female-NoSpec-Check	
	Female-Spec-Check	

Fig. 6. Results obtained during testing

Confusion Matrix	FSC	FNCS	FSP	FNPS	MSC	MNSC	MSP	MNSP
FSC	0.88	0.03	0	0	0.04	0.5	0	0
FNCS	0.4	0.95	0.1	0	0	0	0	0
FSP	0.2	0	0.94	0	0	0	0.3	0.1
FNPS	0	0	0	0.97	0	0	0.3	0
MSC	0	0	0	0	0.99	0	0.1	0
MNSC	0	0	0	0	0.1	0.96	0	0.3
MSP	0	0	0.2	0	0	0	0.98	0
MNSP	0	0	0	0.3	0	0.2	0	0.95

Fig. 7. Confusion Matrix

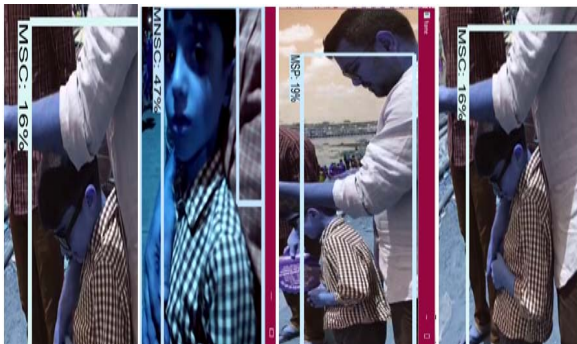


Fig. 8. Output video frames based on Queries

## V CONCLUSION AND FUTURE WORK

In this paper, a smart human detection in video based on gender, checked or unchecked shirt, spectacle or no spectacle criteria was proposed. We used deep learning CNN for classification. Our model detected humans in the video with an accuracy of 87%. Hence, this can be useful in nabbing criminals from the surveillance footage based on the identification marks provided by the witness. As a future direction, we plan to add extra criteria like type of hair the person has like curly or straight, color of the hair, eye color, etc. Adding the detection entities will making it easier for patrol policemen to search an occurrence in video frames. The output from the trained model can be digitized and interfaced with the Arduino hardware kit to give automated alarm warning when an unauthorized person enters the obstructed areas

## REFERENCES

- [1] A. Saglam and A. Temizel, "Real-Time Adaptive Camera Tamper Detection for Video Surveillance," *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, Genova, 2009, pp. 430-435.
- [2] D. Gutchess, M. Trajkovics, E. Cohen-Solal, D. Lyons and A. K. Jain, "A background model initialization algorithm for video surveillance," *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vancouver, BC, Canada, 2001, pp. 733-740 vol.1.
- [3] Zhong Jing and Robert Mariani, "Glasses detection and extraction by deformable contour," *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Barcelona, Spain, 2000, pp. 933-936 vol.2.
- [4] X. Jia and J. Guo, "Eyeglasses removal from facial image based on phase congruency," *2010 3rd International Congress on Image and Signal Processing*, Yantai, 2010, pp. 1859-1862.
- [5] J. Lorenzo, M. Castrillón, E. Ramón, and D. Freire. "Evaluation of LBP and HOG descriptors for clothing attribute description," *Proceedings of Video Analytics for Audience Measurement in Retail and Digital Signage (VAAM)*, 2014.
- [6] C. Bregler, "Learning and recognizing human dynamics in video sequences," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, USA, 1997, pp. 568-574.
- [7] L. Chen, R. Han, S. Xing and S. Ru, "Research on Clothing Image Classification by Convolutional Neural Networks," *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Beijing, China, 2018, pp. 1-5.
- [8] S. Bhatnagar, D. Ghosal and M. H. Kolekar, "Classification of fashion article images using convolutional neural networks," *2017 Fourth International Conference on Image Information Processing (ICIIP)*, Shimla, 2017, pp. 1-6.
- [9] A. H. Meghdadi and P. Irani, "Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization," in *IEEE Transactions on*

*Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2119-2128, Dec. 2013.

[10] Y. Seo and K. Shin, "Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network," *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, Shanghai, 2018, pp. 387-390.

[11] N. R. Gavai, Y. A. Jakhade, S. A. Tribhuvan and R. Bhattad, "MobileNets for flower classification using TensorFlow," *2017 International Conference on Big Data, IoT and Data Science (BIG)*, Pune, 2017, pp. 154-158.

[12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, 2012, pp. 1097-1105

[13] Andrew G. Howard Menglong Zhu Bo Chen Dmitry Kalenichenko Weijun Wang Tobias Weyand Marco Andreetto Hartwig Adam "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications"

[14] <https://opencv.org/>

[15] <https://www.tensorflow.org/>