International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.SP203

Scopus

Original Research

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

A Real-Time Deep Transfer Learning-Based Classification and Social Distance Alert Framework Based on Covid-19

Anurag Singh, Naresh Kumar, Tapas Kumar

School of Computing Science, Galgotias University, Yamuna Expressway, Greater Noida, Gautam Buddha Nagar, India Uttar Pradesh, India

ABSTRACT

Introduction: Covid-19 is a novel virus that has exponentially increased the number of infected persons and the death of human beings into millions within a few months. This virus spreads when a person comes into contact with another person, coughing, sneezing and droplets.

Objective: To avoid loss of lives, human direct assistance and early precaution, automated systems required for reducing the number of cases. Deep Learning can facilitate human life much better way by automating human visual intelligence into machine intelligence.

Methods: In this novel research work, we are implementing transfer learning methodology to improve the learning of a related objective task on top of base deep learning model in developing a mask/non-mask detection model along with changing the hyperparameters and data augmentation technique by using less input dataset for smart healthcare, smart home in reducing and detecting corona cases.

Results: We used object detection model Single Shot Multi-box Detector and classification model mobile net, which achieved significant accuracy and much faster for both training and inference with prediction accuracy of 87% with IOU=.75 on our own created trained dataset comparable with other real-time object detection model such as Faster Regional Convolutional Neural Network by tuning the hyperparameters.

Conclusion: The automated system not only reduces the false alarm but also enhanced the performance accuracy by detecting the mask and non-mask due to which the number of covid-19 cases can be reduced at an early stage.

Key Words: Single Shot Multi-box Detector, Convolutional Neural Network, Transfer Learning, Image Annotation, Deep Learning, Covid-19

INTRODUCTION

World at this stage facing a huge pandemic due to the novel coronavirus Covid-19. Millions of peoples are suffering and millions of people have lost their lives due to this virus. The main guidelines by various health agencies are to keep social distancing, wearing a mask and remain in isolation. But still, people are not following the guidelines and no proper precautions, due to which the number of cases is increasing exponentially.[1]

Governments using surveillance drones to capture crowd and tracking of human in public places to avoid mass gathering which are initial steps to fight against covid-19. Currently, these devices require human assistance to take decisions. For immediate action and accuracy in detection, technologies will play a vital role in the smart healthcare area to avoid the spreading of covid-19 in the community.[2-5]

For the last many years, many researchers have worked on machine learning and computer vision-based algorithms for multiple applications such as driverless cars, healthcare systems, agriculture domain, medical imaging, surveillance and various automation systems. For such applications, Deep Learning (DL) and Machine Learning (ML) are the key domains, which are inspired by the human brain for classifying and solving complex problems.[6-8] For all vision-based applications, computer vision plays a vital role in bridging a semantic gap between human understandings with the help

Corresponding Author:
Singh Anurag, School of Computing Science, Galgotias University, Yamuna Expressway, Greater Noida, Gautam Buddha Nagar, Uttar Pradesh, India; Phone: 9536841474; Email: anurag.singh485@gmail.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 11.02.2021 **Revised:** 26.03.2021 **Accepted:** 12.05.2021 **Published:** 11.06.2021

of tagging and labelling the images in identifying the objects in real-time scenario.[9]

One of the major areas which require more attention and deployment is the advanced surveillance system for smart healthcare systems, smart home security and defence systems but still, the surveillance field is facing some accuracy issues in identifying potential mask, classification and requires more training time to train the machine learning models along with huge datasets.

These topics are closely related to Human-Computer Interaction (HCI), which has developed more interaction between human and computer using Computer Vision (CV) framework and increased its popularity in both academics, healthcare and surveillance systems.[10] Human being perceives information from sense organs and its perception varies from person to person in terms of features such as shape, colour and size. Based on the learning and experience, the human being classifies the object and annotates it either in the form of memories or it makes a dataset using hands and verbal responses for tagging the image frames.[11,12]

All the features in traditional object detection methods are handcrafted and shallow trainable, due to which it has restrictions for a few application and detections while testing. Whereas in deep learning the features and classification are done automatically by the neural network but it required more training with a huge amount of data along with more computational time for training the machines.[13,14]

Literature Survey

Deep Learning Model in Object Detection

Image annotations and tagging are tedious and time taking task required in computer vision related applications by researchers. The majority of tagging is used in large scale retrieval system, managing and organizing multimedia databases which are done manually. Classification of video is more focused on labelling of video clips dependent on their semantic substance such as human activities or complex events. Detection and recognition of objects are carried out by researchers in multiple ways.[15]

- Feature-Based Object Detection.
- Viola-Jones Object Detection.
- Support Vector Machine (SVM) Classifications with Histogram Orientation Gradient (HOG) & HAAR like features.
- Object Detection using deep learning.

But still, the performance of computer vision algorithm lags in many key domains, due to variations in viewpoints, different postures, occlusions and lighting conditions, which is difficult in accomplishing the object detection within the localization process (objects positions are located in a given image). Hence, researchers have more focused on this field in developing various applications by optimizing the algorithms and hyperparameters for better accuracy with less computational process and training.

In this exploration work, the objective is to simulate the human patrolling and decisions system into smart surveillance in detecting potential mask who have not to wear the mask in crowd area and mass gathering with the help of deep learning model and computer vision to replace man operated patrolling with advance automated surveillance system that can be deployed in smart healthcare, smart home for automated mask detection and defence applications, that can be installed at railway stations, Hospitals, Airports, Malls etc to avoid the spread of this virus. Our novel contribution in the research work is the use of very little dataset and optimization of hyperparameters for training the deep learning model using data augmentation technique and detecting the mask/non-mask person, bounding box, captions generation and human pose detection with high accuracy which can be done using transfer learning techniques, which is still not carried out by any researcher in this domain.

In 2012, Convolutional Neural Network (CNN) came into the picture which can represent high-level features and robustness. In deep learning, model object detection grouped into two-stage detection and one-stage detection. Abnormal behaviour detection in a smart surveillance system that majorly discussed in three sections. [16]

- Human detection and discrimination for a subject
- Module-based on posture classification
- Module-based on abnormal behaviour detection
- And the models used for the above three sections are as follows
- You only look once (YOLO) network
- VGG-16 Net
- Long short-term memory (LSTM)
- Single Shot multi-box Detection (SSD)

Researchers have discussed more intelligent video surveillance using deep learning techniques for crowd analysis.[17] Deep Learning techniques provides two major components; training the model and testing the model, which required a huge amount of data for better accuracy. Researchers have discussed in brief about recognition of the person and object detection automatically and deduction of complex events in two ways; low level and high level. People and object detection done under low level whereas the detection from low to high used for event detection.[18]

- Event modelling
- Action modelling
- Detection of action
- Modelling of complex events and detection

The above four are the major architecture used in modelling and detection. Whereas recognizing a system involves:

- Pre- Processing
- Feature Extraction
- Object Tracking
- Understanding of behaviour

The researcher has used a similar method in avoiding suicide attempts in prison using an RGB-D camera and analyzing the body joints which represents the suicidal behaviour. It was suggested that CNN is a better improvement than traditional Neural Network. In the paper computational resources were reduced using dimensionality reduction which happens in reducing computation of 1x1 Convolutional before 5x5 Convolutional.[19-22]

The selection of the best model was carried out based on the high mean Average Precision (MAP) used for the evaluation of the test dataset. As a result, the use of Faster R-CNN along with VGG-16 shows better performance in detecting drones but requires more computational power.[23] Most of the common object detection system follows the following pipeline:

- Potential Object Detection
- Bounding Box
- Feature Extraction
- Classify using good Classifier

Following are the datasets for object detection used by many researchers for object detection in deep learning.

- PASCAL Visual Object Class (PASCAL VOC)
- ImageNet Large Scale Visual Recognition (ILSVR)
- Microsoft Common Object in Context (MS COCO)

In the current scenario, two famous object detection algorithms are the centre of attraction among the researchers for applications based on real-time object detection, which was achieved by You Only Look Once (YOLO) and Single Shot MultiBox Detector. YOLO object detector completely works on region proposal and sliding window-based approach which has divided the images into a grid of cells and each cell predicts the class and bounding box of the object which provides final accuracy for the object class.

Whereas Single Shot MultiBox Detector is completely followed on feed-forward CNN that generates fixed size bounding box along with confidence scores of every class and produces final detection results. Regional based Convolution Neural Network (R-CNN) model is not able to achieve real-time object detection because of its time taking training process and inefficiency of region proposition. Whereas YOLO was developed for object detection and classification which requires single-step process. After 1 input image, it starts evaluating and predicting the class along with the bounding box. YOLO architecture is capable of achieving 45 FPS and YOLOv2 can achieve 244 FPS on CUDA GPU. The simultaneous process of bounding box prediction and class prediction makes YOLO different from other traditional systems.

YOLO and single-shot multi-box detector takes input images and divides them into the grid of S x S and defines a bounding box at each grid cell with a confidence score which is a probability of an object existing in each bounding box discussed in the equation. (2) where IOU is intersection over union and represent the fraction between 0 and 1.Its is an overlapping area between the predicted bounding box and ground truth and it should be close to 1.

$$C = \text{Probability}(\text{Object}) * \text{IOU}_{\text{predict}}^{\text{truth}} \quad (2)$$

Similarly, Class probability C also gets predicted for each grid cell simultaneously for the bounding box and class-specific probability for each grid cell is defined as (3):

$$\text{Probability}(\text{Class}_i|\text{Object}) * \text{Probability}(\text{Object})| \text{IOU}_{\text{predict}}^{\text{truth}}$$

$$= \text{Probability}(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}} \quad (3)$$

Transfer learning enables us to utilize knowledge from previously learned task and applying it to the new model as per research requirement for better performance. As most of the traditional machine learning algorithms are performs a specific task and requires a lot amount of data for training the model which requires more computation and training from scratch. By use of transfer learning the feature, weights can be used for other applications with less computation and data.

There are two principal ways to deal with executing move learning; they are:

- Weight Initialization.
- Feature Extraction.

The weights in re-utilized layers might be utilized as the beginning stage for the preparation procedure and adjusted in light of the new issue. This use treats transfer learning as a kind of weight initialization scheme. This might be helpful when the principal related issue has significantly more marked information than the issue of intrigue and the likeness in the structure of the issue might be valuable in the two settings.

The task t1 from the pre-trained model (weights, features) used for task t2 with fewer data.

Instead of training the CNN model from scratch for our mask warning system, we are using a pre-trained model initially which helps us in detecting mask and mask-based object for our domain and task.

A framework of transfer learning:

$$D = \{X, P(X)\} \quad (6)$$

D is the domain that defines two elements; X a sample data point and P(X) marginal data point.

For a given domain D, a task is defined by:

$$T = \{y, P(y|x)\} = (y, \eta)$$

$$Y = \{y_1, y_2, y_3 \dots \dots y_n\} \quad y \in yi \tag{7}$$

Label space: y

A predictive function: η

Learned from feature vector/label pairs x_i, y_i where $x \in xi, y \in Yi$

For each feature vector in the domain, η predicts its label $\eta(x_i) = y_i$

MATERIALS AND METHODS

In this research work, our novel contribution is to train the deep learning model with a one-shot training technique on images generated from real-time video cameras which are comprised of mask and non-mask faces along with detection of crowds and mass gathering, The trained model will play a vital role in avoiding corona covid-19 type epidemic in society at an early stage and currently have not used by any researcher. We have also optimized and fine-tune the hyperparameters for the pipeline created for this covid-19 application with the help of an inductive transfer learning technique that reduces the overfitting problem on the pre-trained model as a base task over our trained model displayed in Figure 1.

We have optimized the basic hyperparameter that is the dropout value which is also known as the regularization technique in optimizing the model. Initially, we have used TensorFlow and OpenCV for training a pre-trained model and detecting objects in a real-time surveillance system with the help of object detection algorithm single shot multi-box detector and Mobilenet as classification model over MSCOCO dataset with common features from 90 classes and can identify and performs multiple bounding boxes around the objects and compared with Faster Regional Convolutional Neural Network, which is more advance than a single-shot multi-box detector in terms of the number of classes and categories. But the major issue while training a Faster Regional Convolutional Neural Network model was more computational time taking than that of a single-shot multi-box detector model.

In the system, the number of classes/labels was changed to 2(Mask/Non-Mask) and the number of CNN layer was changed to 6 with RELU as activation function with minimum depth 16, batch size 128 and learning rate of 0.004 and 0.04, RMS prop for gradient descent on own dataset with 400 training examples for a person with face covered with cloth pretend to be a mask and not mask. For better performance of machine learning algorithms, hyperparameters play a key role. The followings are the hyperparameters, which are used in enhancing the accuracy of the deep learning-based object detection model.

Table 1: Hyperparameters for training model

Parameter Name	Parameter Type	Required Range
Learning Rate	Continuous parameter	min:1e-6, max:0.5
Batch size	Integer parameter	min:0.8, max:64
Momentum	Continuous parameter	min:0.0 ,max :0.1
Optimizer	Categorical parameter	[SGD, ADAM, RMSPROP]
Weight Decay	Continuous parameter	min:0.0, max;0.999

Table 1 Discuss the hyperparameters and their values are set before the training process in optimizing the performance of the model. Apart from the above parameters number of epochs, dropouts, hidden layers and units, Activation functions are also part of hyperparameters in building accuracy of the model. Optimization of the learning rate is done because it controls the weights after each batch size which helps our model to learn fast and accurate with minimization of the loss function. Optimization of batch size requires less memory because we are training our model on fewer sample data, So a large amount of data cannot be fit into our memory. So small batch size plays an important hyperparameter for the deep learning model so that weights can be updated after each propagation.

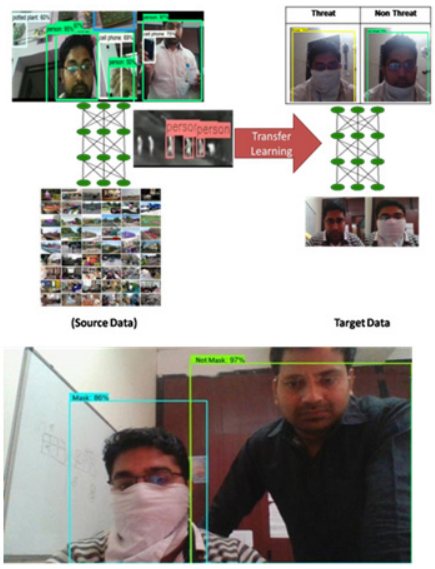

Figure 1: The flow of experimental diagram of implemented transfer learning.

Along with the use of the transfer learning technique, we have used the data augmentation technique which helps our model to overcome the use of a large dataset as a basic need of the deep learning model required. We have used fewer data and this technique fulfils our requirement in converting it to a large dataset with different viewpoint such as Rotating images, Flipping, padding, cropping and transformation with change in colours like brightness, saturation, hue and contrast for training our deep learning model. For data augmentation, we have changed the code of our deep learning model and incorporated the data augmentation codes.

RESULTS

In the figure 2, graph shows the relationship between loss and number of steps. The hyperparameter which controls the values of weights in deep CNN are adjusted in our network concerning loss gradient are known as learning rate in terms of localization loss discussed in figure 2. In table 2, it shows the accuracy of the model varies for different learning rate and value of learning rate determines the travel rate along the slope of the function i.e. if the value is low we move slowly along the slop and high value of learning rate results in a faster movement along the slop. So, while deciding the learning rate value we need to be careful so that we do not miss any local minima. To ensure the coverage of local minima the low learning rate is preferable but at the same time, it would take a longer time to converge. So to keep both aspects in mind, we start training with a relatively large learning rate. The reason behind selecting large learning is that the initial random weight assigned to the network is far away from the actual optimal value. During the Figure 3 discuss the comparison results between two object detection model single shot multi-box detector and Faster Regional Convolutional Neural Network, which shows the model is trained properly as compared to loss functions.

Table 2: Comparison table on different learning rate

Component	lr=0.004		
Pre-training	✓	✓	
Single-shot multi-box detector	✓	✓	✓
Faster RCNN	✓	✓	
mAP%	77.6	83	87

The training loss is done with the help of the loss function which a method to describe how a particular algorithm earns using a loss function (9) (10). It's a method of evaluating how well specific algorithm models the training data. In figure 4, we have observed that the initial loss value is quite high and it gradually decreases and approaches to 1 after a certain step. It means that a decrease in loss leads to an increase and accuracy in classification.

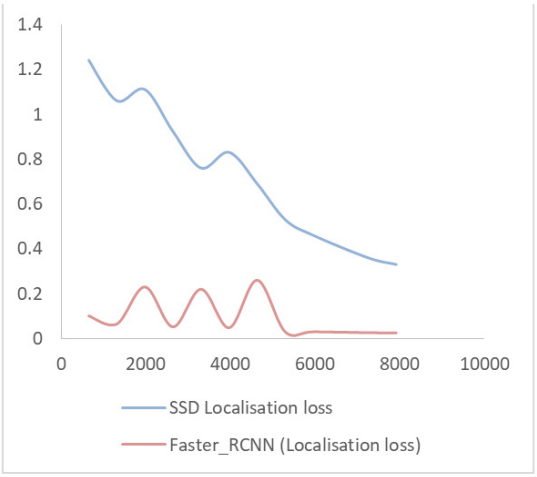

Figure 2: Localization loss for training the model.

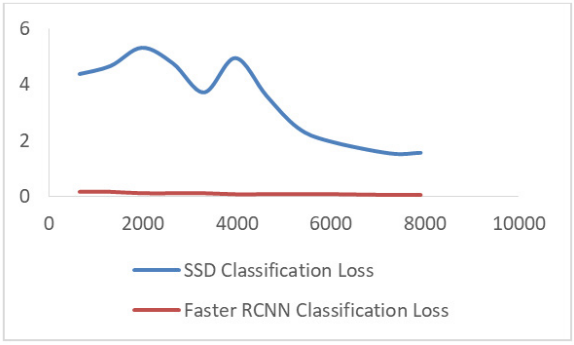

Figure 3: Classification loss for training the model.

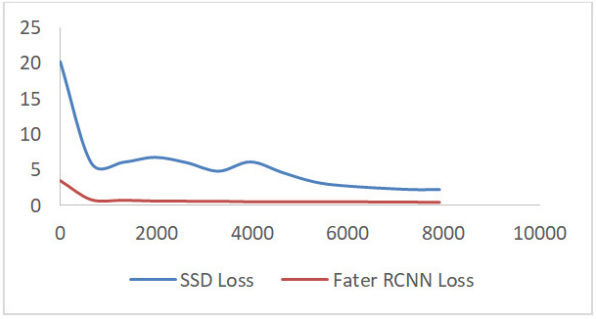

Figure 4: Total loss for training the model.

Initially, the predicted value deviated too much from the actual value that is why the initial loss value is too high. We used the cross entropy loss function which gives a better convergence rate as compared to other loss function hinge. In classification, it is tried to predict output from a set of finite categorical values that are given large data set of images of

mask and non-mask, categorizing them into mask class and non-mask class.

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{conf}(x,l,g) \quad (9)$$

$$L_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} log(c_i^p) - \sum_{i\in Neg} log(c_i^0) \; where c_i^p = \frac{exp(c_i^p)}{\sum_p (c_i^p)} \quad (10)$$

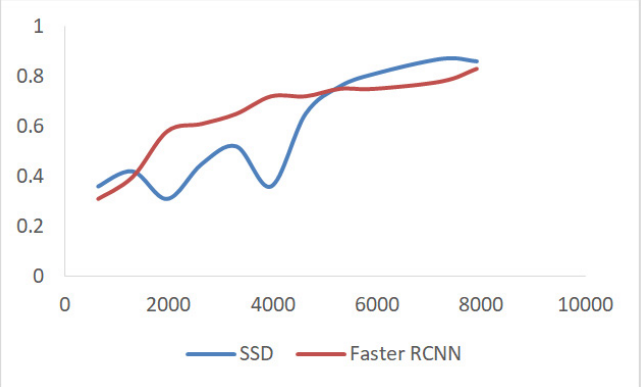

Figure 5: Comparative accuracy between both the models.

DISCUSSION

The implemented transfer learning technique and experimental results in figure 5, shows the success of the unmanned threat warning system and detection of the mask to avoid covid-19 with accuracy mAP 87% on testing validation using own tuned single shot multi-box detector model on top of the pre-trained model. Whereas Faster RCNN is taking too much time in training and reaching the accuracy and takes a lot of time while detecting the mask. Overall, the proposed single shot multi-box detector system shows improved accuracy, learning rate, classification loss at the time of training of a model and takes less time in training than that of training the model from scratch and compared to the Faster RCNN model which takes more time and makes it computationally efficient. Therefore, any type of threat can be annotated by the model and can perform automatic mask detection and performs better with IOU=.75 on our dataset, which shows the successful implementation of transfer learning techniques and optimization/tuning of hyperparameters on the single-shot multi-box detector_mobilenet_v1 model. Our adopted training strategies lead to improved performance in choosing appropriate bounding box, sampling of various location, scaling and aspect ratio than that of existing methods.

The future work will be more focused on improvising the model with different techniques such as loss function and fine-tuning the optimizer on the different pre-trained model for enhancing the performance along with IoT devices and sensors.

ACKNOWLEDGEMENT

I would sincerely thank my Guide, Supervisor and research mates for their motivation and guidance.

Conflict of interest: There is no conflict of interest

Source of Funding: Nil

REFERENCES

1. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states. Health Affairs. 2020; 10–1377.
2. Nguyen CT, Saputra YM, Van Huynh N, Nguyen NT, Khoa TV, Tuan BM, et al. Enabling and emerging technologies for social distancing: A comprehensive survey. 2020; arXiv preprint:2005.02816.
3. Agarwal S, Punn NS, Sonbhadra SK, Nagabhushan P, Pandian K, Saxena P. Unleashing the power of disruptive and emerging technologies amid covid 2019: A detailed review. 2020; arXiv preprint arXiv:2005.11507.
4. Punn NS, Sonbhadra SK, Agarwal S. Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deep sort techniques. 2020;arXiv preprint arXiv:2005.01385.
5. Cristani M, Del Bue A, Murino V, Setti F, Vinciarelli A. The visual social distancing problem. 2020;arXiv preprint:2005.04813.
6. Chan AB, Liang ZSJ, Vasconcelos N. Privacy-preserving crowd monitoring: Counting people without people models or tracking. IEEE Conference on Computer Vision and Pattern Recognition. 2008;1–7.
7. Qureshi FZ. Object-video streams for preserving privacy in video surveillance. Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. 2009;442–447.
8. ParkS, Trivedi M. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. Conference on Advanced Video and Signal Based Surveillance. 2005:71–176.
9. Ren S, He K, Girshick R, Sun J. Faster r-CNN: Towards real-time object detection with region proposal networks. Adv Neural Information Proce Syst. 2015:91–99.
10. Zou Z, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. 2019;arXiv preprint:1905.05055.
11. Singh A, Jotheeswaran J. Cognitive science-based inclusive border management system. MIC, Muscat. 2018;1-5.
12. Singh A, Jotheeswaran J. P300 Brain Waves Instigated Semi-Supervised Video Surveillance for Inclusive Security Systems. In: Ren J. (eds) Advances in Brain Inspired Cognitive Systems. BICS 2018; Lecture Notes in Computer Science, vol 10989. Springer, Cham.
13. Borji A, Sihtie DN. Quantitative analysis of human-model agreement in visual saliency modelling: A comparative study IEEE T-IP 2013;22:55-69.
14. Borji A. Human vs computer in scene and object recognition in IEEE CVPR;2014.
15. Zhao ZQ, Zheng P, Xu ST, Wu X. Object detection with deep learning: A review. IEEE Transact Neural Netw Learn Syst 2019;30(11):3212– 3232.
16. Sreenu G, Durai S. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data. 2019;6(1).

17. Kardas K, Cicekli NK. SVAS: Surveillance Video Analysis System. Expert Syst Appl. 2017;89:343–361.
18. Jackson D, Samuel R, Fenil E, Manogaran G, Vivekananda GN, Thanjaivadivel T, et al. Real-time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM. Computer Netw. 2019;151:191–200.
19. Bouachir W, Gouiaa R, Li B, Noumeir R. Intelligent video surveillance for real-time detection of suicide attempts. Pattern Recogn Lett. 2018; 110:1–7.
20. Ribeiro M, Lazzaretti AE, Lopes HS. A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recogn Lett. 2018;105:13–22.
21. Babaee M, Dinh DT, Rigoll G. A deep convolutional neural network for video sequence background subtraction. Pattern Recogn. 2018; 76:635–49.
22. Cue H, Liu Y, Cai D, He X. Tracking people in RGBD videos using deep learning and motion clues. Neurocomputing. 2016;204:70–6.
23. Zhao Z, Zheng P, Xu S, Wu X. Object Detection With Deep Learning: A Review. IEEE Transact Neural Netw Learn Syst. 2018;30(11):3212-3232.
24. Huang R, Pedoeem J, Chen C. YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers. IEEE International Conference on Big Data. 2018; doi:10.1109/bigdata.2018.8621865
25. Liu W. SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV;2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham
26. Zeng M, Li M, Fei Z, Yu Y, Pan Y, Wang J. Automatic ICD-9 coding via deep transfer learning. Neurocomputing.2018;324:43-50.
27. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. J Big Data 2019;6:60.