Walchand College Of Engineering, Sangli.

(An Autonomous Institute)

Department Of Computer Science and Engineering

A Project Report On

Threat Detection with Facial Expression

Submitted by

Pranav Sunil Raut (2016BTECS00077)

Sourabh Shashikant Pukale (2016BTECS00076)

Dhanashree Shekhar Phulkar (2016BTECS00019)

Under the Guidance Of

Mr. K. P. Kamble	Dr. B.F. Momin
Project Guide	HoD, Dept CSE

Walchand College of Engineering, Sangli

(An Autonomous Institute)

Department Of Computer Science and Engineering

CERTIFICATE

This is to certify that the Project Report entitled, "**Threat Detection with Facial Expression**" submitted by

Mr. Pranav Raut, Mr. Sourabh Pukale, Miss. Dhanashree Phulkar

to Walchand College of Engineering, Sangli, India, is a record of Bonafide Project work carried out by him/her under my/our supervision and guidance and is worthy of consideration for the award of the degree of Bachelor of Technology in Computer Science & Engineering of the Institute.

Mr. K. P. Kamble	Dr. B.F. Momin
Guide	Head Of Department
Computer Sci. & Engg. Dept,	Computer Sci. & Engg. Dept,
WCE, Sangli.	WCE, Sangli

Acknowledgment

We would like to thank our project guide Mr. K. P. Kamble, H.O.D. Dr. B.F. Momin for their guidance throughout the Project period. We would like to thank our institute for helping us through the period. We would like to thank all people who helped us in pursuing the project successfully.

Declaration

We hereby declare that work presented in this project report titled "**Threat Detection with Facial Expression**" submitted by us in the partial fulfilment of the requirement of the award of the Mini-project Submitted in the Department of Computer Science & Engineering, Walchand College of Engineering, Sangli, is an authentic record of my project work carried out under the guidance of (Mr. K. P. Kamble).

Date: 18/04/2019

Place: W.C.E., Sangli

Pranav Raut (2016BTECS00077)

Sourabh Shashikant Pukale (2016BTECS00076)

Dhanashree Shekhar Phulkar (2016BTECS00019)

THREAT DETECTION WITH FACIAL EXPRESSION

Pranav Raut, Sourabh Pukale, Dhanashree Phulkar, Kiran Kamble

Department of Computer Science and Engineering,
Walchand College of Engineering, Sangli, Maharashtra, India

pranavr7700@gmail.com,
souravp7777@gmail.com,
dphulkar@gmail.com,
kirankamble5065@gmail.com

Abstract

Recently criminal activities such as robbery with the threat of life using weapons has increased exponentially. CCTV cameras were used for providing proof of criminal activities happened in the past. But due to the technological advancements in areas like deep learning, image processing, etc. various ways have been put up to prevent those criminal activities. Weapon detection was the first criteria introduced to define an activity suspicious. But it had many drawbacks, such as dummy weapons were being classified as a threat and also the system failed where there was frequent use of weapons. As the phrase “Suspicious Activity” is a relative entity, a person can identify a friendly environment, where people might carry weapons but the existing system lacks of this context. We have tried to introduce the existing system of weapon detection to this context. This enhances the ability of the system to think as if it is a human brain. A CNN model with accuracy 56% was used for facial expression detection, and Faster RCNN and SSD models were used for weapon detection with accuracy 82.48% and 82.84% respectively.

Keywords: Object Detection, Faster RCNN, SSD, Suspicious Activity, Human facial expression detection.

Introduction

From many years people are getting robbed, banks are getting robbed by threatening them with weapons. Many people have lost their life just to create an alert at that point of time. With the increase in development of technologies, every problem keeps on demanding a better solution after a specific interval. CCTV’s are being commonly used in our society, the purpose of using it was to have proof if something goes wrong. But a CCTV along with a brain can now prevent many such incidents. To take actions immediately against a suspicious activity, we first need to be sure that the activity is actually suspicious. Children playing with toy weapons is not a suspicious activity as we can see their facial expressions, they are happy. This analysis of the environment done by our human brain by taking into consideration many possible answers and choosing the appropriate is what we want to embed into our systems. We are not far from the day when drones can be used as safety guards floating around the city and taking appropriate actions if things are seemed to be fishy. It is not possible to have a human as the security guard at every possible place, but machines can be used for serving this purpose if they are as smart enough as humans. A human carrying a weapon is normal in some states/countries, here a complex system is needed to serve the purpose of detecting suspicious activity, because claiming the activity to be suspicious right after detecting would be a dumb way in today's world. We have many systems that work as independent solutions to simple problems. Here we have tried to combine the solutions of simple problems to solve this complex problem.

Literature Study

Object retrieval methods usually include two steps, i.e., searching for images containing the query object and locating the object in the image with a bounding box. The former step is essential to the image classification, and many high performance results are achieved by using convolutional neural network by Hossein Azizpour [2]. However, R-CNN consumes substantial time to process every object proposal without sharing computation.

Convolutional Neural Networks achieve better accuracy with big data. However, there are no publicly available datasets with sufficient data for facial expression recognition with deep architectures. Therefore, to tackle the problem, Andre [7] applied some pre-processing techniques to extract only expression specific features from a face image and explore the presentation order of the samples during training.

Fast R-CNN shares features among object proposals. A Region of Interest (RoI) pooling layer is designed to obtain faster detection speed by Lina Xuna [6]. As a result, features are extracted only once per image by Hailiang Li [3]. While accurate, these approaches have been too computationally intensive for embedded systems and, even with high-end hardware, too slow for real-time applications.

Wei Liu [1] paper presents the first deep network based object detector that does not resample pixels or features for bounding box hypotheses and is as accurate approaches that do. This results in a significant improvement in speed for high-accuracy detection (59 FPS with mAP 74.3% on VOC2007 test, vs. Faster R-CNN 7 FPS with mAP 73.2% or YOLO 45 FPS with mAP 63.4%). The fundamental improvement in speed comes from eliminating bounding box proposals and the subsequent pixel or feature resampling stage.

In our study we have compared both the models i.e. Faster RCNN and Single Shot Detection for object detection.

Proposed Approach

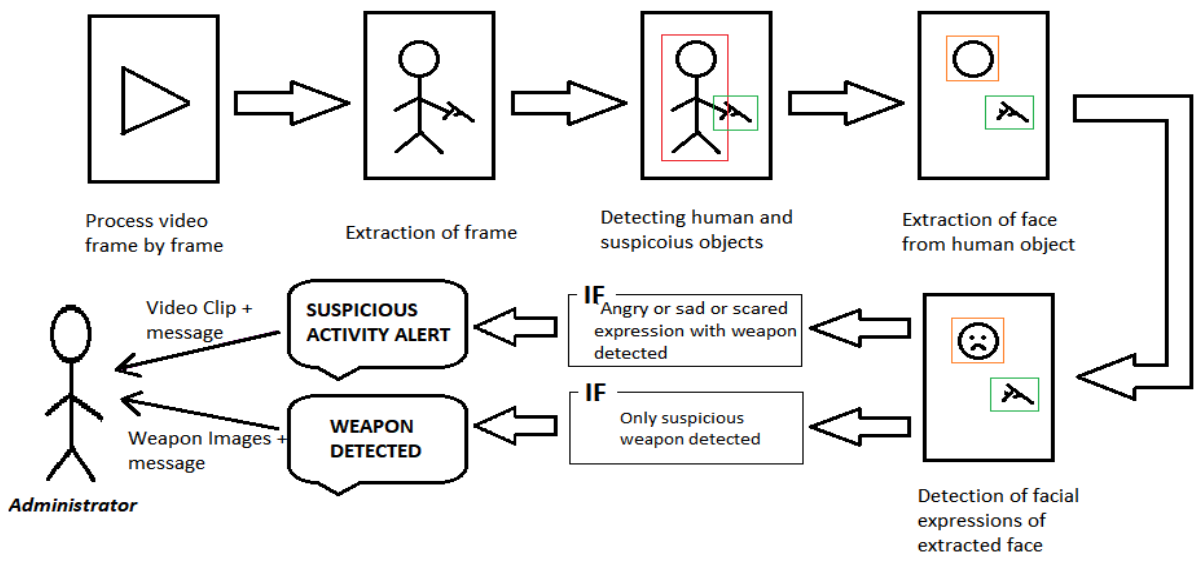

Figure 1: Flowchart of proposed approach

As per figure 1,

1) Video is processed frame by frame.
2) Detect human object in the frame
3) Obtain the face of human if present.
4) Obtain the facial expressions of the human.
5) Check if a weapon is being introduced in the frame with help of Faster RCNN or SSD model, whichever suitable for the respected deployable area.
6) If there is presence of weapon, and humans present in the frame are detected to be scared, angry or sad.
    i. Send a high alert to the respective security administrator, along with the cropped video where the suspicious activity was detected.
    ii. The type of alert, may depend on the probability of the activity being highly suspicious.

   If there is only presence of weapon, keep sending the images captured to the respective security administrator.

Methodology

1. Dataset Generation

The dataset consists of images for face recognition, facial expression detection and suspicious weapons detection. Dataset used for face recognition is Fer2013 which consists of 35.887 grayscale, 48x48 sized face images with 7 emotions. Emotion labels in the dataset:
0: 4593 images- *Angry*
1: 547 images- *Disgust*
2: 5121 images- *Fear*
3: 8989 images- *Happy*
4: 6077 images- *Sad*
5: 4002 images- *Surprise*
6: 6198 images- *Neutral*
The dataset for suspicious weapons were generated by collecting images of guns (majorly pistols) and knives and were labeled using LabelImg tool. 3187 images were used for training the weapon detection model and tested over 450 images.

Labelling Images with **LabelImg tool [9]**:

a. Open the image in LabelImg tool.
b. Draw the rectangle such that the object (weapon in our case) exactly fits in it.
c. Now label the rectangle in the class option by adding the class name (“gun” or “knife”) in our case.
d. Finally, after labelling all the objects present in that frame hit SAVE. This will generate an .xml file (with same name as that of image file). This will contain the information about bounding boxes in xml format. This is used further for training the model.

2. Custom training of the SSD architecture

After the generation of dataset, the model was trained on more than 3000 images depicting suspicious weapons both guns and knives. As per Figure 3, SSD’s architecture [3] builds on the venerable VGG-16 architecture, but discards the fully connected layers. The reason VGG-16 was used as the *base network* is because of its strong performance in high quality image classification tasks and its popularity for problems

where *transfer learning* helps in improving results. Instead of the original VGG fully connected layers, a set of *auxiliary* convolutional layers (from *conv6* onwards) were added, thus enabling to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer.

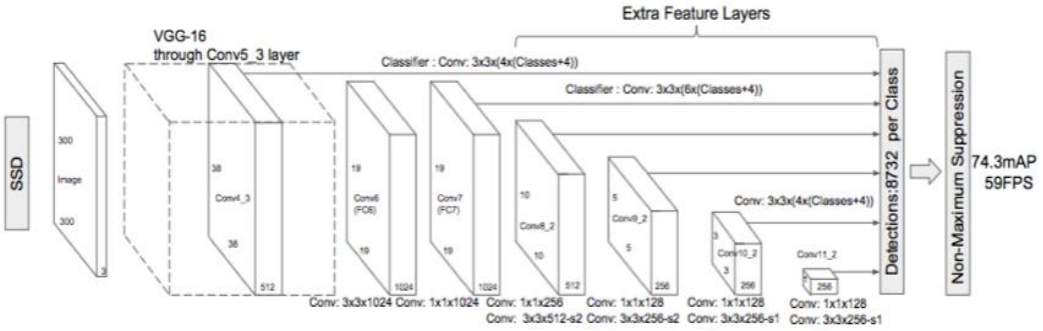

Figure 3: Architecture of SSD model

3. Custom training of the Faster RCNN architecture

As per Figure 4, Faster RCNN [2] is one of the powerful model used for object detection and uses convolution neural networks. It consists of 3 parts:

a. Convolution layers:
   In this layers, filters are trained to extract the appropriate features from the image. Convolution networks are generally composed of Convolution layers, pooling layers and a last component which is the fully connected or another extended thing that will be used for an appropriate task like classification or detection. It preserves the relationship between pixels by learning image features using small squares of input data.
   Pooling layers: Pooling consists of decreasing quantity of features in the features map by eliminating pixels with low values.

b. Region Proposel Network (RPN): RPN is small neural network sliding on the last feature map of the convolution layers and predict whether there is an object or not and also predict the bounding box of those objects.

c. Classes and Bounding Boxes prediction: Now we use another fully connected neural networks that takes as an input the regions proposed by the RPN and predict object class (classification) and Bounding boxes (Regression).

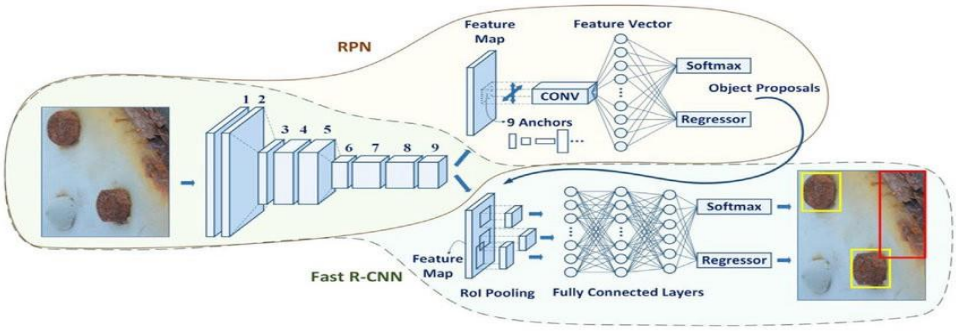

Figure 4: Architecture of FASTER RCNN model

Both SSD and FASTER RCNN models are trained on same dataset of suspicious weapons to compare results. Faster RCNN is also used for face detection and expression recognition.

4. Training and Testing the Custom Model:

The dataset generated consisted of 2500+ images of class – 'gun' and 800+ images of class 'knife. The model built was trained for 10,000+ iterations for SSD and 8000+ iterations for Faster RCNN model and a loss of about 1.5 and 0.01 was achieved for SSD and Faster RCNN respectively.

Results

Confusion Matrix - A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.
It allows easy identification of confusion between classes e.g. one class is commonly mislabelled as the other. Most performance measures are computed from the confusion matrix.

Table depicts Confusion matrix for suspicious weapon detection.

		Predicted		
		Gun	**Knife**	**No Weapon**
Actual	**Gun**	66	0	28
	Knife	4	74	11
	No Weapon	3	1	87

Table 1: Confusion matrix of SSD for suspicious weapons

		Predicted		
		Gun	**Knife**	**No Weapon**
Actual	**Gun**	91	0	3
	Knife	16	61	12
	No Weapon	17	0	74

Table 2: Confusion matrix of FASTER RCNN for suspicious weapons.

Figure 5: Results of Weapon detection

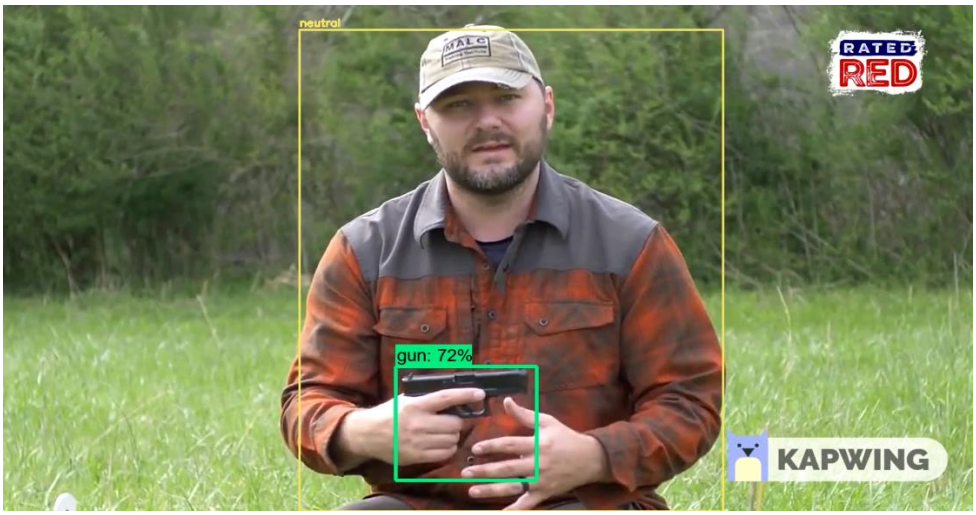

Figure 6: Results of Subject with GUN with neutral expressions

Figure 7: Results of Alerting on WEAPON Detected and Alerting on GUN POINT

The trained SSD model has obtained the accuracy of 82.84% with 86% precision and 83% recall and the trained FASTER RCNN model has obtained the accuracy of 82.48% with 85% precision and 82% recall.

Definition of the terms:

- Positive (P): Observation is positive (for example: is a gun).
- Negative (N): Observation is not positive (for example: is not a gun).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.

• True Negative (TN): Observation is negative, and is predicted to be negative.
• False Positive (FP): Observation is negative, but is predicted positive.

Precision: Precision is given by the ratio of total number of correctly classified positive examples by the total number of predicted positive examples.

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

Recall: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples.

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

Accuracy is given by:

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

Conclusion

Suspicious activity detection were analysed from the weapons. But our work presented a new approach which would use facial expressions of subjects along with the suspicious weapons for analysing the context to better predict the suspicious activities and reduce false predictions. Both SSD and Faster RCNN provided results with 82.84% and 82.48% accuracy. Faster RCNN being good in detecting small guns (91 of 94) whereas (66 of 94) in case of SSD and SSD being good in providing results in less amount of time (1.5 seconds)to process one frame whereas (2.5seconds) required for Faster RCNN . This system can be used to monitor any public place, private shops, can be used by government agencies to detect suspicious activities and prevent it.

I. If the system is to be used at places where distance of subjects would be less than 10m, SSD model can be used for quick response.
II. If the system has to be used in public places where small and distant objects are of our concern then Faster RCNN model can be used for better precision.

Assumptions:

1. In day light, an appropriate camera with high configuration that can capture objects from the required distance.
2. To detect threats at night , a night vision camera should be used for the system to work properly,
3. The system fails and acts only as a weapon detection system when the faces of the subjects are not visible.
4. The system works with acceptable performance with camera radius about 15 m, but can work for more than that if more highly configured camera is used.

Future Scope

1. The facial expression accuracy achieved is about 60% and in our system we came up with 56% accuracy. So there is a chance to improve the accuracy by considering more features of face by taking fine details of faces to classify expression to boost overall threat detection.
2. As the CCTV cameras are fixed, the facial expression of a person facing against camera cannot be detected or there is limitation on area of surveillance, we can also think of using drones to avoid such cases.
3. As the real time system is time consuming we can improve it by using parallel computing wherein the frames will be processed in parallel which will help in minimising time required to process a frame.

References

[1] **CNN Features off-the-shelf: an Astounding Baseline for Recognition**
Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson CVAP, KTH (Royal Institute of Technology) Stockholm, Sweden
http://openaccess.thecvf.com/content_cvpr_workshops_2014/W15/papers/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.pdf

[2] **An Improved Faster R-CNN for Same Object Retrieval**
HAILIANG LI1 , YONGQIAN HUANG2 , AND ZHIJUN ZHANG2 1School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510000, China 2School of Automation Science and Engineering, South China University of Technology, Guangzhou 510000, China
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7986979)

[3] **SSD: Single Shot MultiBox Detector**
Wei Liu1, Dragomir Anguelov2 , Dumitru Erhan3 , Christian Szegedy3 , Scott Reed4 , Cheng-Yang Fu1 , Alexander C. Berg1 1UNC Chapel Hill 2Zoox Inc. 3Google Inc. 4University of Michigan, Ann-Arbor
https://www.cs.unc.edu/~wliu/papers/ssd.pdf

[4] **Developing a Real-Time Gun Detection Classifier**
Justin Lai Stanford University jzlai@stanford.edu Sydney Maples Stanford University smaples@stanford.edu
http://cs231n.stanford.edu/reports/2017/pdfs/716.pdf

[5] **Automatic Handgun Detection Alarm in Videos Using Deep Learning**
Roberto Olmos 1, Siham Tabik 1, and Francisco Herrera1, 2 1. Soft Computing and Intelligent Information Systems research group 2. Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain. Email: siham@ugr.es, herrera@decsai.ugr.es February 20, 2017
https://arxiv.org/pdf/1702.05147.pdf

[6] **Facial Expression Recognition with Faster R-CNN**
https://www.sciencedirect.com/science/article/pii/S1877050917303447

[7] **Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order**
https://www.researchgate.net/publication/305483977_Facial_Expression_Recognition_with_Convolutional_Neural_Networks_Coping_with_Few_Data_and_the_Training_Sample_Order

[8] Object detection with deep learning and OpenCV by **Adrian Rosebrock,**
https://www.pyimagesearch.com/2017/09/11/object-detection-with-deep-learning-and-opencv/

[9] LabelImg: A Graphical Image annotation tool.
https://github.com/tzutalin/labelImg
Reference Video Link: https://youtu.be/p0nR2YsCY_U