Targeted Face Recognition and Alarm Generation for Security Surveillance using Single Shot Multibox Detector (SSD)

K. M. Tawsik Jawad
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Maisha Binte Rashid
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Nazmus Sakib
Assistant Professor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ABSTRACT
Face recognition has been considered as one of the most important means of security in prevention of crimes in this era. Surveillance cameras in crowded areas keeps good monitoring of all activities. So, it can be used as a witness against criminals or can be used to prevent crimes before happening. With the advancement in deep neural networks in surveillance cameras, face recognition accuracy has increased in challenging environments too. But this country is still lagging behind in this regard. So, the proposed work focuses mainly on face recognition with custom Bangladeshi dataset that can be robust enough against blurriness, pose variations and occlusions. Single Shot Multibox Detector (SSD) model was chosen since it produced significant improvement in accuracy compared to many state of the art models. Tensorflow API was used with SSD-Mobilenet-FPN model config to generate alarms when targeted face was recognized among many faces in crowd.

Keywords
Bangladeshi Face Dataset, Targeted Face Recognition, Single Shot Multibox Detector, Tensorflow API, Alarm.

1. INTRODUCTION
Face Recognition means to localize face regions in an image and extract necessary features from those faces to match with faces available in a database for classification [1]. It has various applications in prevention of falsifying data, attendance management and most importantly in surveillance systems. Proper facial recognition system in surveillance systems can help prevent crimes as face variation is the strongest bio-metric identification tool.

Face recognition as a security measure has been heavily researched upon. Alert generation on detection and recognition of wanted individuals has been done through various techniques. Use of machine learning algorithms like Viola Jones, Principal Component Analysis (PCA) for face recognition along with GSM modules for sending message to authorities has been done in [2].

Crime rate in Bangladesh is on the rising side in recent years. With the advancement of modern technology, criminals are finding new means of committing crimes. As a counter-measure, law enforcement authorities also have to be smart enough to prevent a crime before happening. In Fig. 1, a study shows that crime rate in UK decreased in a great number after they started using surveillance cameras [3]. So, a proper surveillance system can go a long way as a solution to the rapid increase in crime rate in the country. As there will be records of every movement in public areas like airports, shopping malls or concerts, criminals will think twice before committing a crime. Surveillance cameras can work as a strong tool for detection of criminals in busy or remote areas. They can detect activities which might go unnoticed by human eyes [4]. Many occurrences of street crimes like theft, burglary, harassment etc. happen in night time where less people are available for help. The motive of the proposed system is to recognize criminals in these areas and alert the law enforcement authorities as early as possible. In 2013 Boston Marathon a massive bombing incident took place where criminals fled the scene. The law enforcement agencies ultimately were able to find out the culprits from the video footage of surveillance cameras in a departmental store after 3 days [5]. These drawbacks in detection acted as a strong motivator to create a robust system of face recognition to identify criminals.

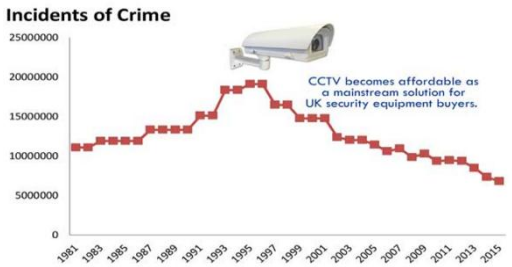

Fig. 1: Decreasing crime rate in UK after using surveillance cameras [3].

Implementation of such a robust system is a computationally expensive process. That is why selection of appropriate model, dataset and fine-tuning model parameters for the dataset was crucial to the cause. With the strong success of deep learning neural networks in object recognition over previous approaches, it was logical to select deep learning approach on proposed system [6]. Immense research in object recognition has been done on deep learning; specially Convolutional Neural Networks (CNN), Region CNN (R-CNN), Fast R-CNN, Faster R-CNN, You Only Look Once (YOLO), Mask R-CNN etc. But, object recognition with Single Shot Multibox Detector (SSD) has been proven to perform with more accuracy compared to YOLO and Faster R-CNN on many datasets [7]. The main approach was to create a custom Bangladeshi dataset with annotation, train the dataset fine tuning necessary SSD parameters in the config file and finally use the classifier in the inference graph to recognize faces of individuals in different scenes.

Face recognition itself is a complex procedure with many factors to be taken into consideration. When challenges like pose variation, blurriness, occlusion, lighting conditions etc. are added then it becomes even more difficult to detect and recognize. That is why the approach in proposed system is divided into two parts. First, to prepare a robust system that can recognize all the possible faces in a crowd. Next, by putting input, the system to recognize any particular face (targeted face) and discard others. When the system can do this, it will simultaneously generate an alarm to alert the authorities.

2. RELATED WORKS

Several pre-processing techniques have been applied for better dataset generation for face recognition. In **[8]** they have focused towards making strong enough dataset for face recognition by image enhancement.

Significant works regarding blurriness, occlusions, pose variations etc. that stand against face detection have been done by Jeremiah R. Barr, Kevin W. Bowyer and Patrick J. Flynn in **[5]** using Multi Pose Algorithm. This algorithm uses left, right, frontal face scores to pass to cascade object detector. The score normalization gets the number of stages the image segment passes and the final score of the base classifier. If the combined score exceeds a threshold then a decision can be inferred as the region being a face [9].

Blurriness was not properly dealt with but it was measured by edge density values. True positive rate was 14.6% and false negative rate was 62.6% in blurry image segments. Both blurriness and occlusion led to large misclassification rates in FDDB dataset [10].

Pose variation is a significant factor to consider for recognizing faces in crowded areas. In order to make a system robust enough against pose variation, major ideas were presented by Changxing Ding, Chang Xu and Dacheng Tao in their paper [11]. Pose variation in different images tend to have 3 general conditions: First, normalized images from different poses will naturally have variation in image quality. So, transformation will get difficult. Second, since pose variations are the data of same subjects, there will be similarity in features and lastly, collection of large dataset with pose variations is difficult and face recognition algorithms usually require significantly large data in order to perform with reasonable accuracy [12].

Identification of large number of faces in a scene has major challenges to deal with. Especially in crowded scenes, where faces are much distant from camera and appear small with so many pose variations. Convolutional Neural Network (CNN) has recent developments in this regard for face recognition [13]. Key frame extraction approach by CNN was quite remarkable work in face recognition from videos [14]. While CNN has vast contributions in object recognition, there were some crucial drawbacks which were visualized in detail by Eric Kauderer-Abrams in [15]. Translation invariance is not up to the mark for CNN classification. They have inferred that data augmentation removes this drawback to a great extent but can't totally overcome it.

Localization is one of the primary stages of object recognition. In [16], they have developed the Selective Search Algorithm into a robust, stable and independent of object-class algorithm adapting segmentation. The Selective Search Algorithm was data driven and combined hierarchical grouping strategies to generate less number of regions for localization [17].

Region proposals by selective search greedy algorithm was combined with CNN for object detection and segmentation improvements. Region based CNN (R-CNN) approach improved over previous objection detection accuracy by a significant margin on VOC 2012 dataset. Category independent region proposal along with convolutional filters that extract feature vectors from each region has been known to be a strong object detector approach than image classifier [18].

The use of Selective Search for Object Recognition didn't prove much effective for object recognition. After comparing it to efficient detection networks it was an order of magnitude slower at 2 seconds per image. In [19], they have introduced a Region Proposal Network (RPN) that eradicates computationally expensive R-CNN for object recognition. With the sharing of deep convolution layers, high quality region proposals were generated and real time frame processing was possible with results evaluated on PASCAL VOC 07 test set, Minival 07+12 train set using model VGG-16 [20]. The running time to obtain results using Faster R-CNN with RPN was 198ms per image.

While Faster R-CNN was able to produce real time results in object recognition, there were rooms for improvement in Frame Processing Time (FPS). Single Shot Detector (SSD) was able to achieve 59 FPS on VOC 2007 test with 74.3% mean Average Precision (mAP) where mAP for Faster R-CNN was 73.2% with 7 FPS. SSD outperforms YOLO by a large margin where mAP for YOLO was 63.4% with 45 FPS [7]. This fast processing of frames comes from elimination of bounding box proposals and subsequent pixel or feature resampling stage. Improvements added by them were basically applying small convolution filter for predicting object class, use of separate filters for separate aspect ratio detections and applying these filters to different feature maps for detection in different scales.

2.1 Single Shot Multibox Detector

SSD Model implemented by [7] has extra feature layers which use a set of convolutional filters to produce a set of detection predictions their base network. The offsets relative to the shapes of default boxes at each cell are determined at each feature map cell [7]

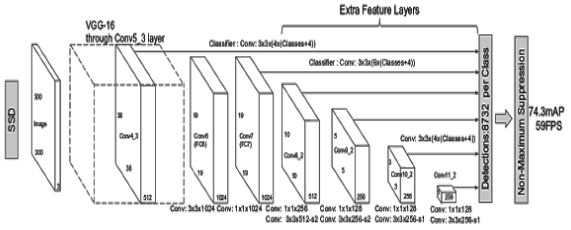

Fig. 2: Application of default boxes to several feature maps of different resolutions to represent possible locations of output box shapes [7].

The base network was VGG-16 at the end of which extra feature layers were added. Default bounding boxes were associated with each feature map cell for multiple feature maps. Faster R-CNN's anchor boxes were similar to these default bounding boxes but application of bounding boxes was done to several feature maps in different resolutions. In this way, possible locations of output box shapes were represented discretely.

3. PROPOSED SYSTEM AND METHODOLOGY

The main focus is to make a system that can be useful for law enforcement agencies by building a face recognition system using surveillance camera. The target is to recognize any criminal face from crowded place using video surveillance camera. But in crowded places, the faces are not directly facing to the camera. So, the main motive is to build a system that can recognize faces handling the factors like off frontal faces, obstacles in front of faces, pose variations, different expressions etc. When police want to find any criminal, they have to give input the ID of the criminal that is already saved in the database. When the system can find the targeted person in the crowd it will generate an alarm to alert police the minute it finds the person's face in the crowd.

After analysis of recent developments of deep learning models with their significant performances on popular datasets and major drawbacks, it was decided to select Single Shot Detector (SSD) as the model to train and evaluate on the Bangladeshi Dataset. The whole process is given in the flowchart given in Fig. 3.

The procedure will follow by processing frames one by one to match faces in the input frame to already labelled faces in the dataset. If the confidence score of recognition is greater than 0.6 then the proposed system will output name of the person along with the confidence score in the output bounding box over his face.

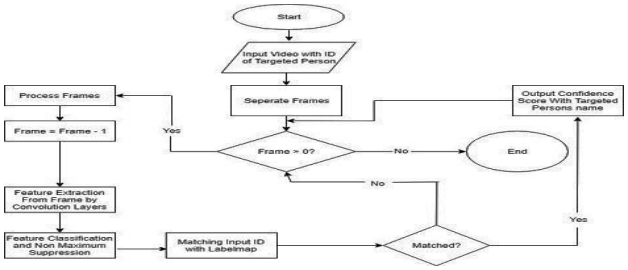

Fig. 3: Flowchart of face recognition by SSD

3.1 Multiple Face Detection

The first step to recognize a face is to detect the face region. Once the system detects the face region, then it matches the face with the faces saved in the dataset. As there are faces labeled in the dataset, after training all the images with SSD the system can detect multiple face regions from videos.

3.2 Multiple Face Recognition

Major focus was on the face recognition accuracy with real-time video processing where SSD has strong enough reputation. It was important to tune SSD parameters for proper training of the dataset. Learning Rate and Batch Size must be changed from the pre-trained config file of selected model for this process. Robust enough dataset preparation with different lighting conditions in different scenes was required for the evaluation of the model.

3.3 Targeted Face Recognition

After recognizing all the possible faces from a scene, proposed system must recognize the individuals that user wants it to recognize. When the system can successfully recognize those particular faces, it will instantly generate an alarm to alert authorities. In this way crime detection will become easier from areas with surveillance cameras with SSD framework for face recognition.

4. DATASET GENERATION

As this work proposes to recognize faces of different people in crowded places large dataset having multiple faces from different angle of different people is necessary. While making this dataset some important factors had to be considered to make a proper face dataset. As it needs to recognize face in the crowd there can be so much variation in background, lighting conditions etc. While recognizing face in crowd key factors to be considered are different expressions, pose variations, occlusion etc.

The steps of making dataset is shown in Fig. 4.

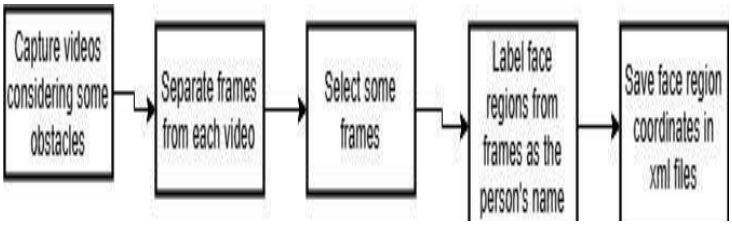

Fig. 4: Steps of making dataset

- Videos were made with different background as in crowded places there can be so many different backgrounds.
- Videos were made in different lighting conditions like – some videos in daylight, some videos where there was insufficient light, some videos in room light were made.
- The videos were made where faces are not facing directly at the camera. The faces are in different angle, some obstacles near the faces, faces with different expressions.
- The videos were prepared with people of different skin color.

Some sample faces of different persons from different angle from the dataset is shown in Fig. 5.

Fig. 5: Sample faces from the dataset

4.1 Dataset Specifics

This dataset has 40 videos of different persons. Multiple persons were present in each video. Each video duration was no greater than 40 seconds. It was done by separating each video into image frames.

Dataset consisted of 2500 images of 44 subjects. Equal number of images could not be collected for each subject but approximately there were 35-40 images per person in the dataset.

Each image size was less than 450KB and resolution was 1280
X 720. Videos were made in different background and different lighting conditions and faces with different angles. The dataset was divided into training set and testing set with the ratio 8:2.

Fig. 6: Labeled face regions

After gathering videos and separating frames, each face region was selected in an image with a box and labeled them as shown in Fig. 6. As this system will recognize faces from crowded scenes, so face regions were made that have different lighting and image quality.

5. IMPLEMENTATION

This section provides reasons for selection of SSD model, followed by training illustration and finally evaluation on the trained dataset.

5.1 Model Selection

Model selected for the system was ssd mobilenet v1 fpn coco among the pre-trained models on COCO dataset found in Tensorflow object detection model zoo. Reason for selecting this model was its great accuracy and train-test time which was much better compared to other Faster R-CNN, Mask R-CNN and SSD models. Running time on COCO dataset was considered to be suitable for the system. Another reason for selecting this model was this model can be trained using Tensor Processing Unit (TPU) which might use to implement for faster training time in future.

5.2 Training with Parameter Tuning

Model was trained using Tensorflow Object Detection API in a Windows Machine having NVIDIA GeForce GTX 1050 GPU. Total time allowed for training was around 9 hours. Two parameters were tuned for training. Batch size 64 proved to be too large to train the dataset that's why it was reduced to 32 which is considered to be one of the standard batch sizes to train a neural network for custom object recognition. Learning rate base was 0.04 which then allowed for couple of hours for observation of loss graphs. It was found out this base learning rate was quite slow and converged on a local minima. Later on, base learning rate as shown in Fig. 8, was changed to 0.06 and again to 0.12 to observe the performance of the model.

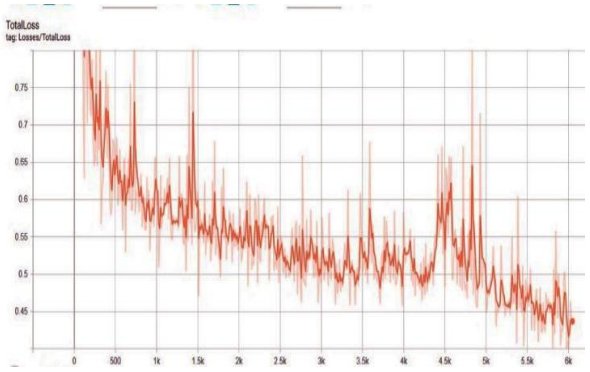

Fig. 7: Total loss graph after 6k steps. Loss being consistent around 0.45 for a long time where training was stopped.

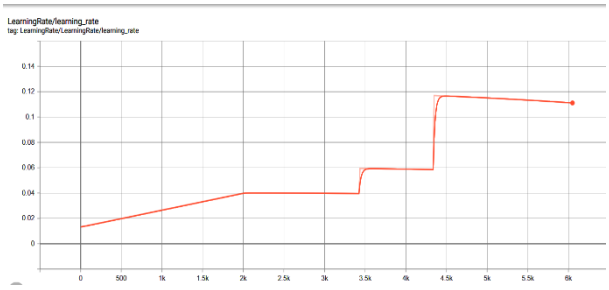

Fig. 8: Learning rate graph transition to base.

In Fig. 7, it shows the total loss graph had some big spikes in the beginning when learning rate was rising from 0.04 to 0.06 and from 0.06 to 0.12. But as learning rate reached close to 0.12, the loss graph spikes became more consistent around 0.40 to 0.45. This process continued on for quite some time and finally the training was stopped when total loss value was consistently below 0.45 and the total loss values did not degrade any further for a considerable amount of time.

5.3 Model Evaluation

Completion of training was followed by exporting inference graph that contained the classifier for face recognition. Input videos were captured in different environments to find out how well the classifier performs. Input videos were captured in resolution 1280 * 720. The flowchart in Fig. 3 works in the following ways for targeted face recognition:

- Video and the ID of the person to be identified are provided as input to the system.
- Proposed system extracts frames one by one and those are input to the convolution layers for feature extraction.
- Feature Classifier contained in the Inference Graph classifies dominant features and matches results with features in the trained dataset.
- If minimum score of faces in the input is greater than 0.6 then bounding boxes with detection score and matched ID with Label map are shown over the recognized face.
- When targeted face is recognized with confidence score greater than 0.6 then an alarm is generated by the system.
- If face is not recognized in one frame, process moves on to the next and finishes when frame count reaches 0.

Each frame after being processed one by one is saved and an output video in real time is generated with the recognition results. Frame processing works the same way from any smart-phone videos with different qualities.

6. EXPERIMENTAL RESULT

Main purpose is to recognize targeted faces from crowded places. So first it was checked to see how many faces the system can recognize. All the faces were not facing directly to the camera but when the system compares any face with the faces in the dataset and finds match then it shows result with a bounding box and confidence score.

6.1 Result Illustration

Input videos were taken with subjects engaged in conversation or day to day activities to represent action of a surveillance camera in real world scenarios. Daily activities faced the usual obstacles in face recognition so the accuracy of the model in proposed system dataset can be visualized properly.

Fig. 9: Recognized faces in daylight

In Fig. 9, it shows that the system can recognize faces in daylight and shows the recognized person's name with a bounding box.

Fig. 10: Recognized faces in insufficient light

In Fig. 10, it shows that the system can recognize faces in low light where the subjects are immersed in conversations. This system can recognize faces in hazy or blurry images. While making the dataset, some videos were made where frames were a bit hazy or blurry so that the system can recognize faces in these difficult situations.

Fig. 11: Recognized faces in blurry image

In Fig. 11, the image is a bit blurry but the system can recognize some faces correctly. Sometimes this system recognizes faces incorrectly. In Fig. 12, it can be seen that the system recognized two faces where it recognized one correctly and the other incorrectly.

Fig. 12: Incorrectly recognized face

After testing that this system can recognize multiple faces from video frames, approach was made for next proposal and that is searching and recognizing targeted faces from a video. When the system recognizes that particular face or faces it generates an alarm.

In Fig. 9, it shows that the system recognized multiple faces but in Fig. 13 which is the same image used in Fig. 9, it only recognized one person.

Fig. 13: Finding and recognizing a particular face among many faces

In Fig. 13, it can be seen that there are many faces that are saved in the dataset but as the system needed to search for a particular person so when the ID of the targeted person is provided as input, the system only recognizes that person's face among all the faces in the scene.

6.2 Analysis

After analysis on a number of input videos in different conditions performance metrics were measured in terms of average performance in three different scenarios: daylight conditions, room light and low-light conditions; which are shown in the tables below :

Table. 1: Rates of Performance Metrics of the SSD model in daylight.

Performance Rates in daylight	
True Positive	41%
True Negative	29%
False Positive	12%
False Negative	18%

Table. 2: Rates of Performance Metrics of the SSD model in room light.

Performance Rates in room light	
True Positive	28%
True Negative	17%
False Positive	24%
False Negative	31%

Table. 3: Rates of Performance Metrics of the SSD model in low light.

Performance Rates in lowlight	
True Positive	16%
True Negative	14%
False Positive	28%
False Negative	42%

Due to small dataset, recognition results were not up to the mark. It was found that model had 70% accuracy on average in daylight conditions where it dropped to 45% on room light and 30% in low light conditions. Frame processing time was slow with only 5 frames per second on the input videos. So, real-time face recognition could not be done with desired accuracy.

7. LIMITATIONS AND FUTURE WORK

Proposed system has some limitations. That is why desired goal was not achieved as it was planned before.

- System needed more data for getting more accurate result. For this reason, the accuracy level of this system is not up to the mark.
- Due to high resolution in input videos, frame processing speed is low.
- Because of not having proper available resources mobile camera was used for recording the videos. If surveillance cameras were used, the system would have been more suitable for real world scenarios.

Future goal is to make this system accurate enough to help the law enforcement unit of this country for preventing the crime and detecting the criminals as early as possible. If this system can be established in this country, then the criminals will be afraid to commit any kind of crime. So in future, a large dataset for making this system more effective is needed. Without a sound amount of dataset, the desired result cannot be gained. Huge facial variation is one the main reasons for not up to the mark performance.

8. CONCLUSION

Crime rate around the country in recent years is a huge concern for the citizens. A proper surveillance system requires a strong facial recognition system for all possible scenarios. The limitations faced by this system can be overcome with more dataset collection and surveillance camera modifications using this system. Implementation of proposed system in surveillance cameras after overcoming the limitations can be a pretty useful tool in reducing crime rate and making the citizens feel safe in their home country.

9. REFERENCES

[1] W. a. C. R. a. P. P. J. a. R. A. Zhao, "Face recognition: A literature survey," ACM computing surveys (CSUR)," 2013.

[2] A. J. P. J. A. Aswathy Wilson, "Security Alert Using Face Recognition," 2017.

[3] "Office for National Statistics," [Online]. Available: https://www.ons.gov.uk.

[4] D. Williams, "Effective CCTV and the challenge of constructing legitimate suspicion using remote visual images," 2007.

[5] J. R. a. B. K. W. a. F. P. J. Barr, "The effectiveness of face detection algorithms in unconstrained crowd scenes," 2014.

[6] Q. D. Xiao Han, "Research on Face Recognition Based on Deep," 2018.

[7] W. a. A. D. a. E. D. a. S. C. a. R. S. a. F. C.-Y. a. B. A. C. Liu, "Ssd: Single shot multibox detector," 2016.

[8] A. H. Ade Nurhopipah, "Motion Detection and Face Recognition," 2018.

[9] Y.-Y. a. L. T.-L. a. F. C.-S. Lin, "Fast object detection with occlusions," 2004.

[10] V. a. L.-M. E. Jain, "Fddb: A benchmark for face detection in unconstrained settings," 2010.

[11] C. a. X. C. a. T. D. Ding, "Multi-task pose-invariant face recognition," in IEEE Transactions on Image Processing, 2015.

[12] a. R. A. P. Alexander A. S. Gunawan, "Face Recognition Performance in Facing Pose," 2017.

[13] Y.-Y. O. L.-Y.-C. H. J.-F. W. An-Chao Tsai, "Efficient and Effective Multi-person and Multiangle Face Recognition based on Deep CNN," 2018.

[14] C. L. a. S. S. Xuan Qi, "Boosting Face in Video Recognition via CNN based Key Frame Extraction," 2018.

[15] E. Kauderer-Abrams, "Quantifying translation-invariance in convolutional neural networks," 2017.

[16] J. R. a. V. D. S. K. E. a. G. T. a. S. A. W. Uijlings, "Selective search for object recognition," in International journal of computer vision, 2013.

[17] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," 2013.

[18] J. D. T. D. a. J. M. Ross Girshick, "Region-Based Convolutional Networks for," 2016.

[19] S. a. H. K. a. G. R. a. S. J. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015.

[20] M. a. V. G. L. a. W. C. K. a. W. J. a. Z. A. Everingham, "The PASCAL visual object classes challenge 2007 (VOC2007) results," 2007.