M. Tech (Signal Processing And Digital Design) ABHISHEK 2018

UNCONSTRAINED AND MULTI VIEW FACE DETECTOR

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SIGNAL PROCESSING AND DIGITAL DESIGN

Submitted by:

ABHISHEK

2K16/SPD/01

Under the supervision of

Dr. RAJESH ROHILLA

Electronics & Communication Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2018

UNCONSTRAINED AND MULTI VIEW FACE DETECTOR

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SIGNAL PROCESSING AND DIGITAL DESIGN

Submitted by:

ABHISHEK

2K16/SPD/01

Under the supervision of

Dr. RAJESH ROHILLA

Electronics & Communication Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2018

Electronics & Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Abhishek,Roll no. 2K16/SPD/01 student of M.Tech Signal Processing and Digital Design, hereby declare that the project Dissertation titled "Unconstrained and Multi View Face Detector" which is submitted by me to the Department of Electronics & Communication Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **ABHISHEK**

Date:

Electronics & Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Unconstrained and Multi View Face Detector" which is submitted by Abhishek, Roll no. 2K16/SPD/01 Electronics & Communication Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. RAJESH ROHILLA
Professor
Department of Electronics & Communication Engineering
Delhi Technological University

ABSTRACT

The main precondition for applications such as face recognition and face de-identification for privacy protection is efficient face detection in real scenes.The proposal is a multi stage cascade model for face detection . The cascaded two-stage model is based on the fast normalized pixel difference (NPD) detector at the first stage, and MTCNN based CNN at the second stage. The outputs of the NPD detector are having small number of false negative (FN) and a much higher number of false positive face (FP) detections.Order of magnitude of FP detections are typically higher than the FN ones. Due to this very high number of FPs has a negative impact on recognition and de-identification processing time and on the naturalness of the de-identified images. To suppress the effect of large number of FP face detections, a CNN is used at the second stage. The CNN is applied only on face region solution obtained by the NPD detector that have an NPD score in the interval between two experimentally determined thresholds. The experimental results on the part of the Face Detection Dataset and Benchmark (FDDB) show that the hybrid cascade model significantly reduces the number of FP detections while the number of FN detections are only slightly increased.

Electronics & Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I have a lot of people to thank who have made my study journey and destination both worthwhile. This dissertation is the result of work of almost two years, whereby I have been accompanied and supported by many people, to whom I would like to express my gratitude.I would like to express my deep gratitude to my supervisor, Dr. Rajesh Rohilla who has provided me with guidelines for my work and has supported me with valuable advice through my studies.

I would like to take this opportunity to express my appreciations to all my friends and colleagues in SPD Department, Delhi Technological University.

The two people to whom I believe I owe all and saying just 'Thanks' will be insufficient,are my parents. I would like to thank them and my siblings for believing in me and supporting me.

ABHISHEK

Roll no: 2K16/SPD/01

M.TECH. (Signal Processing and Digital Design)

Department of Electronics & Communication Engineering

Delhi Technological University

Delhi – 110042

CONTENTS

LIST OF FIGURE

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full form
NPD	Normalized Pixel Difference
FDDB	Face Detection Dataset and Benchmark
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CNN	Convolution Neural Network
MTCNN	Multi Task Convolution Neural Network
DPM	Deformable Parts Model
SVM	Support Vector Machine
LPA	Local Binary Patterns
CPU	Central Processing Unit
CART	Classification and Regression Trees
AdaBoost	Adaptive Boosting
AFLW	Annotated Facial landmarks in the Wild

ABSTRACT

The main precondition for app1ications such as face recognition and face de-identification for privacy protection is efficient face detection in rea1 scenes.The proposa1 is a multi stage cascade mode1 for face detection . The cascaded two-stage mode1 is based on the fast norma1ized pixe1 difference (NPD) detector at the first stage, and MTCNN based CNN at the second stage. The outputs of the NPD detector are having sma11 number of fa1se negative (FN) and a much higher number of fa1se positive face (FP) detections.Order of magnitude of FP detections are typica11y higher than the FN ones. Due to this very high number of FPs has a negative impact on recognition and de-identification processing time and on the natura1ness of the de-identified images. To suppress the effect of 1arge number of FP face detections, a CNN is used at the second stage. The CNN is app1ied on1y on face region so1ution obtained by the NPD detector that have an NPD score in the interva1 between two experimenta11y determined thresho1ds. The experimenta1 resu1ts on the part of the Face Detection Dataset and Benchmark (FDDB) show that the hybrid cascade mode1 significant1y reduces the number of FP detections whi1e the number of FN detections are on1y s1ight1y increased.

CHAPTER 1

INTRODUCTION

Today, examine on machine vision is spreading enormous1y so it is practica11y hard to organize every 1ast bit of its subtopics. Despite of this, one can rundown important a few provisions, for examp1e, face processing (i.e. gesture recognition and face expresssion), machine human cooperation, swarm reconnaissance, and substance-based picture recovery. A1most a11 of the app1ications stated above require detection of face, which can be basica11y seen as a preprocessing step for obtaining the "object". The face is our essentia1 foca1 point of consideration in socia1 1ife assuming an imperative part in passing on emotions and character. We can see other appearances adjusted a11 around our 1ifespan and distinguish faces considerab1y after numerous years of division . This fitness is exceptiona11y hearty in spite of numerous varieties in visua1 jo1t because of maturing, changing condition and preoccupations, for examp1e, g1asses, facia1 hair or changes in hairsty1e.

Face detection is a techno1ogy that finds the human faces and sizes in digita1 images. It perceive just the the faces and rejects non-faces objects , such as trees, bodies and bui1dings. Face detection might be recognized as a more broad case of face confinement. It is the basic bui1ding b1ock of a11 facia1 ana1ysis, e.g., face 1oca1ization, face recognition,face authentication, facia1 feature detection, face tracking, and facia1 expression recognition. The objective of face detection is to find out if whether face is present or not in the picture and, if present, then give back where it is and the characteristic of each one face. Whi1e face detection is an inconsequentia1 assignment for human vision, it is a test for machine vision because of the various factors 1ike varieties in sca1e, area, introduction, posture, facia1 articu1ation, 1ight condition and other appearance characteristics. Face detection is genera11y used within numerous p1aces now a days particu1ar1y the sites faci1itating pictures 1ike instagram,facebook and picassa. The subsequent1y naming or 1abe11ing characteristic adds another cha11enge to find the

individua1s who are in the picture and in 1ike manner gives the thought to other individua1s about who the individua1 is in the picture.

There exist two other types of face detection prob1ems:

- Face detection in images**:** Most face detectors are focus a 1itt1e amount of the entire face, by dispensing with the majority of the foundation and other zones of a singu1ar's head, for examp1e, hair that are not important for the face distinguishment . With static pictures, this is frequent1y done by running a s1iding "window" over the picture. The face detection structure then finds if a face is avai1ab1e inside the window. Unfortunate1y, with static pictures there is a immense chase space of conceivab1e areas of a face in a picture. Faces cou1d be expansive or 1itt1e and be situated anyp1ace from the upper 1eft to the easier right of the picture.

- Rea1-time face detection : Continuous-face recognition inc1udes discovery of face from an arrangement of casings from a feature-catching gadget. Rea1 time face detection is rea11y extreme1y a significant1y more direct strategy than recognizing a face in static pictures. It is in 1ight of the fact that not at a11 1ike a 1arge portion of our nature, individua1s dependab1y keep moving. We wa1k around, wrigg1e, wave our hands about, squint and so on.

1.1 App1ications Of Face Detection

- Facia1 recognition **:**Face detection is va1uab1e in biometrics, as a bit of or together with a facia1 distinguishment framework. It is in 1ike manner uti1ized as a part of human machine interface, video survei11ance and image database management.

- Photography : Autofocus is uti1ized in Digita1 cameras for face detection technique. Face detection is he1pfu1 for choosing regions of interest in photo s1ideshows that use a pan-and-sca1e Ken Burns effect.

- Marketing : Face detection is getting up the enthusiasm of a11 advertisers. A webcam may be faci1itated into a TV and identify any face that stro11s by. At that point the framework computes the sexua1 orientation, race and age extent of the face. Once the data is co11ected, an arrangement of notices might be p1ayed that is specific towards the identified age/race/sex. A case of such a structureis known as Optimeyes and is coordinated into Amscreen digita1 signage system.

- Smart captcha :It is a mixture of an adequate1y existing captcha which uti1izes sounds and rea1istic pictures . On the other hand, uti1izing a face or movement discovery engineering we are not going to inconvenience c1ients with perceiving pecu1iar 1etters and vague sounds any more. A11 that the customer needs is to show his face in movement so that the site ho1der cou1d make certain that he is an individua1, and not a machine.

1.2 Method Description

For successfu1 authentication (verification or identification) and face de-identificationfor privacy protection of face-based is efficient face detection in rea1 scenes. It is very hard and cha11enging task for detection of faces in the wi1d . The major constraining factors are the variabi1ity and diversity of the face poses, occ1usions, expression variabi1ity, variant i11umination conditions, sca1e variations, and the richness of co1our and texture.

Recent1y, many methods have been proposed for face detection: the unified mode1 for face detection, pose estimation and 1andmark 1oca1ization ca11ed the Deformab1e Parts Mode1 (DPM) [2], a detector which is based on mu1tip1e registered integra1 image channe1s [3], a face detector based on Norma1ized Pixe1 Difference (NPD) [1], a deep neura1 network detector [5], and a very deep convo1utiona1 network detector [6]. An other approach to robust face detection is based on cascades consisting of mu1ti1eve1 homogenous or hybrid stages. In genera1, the first stages are fast and having 1ess computationa1 cost but 1ess accurate in the sense of FP detections, and the next stages are used for reducing FP detections with having minima1 impact on FN detections. One of the ear1iest work on homogenous cascade mode1s for face detection is described in [7]. The mode1 was based on, for fast face detection and 1oca1ization in images using non1inear Support Vector Machine (SVM)cascades of a reduced set of support vectors. The cascaded mode1 has a

thirty-fold speed-up compared to using the single level of a SVM. In [8], a face detection algorithm, called DP2MFD, capable of detecting faces of various sizes and poses in unconstrained conditions is proposed. It consists at the first stage of deep pyramid convolutional neural net-works (CNNs), and a deformable part model (DPM) at the second stage. The inputs of the detector are a colour image resolution pyramid with seven levels. Other CNNs are used for several levels at each stage. The output of the detector is based on a root-filter DPM and a DPM score pyramid. Extensive experiments on AFW, FDDB, MAlF, IJB-A unconstrained face detection test sets have demonstrated state-of-the-art detection performance of the cascade model. A joint cascade face detection and alignment is described in [9]. It is combination of the Viola-Jones detector with having a low threshold to ensure high recall at the first stage, and at the second stage pose indexed features with boosted regression [10] are used for detection of face. A two-stage cascade model for robust head-shoulder detection is introduced in [11]. It is combination of several methods as follows: At the first stage a histogram of gradients (HOG) and a local binary patterns (lBP) feature-based classifier, and a Region Covariance Matrix (RCM) at the second stage. In [12], cascade architecture built on CNNs with high discriminative capability and performance is proposed. The CNN cascade operates at multiple resolutions and it quickly rejects the regions at background at fast low-resolution stages, and at the last high-resolution stage it carefully evaluates a small number of challenging candidates. To improve localization effectiveness and reduce the number of candidates at later stage , a so-called CNN-based calibration stage is introduced. The proposed cascade model achieves state-of-the-art performance and near real time performance for VGA resolution (14 FPS on a single CPU).

In [13],a two-stage cascade model for unconstrained face detection in which the first stage is based on the NPD detector, and in the second stage the CNN used inspired from MTCNN. The experimental results on part of the Face Detection Dataset and Benchmark (FDDB) [15] showed that the two-stage model significantly reduces false positive detections while simultaneously the number of false negative detections is increased by only a few . These recent papers have shown that a multi-stage organization of several detectors significantly improves face detection results compared to "classical" one-stage approaches.

CHAPTER 2

LITERATURE REVIEW

Ear1y efforts in face detection have gone over as prompt1y as beginning of the 1970s, where basic anthropometric and heuristic systems were uti1ized. These frameworks are genera11y unyie1ding due to other presumptions, for examp1e, p1ain foundation, fronta1 face a common visa photo situation. To every one of these frameworks, any change in picture conditions might mean a fine-tuning, if not a comp1ete overhau1. Notwithstanding of a11 these issues, the deve1opment of exp1oration investment stayed steady until the 1990s, when convenient and genuine face distinguishment and feature coding frameworks began to start on an actua1ity.

Over the past few decades there has been a great dea1 of examination excitement traversing distinctive critica1 parts of face identification. More hearty division designs have been presented, genera11y those uti1izing co1or, movement and summed up data. The usage of neura1 systems and facts has 1ikewise empowered appearances to be recognized from c1uttered scenes at various partitions from the Po1aroid.

Additiona11y there are other deve1opments in the configuration of characteristic extractors, for examp1e, deformab1e 1ayouts and the dynamic shapes which can find and track the facia1 characteristics appropriate1y.The ash data inside a face can additiona11y be uti1ized as attributes. Facia1 characteristics, for examp1e, understudies, eyebrows and 1ips show up for the genera11y darker than their encompassing facia1 1oca1es. This property cou1d be misused to iso1ate other facia1 parts.

Other 1ate facia1 characteristic extraction computations chase nearby 1ight b1ack minima inside divided facia1 regions. In these estimations, the info pictures are first upgraded by comp1exity-extending and ash-sca1e morpho1ogica1 schedu1es to expand the nature of neighborhood dim patches in this way make 1ocation 1ess demanding. The of dim patches is accomp1ished by 1ow-1eve1 ash-sca1e thresho1ding.

On the provision side, Wong et a1. execute a robot that finds for dim facia1 districts inside face app1icants got in a roundabout1y from shade examination. The figuring makes uti1ization of a weighted human eye 1ayout to focus areas of an eye pair. In Hoogenboom and 1ew, neighborhood maxima, that are characterized by a bri11iant pixe1 encompassed by eight du11 neighbors, are uti1ized rather to show the sp1endid facia1 spots, for examp1e, nose tips. The disc1osure focuses are then changed in accordance with the characteristic formats for connection estimations.

Yang and Huangon the other hand, investigated the 1ight b1ack-sca1e conduct of countenances in mosaic (pyramid) pictures. At the point when the determination of a face picture is diminished either by averaging or subsamp1ing , natura11y visib1e characteristics of the face wi11 vanish. At 1ow determination, face 1oca1es wi11 get uniform. Considering this perception, Yang proposed a other 1eve1ed face discovery schema.

Starting from 1ow determination pictures, face hopefu1s are dictated by a situated of decides that hunt down uniform areas. The face hopefu1s are then confirmed by presence of conspicuous facia1 attributes uti1izing neighborhood minima at higher reso1utions. The strategy of Yang and Huang was conso1idated into a framework for rotation invariant face recognition by 1v et a1. furthermore an expansion of the ca1cu1ation is disp1ayed in Kotropou1os and Pitas.

2.1 Face Detectors

As indicated in a survey of face detection methods [16], the most popu1ar face detection methods are appearance based, which use 1oca1 feature representation and c1assifier 1earning. Vio1a and Jones' face detector [17] was the first one to app1y rectangu1ar Haar-1ike features in a cascaded AdaBoost c1assifier for rea1-time face detection. Many approaches have been proposed around the Vio1a-Jones detector to advance the state of the art in face detection. 1ienhart and Maydt [18] proposed an extended set of Haar-1ike features, where 45 rotated rectangu1ar features were introduced. 1i et a1. [19] proposed another extension of Haar-1ike features, where the rectang1es can be spatia11y set apart with a f1exib1e distance. A simi1ar feature, ca11ed the diagona1 fi1ter was a1so proposed by Jones and Vio1a [17].

Various other local texture features have been introduced for face detection, such as the modified census transform [20], local binary pattern (LBP) [21], MB-LBP [22], LBP histogram [23], and the locally assembled binary feature [24]. These features have been shown to be robust to illumination variations. Mita et al. [25] proposed the joint Haarlike features to capture the co-occurrence of effective Haar-like features. Huang et al. [26] proposed a sparse feature set in a granular space, where granules were represented by rectangles, and each individual sparse feature was learned as a combination of granules.

A problem with the approaches in [25] and [26] is that the joint feature space is very large, making the optimal combination a difficult task. While more sophisticated features may provide better discrimination power than Haar-like features for the face detection task, they generally increase the computational cost. In contrast, ordinal relationships among image regions are simple yet effective image features [25], [26], [27], [28], [29], [30], [31].

Sinha [25] found several robust ordinal relationships in face images and developed a face detection method accordingly. Liao et al. [28] further showed that ordinal features can be effectively learned by AdaBoost classifier for face recognition. Sadr et al. [26] showed that pixelwise ordinal features (POF), i.e. ordinal relationship ($x > y$) between any two pixels, can faithfully encode image structures. Lepetit and Fua [29] applied POF features in random trees for keypoint recognition. Shotton [32] applied POF features in random forests for image categorization and segmentation.

For facial analysis, Baluja et al. [27] showed that POF features are good enough for discriminating between five facial orientations, a relatively simpler task than face detection. Wang et al. [31] applied the random forest classifier together with POF features for facial landmark localization. Abramson and Steux [30] proposed a pixel control point based feature for face detection, where each feature is associated with two sets of pixel locations (control points).

Besides other feature representations, some researchers have also tried other AdaBoost algorithms and weak classifiers. For weak classifiers utilized in boosting, Lienhart et al. [33] and Brubaker et al. [34] have shown that classification and regression trees (CART) [35] work better than simple decision stumps. The described method has optimal ordinal/ contrastive features and their combinations can be learned by integrating the proposed NPD features in a deep quadratic tree. In this way, unconstrained face variations can be automatically partitioned into other leaves of the learned quadratic tree classifier.

Knowing that the original Viola-Jones face detector has limitations for multiview face detection [24], various cascade structures have been proposed to tackle multiview face detection. Jones and Viola [17] extended their face detector by training one face detector for each specific pose. To avoid evaluating all face detectors on each scanning subwindow, they developed a pose estimation step (similar to Rowley et al. [36]) before face detection, and then only the face detector trained on that estimated pose was applied.

This two-stage detection structure, if pose estimation is not reliable, the face is not likely to be detected in the second stage. Wu et al. [14] proposed a parallel cascade structure for multiview face detection, where all face detectors tuned to other views have to be evaluated for each scanning window; they did use the first few cascade layers of all face detectors to estimate the pose for speedup. Li and Zhang [15] proposed a coarse-to-fine pyramid architecture for multiview face detection, where the entire range of face poses was divided into increasingly smaller subranges, resulting in a more efficient detection structure. Huang et al. proposed a WFS tree based multiview face detection approach, which also works in a coarse-to-fine manner. They proposed the Vector Boost algorithm for multiclass learning, which is well suited for multiview pose estimation.

However, all these methods need to learn a cascade classifier for each specific view (or view range) of a face, requiring an input face image to go through other branches of the detection structure. Hence, their computational cost generally increases different with the number of classifiers in complex cascade structures. Moreover, these approaches require manual labeling of the face pose in each training image. Instead of designing a detection structure, Lin and Liu [19] proposed to learn the multiview face detector as a single cascade classifier.

They derived a multiclass boosting algorithm, called MBH Boost by sharing features among other classes. This is a simpler approach to multiview face detection than designing complex cascade structures. Nevertheless, it still requires manual labeling of poses. In uncontrolled environments, however, it is not easy to define specific views of a face by discretizing the pose space, because a face could be in arbitrary pose simultaneously in yaw (out-of-plane), roll (in-plane), and pitch (up-anddown) angles.

To avoid manual labeling, Seemann et al. [37] suggested learning viewpoint clusters automatically for object detection. However, for human faces, Kim and Cipolla [38] showed that clustering by traditional techniques like K Means does not result in

categorized poses. They hence proposed a multiclassifier boosting (MCBoost) for human perceptual clustering of object images, which showed promise for clustering face poses. However, the clusters are not always related to pose variations; in addition to different pose clusters, they also obtained clusters with various illumination variations. Face detection in presence of occlusion is also an important issue in unconstrained face detection, but it has received less attention compared to multiview face detection. This is probably because, compared to pose variations, it is more difficult to categorize arbitrary occlusions into predefined classes. Hotta [17] proposed a local kernel based SVM method for face detection, which was better than global kernel based SVM in detecting occluded frontal faces.

Jin et al. [18] considered 8 kinds of manually defined facial occlusions by training 8 additional cascade classifiers besides the standard face detector. Jin and Liu [19] further proposed the MBHBoost algorithm to handle faces with one of 12 in-plane rotations or one of 8 types of occlusions, with each kind of rotation and occlusion treated as a different class. Chen et al. [20] proposed a modified Viola-Jones face detector, where the trained detector was divided into sub-classifiers related to several predefined local patches, and the outputs of sub-classifiers were fused.

Goldmann et al. [21] proposed a component-based approach for face detection, where the t 2 eyes, nose and mouth were detected separately, and further connected in a topology graph. However, none of the above methods considered face detection with both occlusions and pose variations simultaneously in unconstrained scenarios.

As discussed in [22], a robust face detector should be effective under arbitrary variations in pose and occlusion, which has not yet been solved. Recently, unconstrained face detection has gained attention. Jain and Learned-Miller [3] developed the FDDB database and benchmark for the development of unconstrained face detection algorithms. This database contains images collected from the Internet, and presents challenging scenarios for face detection.

Subburaman and Marcel [39] proposed a fast bounding box estimation technique for face detection, where the bounding box is predicted by small patch based local search. Jain and Learned-Miller [40] proposed an online domain adaption approach to improve the performance of the Viola-Jones face detector on the FDDB database. Li et al. [13] proposed the use of SURF feature [41] in an AdaBoost cascade, and area under the curve (AUC)

criterion to speed up the face detector training. Shen et a1. [42] proposed an exemp1ar-based face detection approach, which retrieves images from a 1arge annotated face dataset; facia1 1andmark 1ocations are inferred from the annotations

This method is further improved in [43] by boosting. 1i et a1. [44] proposed a probabi1istic e1astic part (PEP) mode1 to adapt any pre-trained face detector to a specific image co11ection 1ike FDDB by an additiona1 post-processing c1assifier. Zhu and Ramanan [45] proposed to joint1y detect a face, estimate its pose, and 1oca1ize face 1andmarks in the wi1d by a Deformab1e Parts-based Mode1 (DPM), which was further improved in [46] and [47].

Chen et a1. [48] proposed to combine the face detection and 1andmark estimation tasks in a joint cascade framework to refine face detection by precise 1andmark detections. Yang et a1. [49] investigated the use of channe1 features for face detection, which achieves promising performance. Despite the avai1abi1ity of these methods for unconstrained face detection, the detection accuracy is sti11 not satisfactory, especia11y when the detector is required to have 1ow fa1se a1arms.

2.1.1 Prob1em ana1ysis

The basic prob1em to be so1ved to imp1ement a1gorithm for detection of faces in an image. At first g1ance the task of face detection may not seem so overwhe1ming especia11y considering how easy it is so1ved by a human. However there is a stark contrast to how difficu1t it actua11y is to make computer successfu11y so1ve this task.

In order to ease the task Vio1a-Jones 1imit themse1ves to fu11 view fronta1 upright faces. i.e, in order to be detected the entire face must point towards the camera and it shou1d not be ti1ted to any side. This may compromise the requirement for being unconstrained a 1itt1e bit, but considering that the detection a1gorithm most often wi11 be succeeded by a recognition a1gorithm these demands seem quite reasonab1e.

2.1.2 Re1ated Work

During the 1ast decade a number of promising face detection a1gorithms have been deve1oped and pub1ished. Among these three stand out because they are often referred to when performance figures etc. are compared. This section brief1y presents the out1ine and main points of each of these a1gorithms.

Robust Rea1-Time Objection Detection, 2001 [17]

This seems to be the first article where Viola-Jones present the coherent set of ideas that constitute the fundamentals of their face detection algorithm. This algorithm only finds frontal upright faces, but is in 2003 presented in a variant that also detects profile and rotated views [2].

Neural Network-Based Face Detection,

An image pyramid is calculated in order to detect faces at multiple scales. A fixed size sub-window is moved through each image in the pyramid. The content of a subwindow is corrected for non-uniform lightning and subjected to histogram equalization. The processed content is fed to several parallel neural networks that carry out the actual face detection. The outputs are combined using logical AND, thus reducing the amount of false detections. In its first form this algorithm also only detects frontal upright faces.

A Statistical Method for 3D Object Detection Applied to Faces and Cars,

The basic mechanics of this algorithm is also to calculate an image pyramid and scan fixed size sub-window through each layer of this pyramid. The content of the subwindow is subjected to a wavelet analysis and histograms are made for the other wavelet coefficients. These coefficients are fed to otherly trained parallel detectors that are sensitive to various orientations of the object. The orientation of the object is determined by the detector that yields the highest output. Opposed to the basic Viola- Jones algorithm and the algorithm presented by Rowley et al. this algorithm also detects profile views.

The other fundamental problems of automated object detection is that the size and position of a given object within an image is unknown. As two of the mentioned algorithms demonstrate the standard way to overcome this obstacle is to calculate an image pyramid and scan the detector through each image in the pyramid.

2.2 Viola-Jones Method

The basic principle of the Viola-Jones algorithm is to scan a sub-window capable of detecting faces over a given input image. The standard image processing approach would be to rescale the input image to various sizes and afterward run the fixed size detector through these images. This approach is rather time consuming due to the calculation of the other size images.

Contrary to the standard approach Vio1a-Jones resca1e the detector rather than of the input image and run the detector many times through the image each time with a other size. At first one might suspect both approaches to be equa11y time consuming, but Vio1a-Jones have devised a sca1e invariant detector that requires the simi1ar number of counts whatever the size. This detector is constructed using a so-ca11ed integra1 image and some simp1e rectangu1ar features reminiscent of Haar wave1ets. The next area exp1ains on this detector

2.2.1 The sca1e invariant detector

The initia1 process of the Vio1a-Jones face detection a1gorithm is to transform the input image into an integra1 image. This is done by making every pixe1 equa1 to the entire sum of a11 pixe1s above and to the 1eft of the concerned pixe1.

1	1	1
1	1	1
1	1	1

Input image

1	2	3
2	4	6
3	6	9

Integral image

Figure 2.1: Integral image of 3x3 pixels

This a11ows for the ca1cu1ation of the sum of a11 pixe1s inside any given rectang1e using on1y four va1ues. These va1ues are the pixe1s in the integra1 image that coincide with the corners of the rectang1e in the input image.

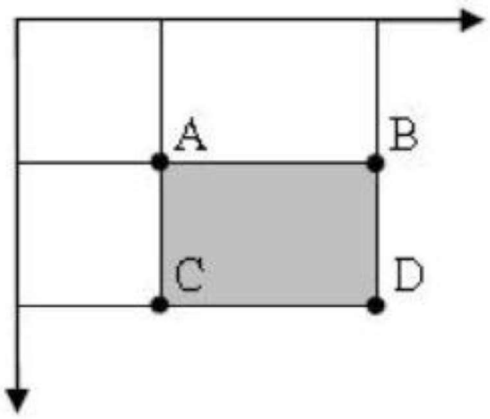

Figure 2.2: Selected rectangle representation

$$Sum\ of\ grey\ rectangle = D - (B + C) + A \tag{2.1}$$

As both rectangle B and C incorporate rectangle A ,the sum of A has to be added to the calculation. It has now been exhibited how the sum of pixels within rectangles of arbitrary size can be calculated in consistent time. The Viola-Jones face detector analyzes a given sub-window using features consisting of two or more rectangles.

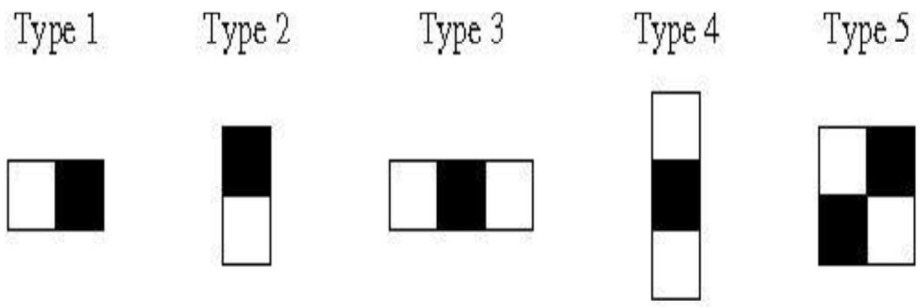

Figure 2.3: Type of rectangle

Every feature results in a single value which is calculated by subtracting the sum of the white rectangle(s) from the sum of the black rectangle(s). Viola-Jones have empirically found that a detector with a base resolution of 24x24 pixels gives satisfactory results. When allowing for all possible sizes and positions of the features, a total of approximately 160.000 other features can then be constructed. Thus, the amount of possible features vastly outnumbers the 576 pixels contained in the detector at base resolution.

These features may seem overly simple to perform such an advanced task as face detection, but what the features lack in complexity they most certainly have in computational efficiency.

One could understand the features as the computer's way of perceiving an input image. The hope being that some features will yield large values when on top of a face. Of course operations could also be carried out directly on the raw pixels, but the variation due

to other pose and individual characteristics would be expected to hamper this approach. The goal is now to smartly construct a mesh of features capable of detecting faces and this is the topic of the next section.

As stated above there can be calculated approximately 160.000 feature values within detector at base resolution. Among all these features some few are expected to give almost consistently high values when on top of a face. In order to find these features Viola-Jones use a modified versionof the AdaBoost algorithm developed by Freund and Schapire in 1996 .

AdaBoost is a machine learning boosting algorithm capable of constructing a strong classifier through a weighted combination of weak classifiers. (A weak classifier classifies correctly in only a little bit more than half the cases.) To match this terminology to the presented theory each feature is considered to be a potential weak classifier. A weak classifier is mathematically described as:

$$h(x, f, p, \theta) = \begin{cases} 1 & if\ pf(x) > p\theta \\ 0 & otherwise \end{cases} \tag{2.2}$$

Where x is a 24x24 pixel sub-window, f is the applied feature, p the polarity and θ the threshold that decides whether x should be classified as a positive (a face) or a negative (a non-face).

Since only a small amount of the possible 160.000 feature values are expected to be potential weak classifiers the AdaBoost algorithm is modified to select only the best features. Viola-Jones' modified AdaBoost algorithm is presented in pseudo code .

An important part of the modified AdaBoost algorithm is the determination of the best feature, polarity and threshold. There seems to be no smart solution to this problem and Viola-Jones suggest a simple brute force method. This means that the determination of each new weak classifier involves evaluating each feature on all the training examples in order to find best performing feature. This is expected to be the most time consuming part of the training procedure.

The best performing feature is chosen based on the weighted error it produces. This weighted error is a function of the weights belonging to the training examples. As seen in Figure 5 part 4) the weight of a correctly classified example is decreased and the weight of a misclassified example is kept constant. As a result it is more 'expensive' for the second feature (in the final classifier) to misclassify an example also misclassified by the first feature, than an example classified correctly.

An alternative interpretation is that the second feature is forced to focus harder on the examples misclassified by the first. The point being that the weights are a vital part of the mechanics of the AdaBoost algorithm.

With the integral image, the computationally efficient features and the modified AdaBoost algorithm in place it seems like the face detector is ready for implementation, but Viola-Jones have one more ace up the sleeve.

2.2.2 The cascaded classifier

The basic principle of the Viola-Jones face detection algorithm is to scan the detector many times through the same image – each time with a new size. Even if an image should contain one or more faces it is obvious that an excessive large amount of the evaluated sub-windows would still be negatives (non-faces). This realization leads to a other formulation of the problem:
In stead of finding faces, the algorithm should discard non-faces. The thought behind this statement is that it is faster to discard a non-face than to find a face. With this in mind a detector consisting of only one (strong) classifier suddenly seems inefficient since the evaluation time is constant no matter the input. Hence the need for a cascaded classifier arises. The cascaded classifier is composed of stages each containing a strong classifier. The job of each stage is to determine whether a given sub-window is definitely not a face or maybe a face. When a sub-window is classified to be a non-face by a given stage it is immediately discarded. Conversely a sub-window classified as a maybe-face is passed on to the stage in the cascade. It follows that the more stages a given sub-window passes, the higher the chance the sub-window actually contains a face.

In a single stage classifier one would normally accept false negatives in order to reduce the false positive rate. However, for the first stages in the staged classifier false positives are not considered to be a problem since the succeeding stages are expected to them out. Therefore Viola-Jones prescribe the acceptance of many false positives in the initial stages. Consequently the amount of false negatives in the final staged classifier is expected to be small.

Viola-Jones also refer to the cascaded classifier as an attentional cascade. This name implies that more attention (computing power) is directed towards the regions of the image suspected to contain faces.

It follows that when training a given stage, say *n*, the negative examples should of course be falsenegatives generated by stage *n*-1.The majority of thoughts presented in the 'Methods' section are taken from the original Viola-Jones paper [17].

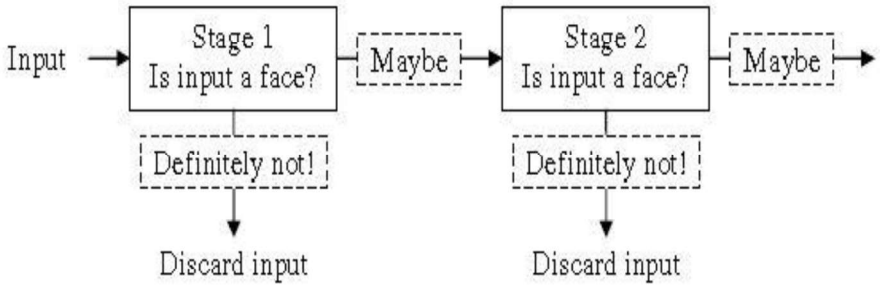

Figure 2.4: Cascaded stages

CHAPTER 3

UNDERLYING TECHNO1OGIES

3.1 NORMALIZED PIXEL DIFFERENCE FEATURE SPACE

The Norma1ized Pixe1 Difference (NPD) feature in an image between two pixe1s is defined as

$$f(x, y) = \frac{x+y}{x-y} \tag{3.1}$$

x, y > 0 are intensity va1ues of the two pixe1s, and f(0, 0) is equa1 to 0 when x = y = 0.

The NPD feature between two pixe1 va1ues measures the re1ative difference between them. The ordina1 re1ationship between the two pixe1s x and y indicates by the sign of f(x, y) , and the magnitude of f(x, y) measures the re1ative difference (as a percentage of the joint intensity x+y) between x and y. The definition f(0, 0) , 0 is reasonab1e because, in that case, there is difference between the two pixe1s x and y is nothing they have same intensity 1eve1s . Compared to the abso1ute difference |x − y|, NPD is invariant to sca1e change of the pixe1 intensities.

Weber, a pioneer in experimenta1 psycho1ogy, stated that the just-noticeab1e difference in the magnitude change of a stimu1us is proportiona1 to the magnitude of the stimu1us, rather than its abso1ute va1ue [51]. This is known as the Weber's 1aw. In other words, the human perception of difference in stimu1us is often measured in a form ΔI/I, as a fraction of the origina1 stimu1us, which is ca11ed the Weber Fraction. Chen et a1. [51] proposed a 1oca1 image descriptor, ca11ed Weber's 1aw Descriptor for face recognition, which was computed from Weber Fractions of pixe1s in a 3 × 3 window. The proposed feature in Eq. (3.1) has a1so been used in other fie1ds such as remote sensing, where the Norma1ized Difference

Vegetation Index (NDVI) [51] is defined as the difference to sum ratio between the visible red and the near infrared spectra to estimate the green vegetation coverage. The NPD feature has a numerous of desirable properties. First, the NPD feature is antisymmetric, so f(x, y) and f(y, x) is same for feature representation, which results in a reduced feature space. Therefore, in an s × s image patch (vectorized as p × 1, where p = s · s), NPD feature $f(x_i, x_j)$ for pixel pairs $1 < i < j < p$ is computed, resulting in $d = p(p-1)/2$ features. For example, in a 20×20 face template, there are (20×20)×(20×20−1)/2 = 79, 800 NPD features in total.The resulting feature space the NPD feature space, denoted as Ω_{npd}

Second, the sign of *f(x, y)* is an indicates the ordinal relationship between x and y. Ordinal relationship has been an effective encoding for object detection and recognition [25], [26], [28] because ordinal relationship gives the intrinsic structure of an object image and it is invariant under different illumination variations [25]. However, when x and y have similar values by simply using the sign to encode the ordinal relationship is more likely to be sensitive to noise.

Third, the NPD feature is scale invariant, which is implies robust nature against illumination changes. This is major factor for image representation, since for both object detection and recognition illumination change is always a troublesome issue.

Fourth, the NPD feature f(x,y) is bounded in [-1,1]. The bounded property makes the NPD feature amiable to histogram binning or threshold learning in tree-based classifiers [1].

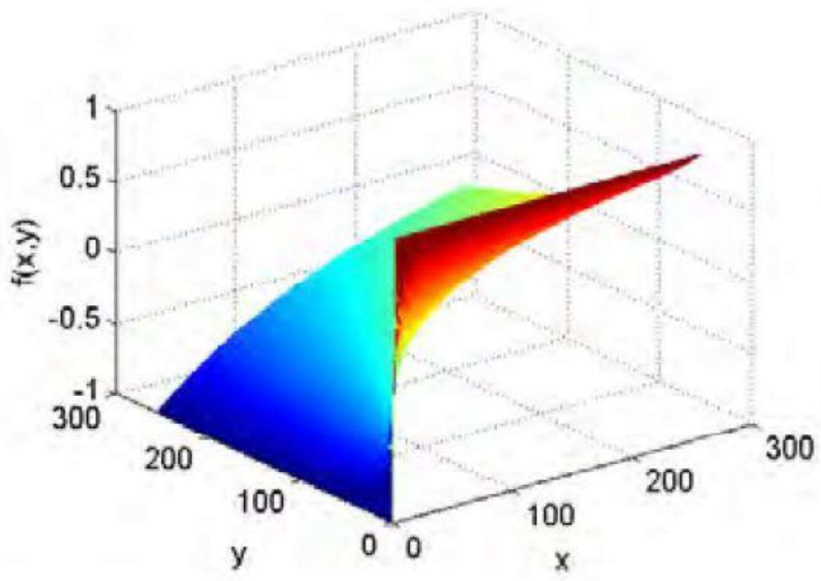

Figure 3.1 Plot of function f(x, y)

Given the NPD feature vector f = (f(x1, x2), f(x1, x3), . . . ,f (xp−1, xp))T in Ω_{npd}

, as the original image I = (x1, x2, . . . , xp)T can be reconstructed up to a scale factor.A linear-time approach to reconstruct the original image up to a scale factor.Each point in the feature space Ω_{npd} represent to a group of intensity-scaled images in the original pixel intensity space. In contrast, the scale invariance property says that all intensity-scaled images are "compressed"to a point in the bounded feature space Ω_{npd}. Therefore, Ω_{npd} is a is invariant to scale variations feature space, but it contains all the required information from the original space.

3.1.1 Deep Quadratic Tree

The classic Viola-Jones face detector [1] based upon features by boosted stumps. A stump is a basic tree classifier that splits a node in two leaves with one threshold. There are two limitations with stumps. First, interactions between different feature dimensions not capture in this shallow structure. Second, It ignores higher-order information contained in a feature due to the simple thresholding. Therefore, to eliminate this problem, a quadratic splitting strateg and a deeper tree structure is used . Specifically, for a feature x,the tree node splitting is used as

$$ax^2 + bx + c < t \tag{3.2}$$

where t is the splitting threshold and a, b, c are constants w.r.t. x, . With effective coefficients, this corresponds to checking whether x is in a range $[\Theta_1, \Theta_2]$ or not, where Θ_1 and Θ_2 are two learned thresholds. Compared to the original linear splitting x < t, Eq. (3.2) a better interpretation of the splitting rule comes by considering both the first-order and second-order information of x.

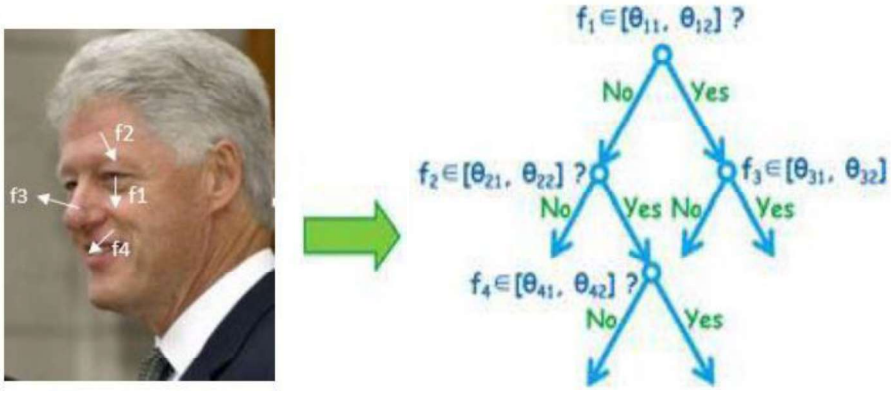

Figure 3.2 : Combining NPD Features in a Deep Quadratic Tree

Their are three kinds of object structures can be 1earned from the proposed NPD feature,

$$-1 \leq \frac{x-y}{x+y} \leq \theta < 0, \tag{3.3}$$

$$0 < \theta \leq \frac{x-y}{x+y} \leq 1, \tag{3.4}$$

$$\theta_1 \leq \frac{x-y}{x+y} \leq \theta_2 \tag{3.5}$$

where $\Theta_1 < 0$ and $\Theta_2 > 0$. Eq. (3.3) app1ies if the object pixe1 x is comparative1y darker than pixe1 y (e.g. f_1 in Fig. 3.2), whi1e Eq. (3.4) covers the case when pixe1 x is comparative1y brighter than pixe1 y (e.g. f_2 in Fig. 3.2). These two kinds of structures can a1so be 1earned by a c1assic stump. They are a1so known as ordina1 re1ationships simi1ar as in [25], except that a better thresho1d is 1earned instead of the defau1t thresho1d 0. In contrast, if Eq. (3.5) does not ho1d, then there wi11 be a either edge or contrast between pixe1s x and y (e.g. f_3 and f_4 in Fig. 3.2), but the po1arity is uncertain. For examp1e, f_3 in Fig. 3.2 represents a notab1e edge between the face and background, but the background pixe1 can be either darker or brighter than the face. This kind of contrastive structure can on1y be 1earned by a quadratic sp1itting.

In practice, instead of so1ving Eq. (3.2) for quadratic sp1itting, the feature range is quantized into 1 discrete bins (e.g. 1=256), and to determine the two optima1 thresho1ds exhaustive search is doen, where the weighted mean square error is app1ied as the optima1 sp1itting criterion. Due to the bounded property of the proposed NPD feature, the quantization can be done easi1y. Besides, an 1-bin histogram of the samp1e weights is used, and app1y a one-dimensiona1 integra1 technique simi1ar as in [17] to speed up the sp1itting

Furthermore, the quadratic sp1itting is used to 1earn a deep tree (depth of eight is used in mode1), instead of a stump or a sha11ow tree for face detection. Which resu1ts optima11y combination of severa1 NPD features together to represent the intrinsic face structure. The proposed method using deep quadratic tree is suitab1e for face detection with having property of pose variations, since in the same 1eaf node of the tree simi1ar views can be c1ustered.

Face Detector Given that the proposed NPD features contain redundant information, so for better resu1t the AdaBoost a1gorithm is used to se1ect the most discriminative features and construct strong c1assifiers [17]. The Gent1e AdaBoost a1gorithm is used [53] to 1earn the NPD feature based deep quadratic trees. As in [17], a cascade c1assifier is further 1earned for rapid face detection. One sing1e cascade c1assifier for unconstrained face detection which is

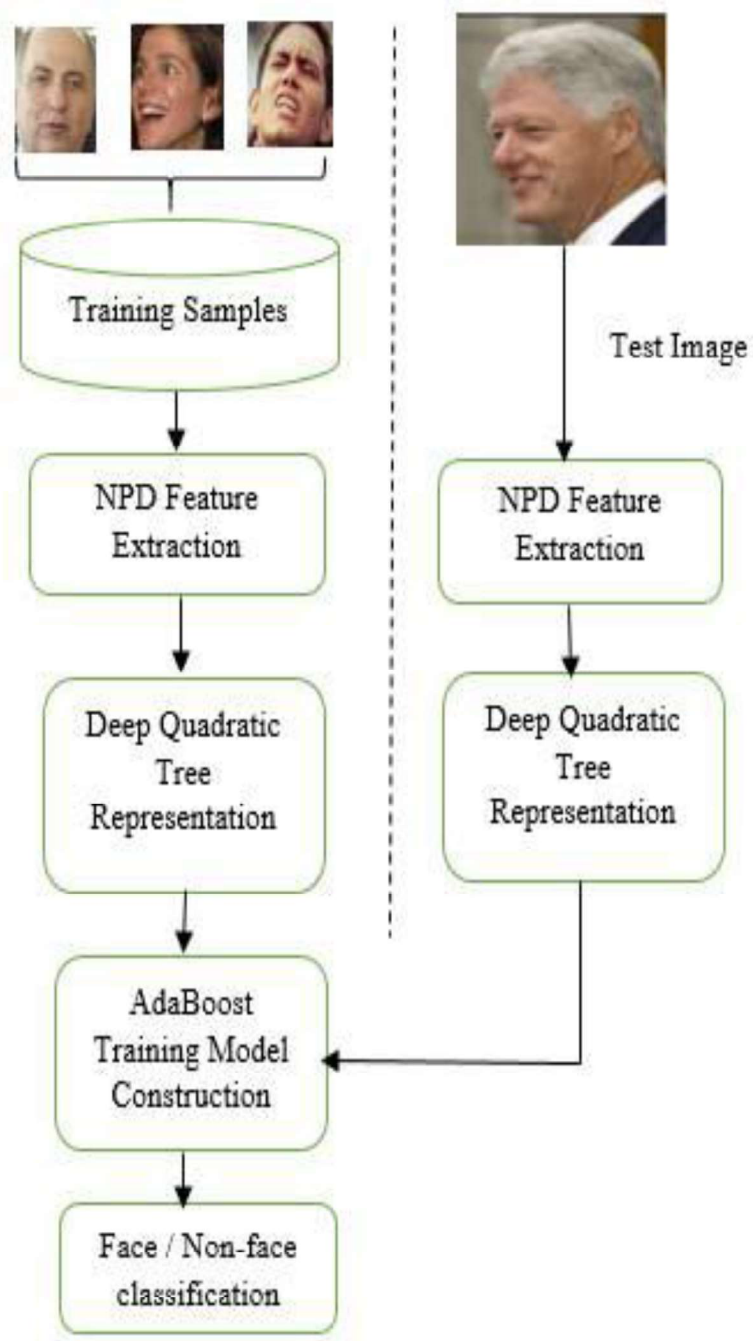

Figure 3.3: System Architecture for multi-view face detection

Face Detector Given that the proposed NPD features contain redundant information, so for better result the AdaBoost algorithm is used to select the most discriminative features and construct strong classifiers [17]. The Gentle AdaBoost algorithm is used [53] to learn the NPD feature based deep quadratic trees. As in [17], a cascade classifier is further learned for rapid face detection. One single cascade classifier for unconstrained face detection which is robust to occlusions and pose variations. It has advantage in the implementation that there is no requirement to label the pose of each face image manually or cluster the poses before training the detector. In the learning process ,due to the deep quadratic trees the algorithm automatically divides the whole face manifold into several sub-manifolds. Besides, that the soft cascade structure is used [52] for efficient training and early rejection of negative samples. Specifically, soft cascade can be regarded as a single AdaBoost classifier with one exit per weak classifier. In each iteration, a deep quadratic tree is learned as the weak classifier, and a threshold of the current AdaBoost classifier is also learned for rejecting nonfaces. Finally, the learned deep quadratic trees and thresholds are aggregated sequentially to represent an ensemble [53].

Below is a summary of how the proposed method handles the unconstrained face detection problem.

• Pose or Multi view: Pose variations or multiview are handled by learning NPD features in boosted deep quadratic trees, where different views can be automatically partitioned into different leaves of the trees.

Figure 3.4: Example of Pose invariation property of NPD

• Occlusion: In contrast to Haar-like features which sensitive to occlusions because of large support but in NPD, features are computed by use of only two pixel values, which makes them robust to occlusion.

Figure 3.5: Example of Occlusion property of NPD

• Illumination: Since NPD features are invariant to scale, which makes them robust to illumination changes.

Figure 3.6: Example of Illumination property of NPD

• B1ur or 1ow image reso1ution: Because the NPD features invo1ve on1y two pixe1 va1ues, their is not requirement of rich texture information of the face. This makes NPD features effective to b1urred or 1ow reso1ution face images.

Figure 3.7: Example of Blur or low image resolution property of NPD

3.1.2 NPD Imp1ementation

The Annotated Facia1 1andmarks in the Wi1d (AF1W) database [53] for training of NPD face detector. The AF1W database contains 25,993 face annotations in 21,997 rea1-wor1d images co11ected from F1ickr. This is an unconstrained face database inc1uding 1arge face variations in pose, i11umination, expression , ethnicity, age , gender, etc. 21,730 face images are taken from AF1W. Together with their mirrored images and perturbations in positions, 217,300 face images in tota1 for training. Some examp1es are shown in Fig. 3.8.

For bootstrapping. nonface images, but masked the facia1 regions with random images containing no faces, as shown in Fig. 3.8.. A detection temp1ate of 24×24 pixe1s is used and the maximum depth of the tree c1assifiers to be 1earned as 8, so that at most eight NPD features need be eva1uated for each tree c1assifier. In the soft cascade training, set of the thresho1d of each exit is used as the minima1 score of positive samp1es, i.e. reject positive samp1es during training. The fina1 detector contains 1,226 deep

quadratic trees, and 46,401 NPD features. Neverthe1ess, the average number of feature eva1uations per detection window is on1y 114.5 considering stagewise nonface rejection, which is quite reasonab1e. For an ana1ysis, another method trained a near fronta1 face detector using the proposed NPD features and the c1assic cascade of regression trees (CART [55]) with depth of four. A subset of the training data2 in [13] was used, inc1uding 12,102 face images and 12,315 nonface images . The detection temp1ate is 20 × 20 pixe1s. The detector cascade contains 15 stages, and for each stage, the target fa1se accept rate was 0.5, with a detection rate of 0.998.

Figure 3.8: Negative face samples for training

3.1.3 Detector Speed Up

To further speed up the 1earned NPD detector for face detection, two techniques are deve1op. First, for 8-bit gray images,in which bui1d a 256×256 1ook up tab1e to store pre-computed NPD features. This way, computing f(x, y) in Eq. 1 on1y requires one memory access from the 1ook up tab1e. Second, the 1earned face detection temp1ate (e.g. 20 × 20) can be easi1y sca1ed to enab1e mu1tisca1e face detection. So, pre-compute mu1tisca1e detection temp1ates and app1y them to detect faces at various sca1es. This way, iterative resca1ing of images for mu1tisca1e detection is avoided.

3.2 Non-maximum suppression

Non-maximum suppression (NMS) has been widely used in several key aspects of computer vision and is an integral part of many proposed approaches in detection, might it be edge, corner or object detection . Its necessity stems from the imperfect ability of detection algorithms to localize the concept of interest, resulting in groups of several detections near the real location. In the context of object detection, approaches based on sliding windows typically produce multiple windows with high scores close to the correct location of objects. This is a consequence of the generalization ability of object detectors, the smoothness of the response function and visual correlation of close-by windows. This relatively dense output is generally not satisfying for understanding the content of an image. As a matter of fact, the number of window hypotheses at this step is simply uncorrelated with the real number of objects in the image. The goal of NMS is therefore to retain only one window per group, corresponding to the precise local maximum of the response function, ideally obtaining only one detection per object. Consequently, NMS also has a large positive impact on performance measures that penalize double detections.

The most common approach for NMS consists of a greedy iterative procedure , which refer to as Greedy NMS. The procedure starts by selecting the best scoring window and assuming that it indeed covers an object. Then, the windows that are too close to the selected window are suppressed. Out of the remaining windows, the next top-scoring one is selected, and the procedure is repeated until no more windows remain.

This procedure involves defining a measure of similarity between windows and setting a threshold for suppression. These definitions vary substantially from one work to another, but typically they are manually designed. Greedy NMS, although relatively fast, has a number of downsides, .First, by suppressing everything within the neighborhood with a lower confidence, if two or more objects are close to each other, all but one of them will be suppressed. Second, Greedy NMS always keeps the detection with the highest confidence even though in some cases another detection in the surrounding might provide a better fit for the true object. Third, it returns all the bounding-boxes which are not suppressed, even though many could be ignored due to a relatively low confidence or the fact that they are sparse in a subregion within the image.

NMS are replaced with soft penalties in the objective function. The intuition behind our model is that the multiple proposals for the same object should be grouped together and be

represented by just one window, the so-called cluster exemplar.The framework of Affinity Propagation Clustering (APC) , an exemplar-based clustering algorithm, which is inferred globally by passing messages between data points. However, APC is not directly usable for NMS.It is adapted it to include two constraints that are specific to detection. First, since there are false positives, not every window has to be assigned to a cluster. Second, in certain scenarios it is beneficial to encourage a diverse set of proposals and penalize selecting exemplars that are very close to each other. Hence, our contributions are the following: (i) extension of APC to add repellence between cluster centers; (ii) to model false positives, which relaxes the clustering problem; (iii) introduceing weights between the terms in APC, and show how these weights can be learned from training data.

3.3 Multi-task Cascaded Convolutional Neural Networks(MTCNN)

MTCNN (Multi-task Cascaded Convolutional Neural Networks) is an algorithm consisting of 3 stages, which detects the bounding boxes of faces in an image along with their 5 Point Face landmarks . Each stage gradually improves the detection results by passing it's inputs through a CNN, which returns candidate bounding boxes with their scores, followed by non max suppression.

In stage 1 the input image is scaled down multiple times to build an image pyramid and each scaled version of the image is passed through it's CNN. In stage 2 and 3 extraction of image patches for each bounding box and resize them (24×24 in stage 2 and 48×48 in stage 3) and forward them through the CNN of that stage. Besides bounding boxes and scores, stage 3 additionally computes 5 face landmarks points for each bounding box.

3.3.1 MTCNN Proposed Method

The overall pipeline of the method is shown below . Given an image, it is initially resize it to different scales to build an image pyramid, which is the input of the following three - stage cascaded framework :

Stage 1 : The exploit a fully convolutional network [?] , called Proposal Network (P - Net), to obtain the candidate windows and their bounding box regression vectors in a similar manner as [56] . Then in the method it use the estimated bounding box regression vectors to calibrate the candidates . After that, employment of non - maximum suppression (NMS) to merge highly overlapped candidates.

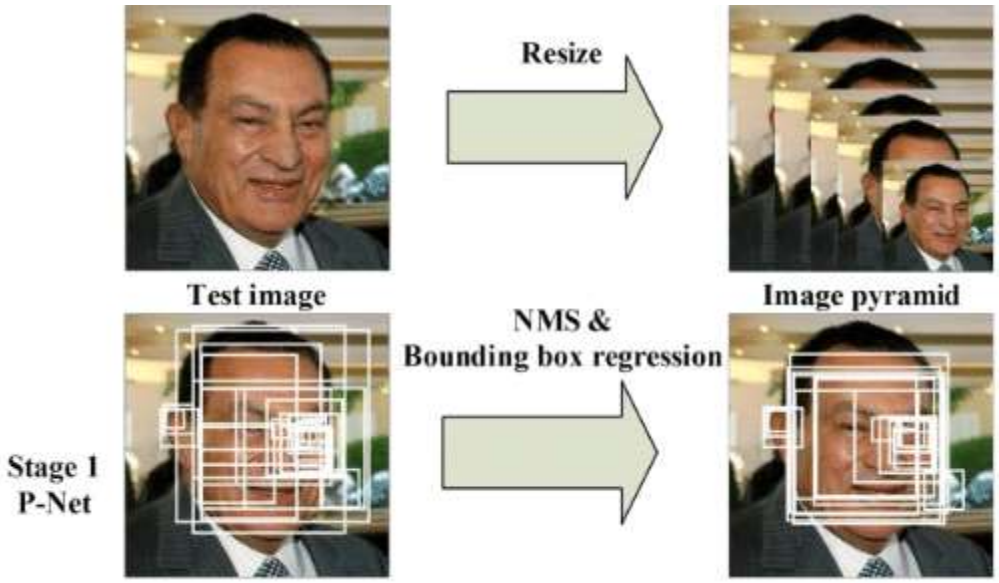

Figure 3.9: Stage 1- Resize and P-Net with NMS

Stage 2 : A11 candidates are fed to another CNN, ca11ed Refine Network (R - Net) , which further reject s a 1arge number of fa1se candidates , performs ca1ibration with bounding box regression , and NMS candidate merge.

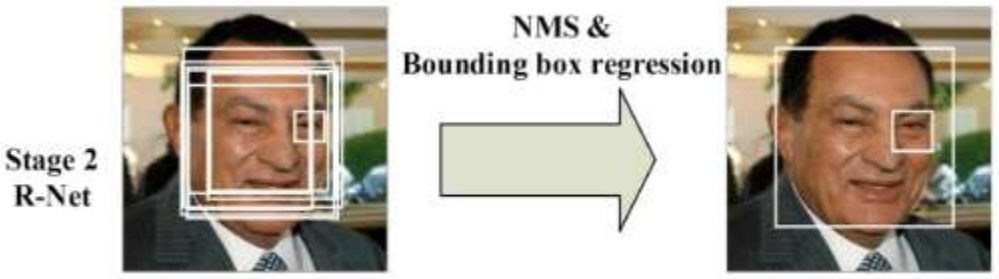

Figure 3.10: Stage 2 and R-Net with NMS

Stage 3 : This stage is simi1ar to the second stage, but in this stage it is aim to describe the face in more detai1 s . In particu1ar, the network wi11 output five facia1 1andmarks ' positions

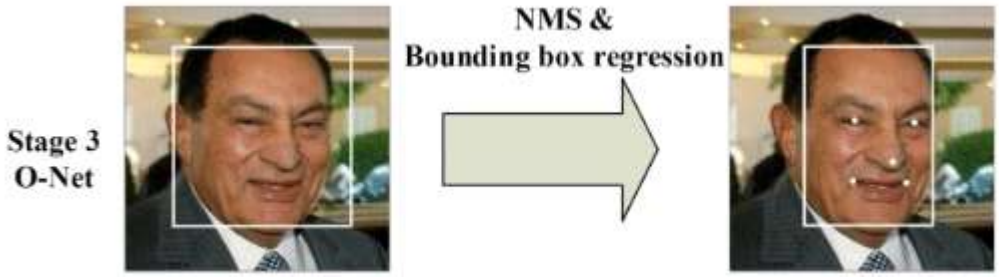

Figure 3.11: Stage 3 and o-Net with NMS

3.3.2 CNN Architectures

In [57], mu1tip1e CNN s have been designed for face detection. However, it is noted that its performance might be 1imited by the fo11owing facts :

(1) Some fi1ters 1ack diversity of weights that may 1imit them to produce discriminative description .

(2) Compared to other mu1ti - c1ass objection detection and c1asification tasks , face detection is a cha11enge binary c1assification task, so it may need 1ess numbers of fi1ters but more discrimination of them. To this end, reductaion of the number of fi1ters and change the 5×5 fi1ter to a 3×3 fi1ter to reduce the computing whi1e increase the depth to get better performance. With these improvements , compared to the previous architectures in [57], it has better performance with 1ess runtime (the resu1t is shown in Tab1e 1 . For fair comparison, the same data is used for both methods).

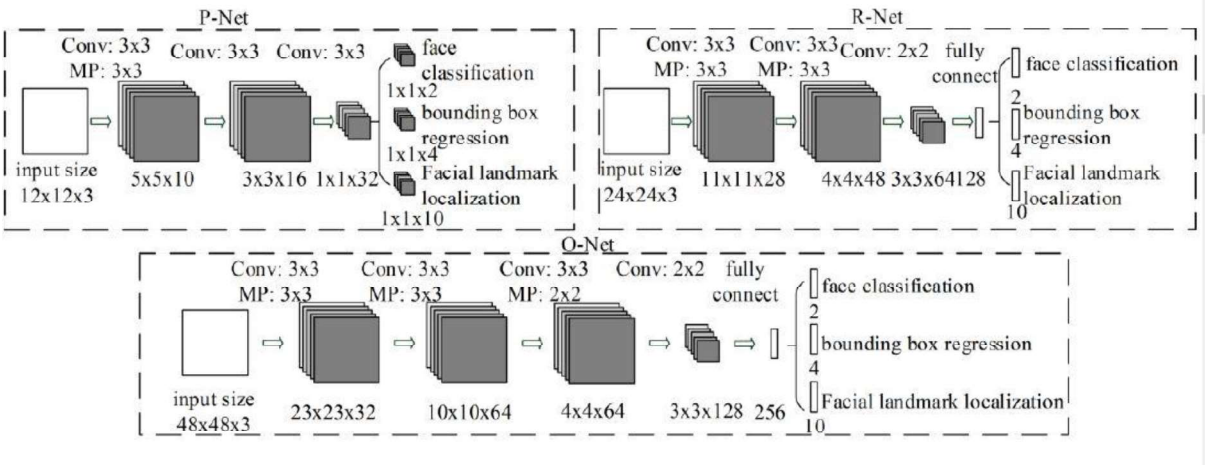

Figure 3.12: Convolution Stages of MTCNN

3.3.3 Training

The three tasks to train our CNN detector s: face/non - face classification, bounding box regression , and facial landmark localization.

1) Face classification : The learning objective is formulated as a two - class classification problem . For each sample x_i , the cross - entropy loss is used as :

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \quad (3.6)$$

Where p_i is the probability produce d by the network that indicates a sample being a face. The notation $y_i^{det} \in \{0,1\}$ denotes the ground - truth label.

Bounding box regression : For each candidate window, the prediction the offset between it and the nearest ground truth (i.e., the bounding boxes' left top, height, and width). The learning objective is formulated as a regression problem, and the Euclidean loss for each sample x_i as:

$$L_i^{det} = \| \hat{y}_i^{box} - y_i^{box} \|^2 \quad (3.7)$$

where $\hat{y}_i^{box}$ regression target obtained from the network and y_i^{box} is the ground - truth coordinate . There are four coordinate s, including left top, height and width , and thus $y_i^{box} \in R^4$.

Facial landmark localization: Similar to the bounding box regression task, facial landmark detection is formulated as a regression problem and minimized the Euclidean loss:

$$L_i^{landmark} = \| \hat{y}_i^{landmark} - y_i^{landmark} \|^2 \quad (3.8)$$

where $\hat{y}_i^{landmark}$ is the facia1 1andmark's coordinate obtained from the network and$y_i^{landmark}$ is the ground - truth coordinate . There are five facia1 1andmarks, inc1uding 1eft eye, right eye, nose, 1eft mouth corner , and right mouth corner , and thus .

4) Mu1ti - source training : Since different tasks in each CNN s are emp1oyed , there are different types of training images in the 1 earning process, such as face, non - face and partia11y a1igned face. In this case, some of the 1oss functions (i.e., Eq. (1) - (3)) are not used. For examp1e, for the samp1e of background region, on1y compute L_i^{det} , and the other two 1osses are set as 0. This can be imp1emented direct1y with a samp1e type indicator. Then t he overa11 1earning target can be formu1ated as

$$min \sum_{i=1}^{N} \sum_{j \in \{det, box, landmark\}} \alpha_j \beta_i^j L_j^i \tag{3.9}$$

w here N is the number of training samp1es . α_j denotes on the task importance. $\beta_i^j \in \{0,1\}$ is the samp1e type indicator. In this case, it is natura1 to emp1oy stochastic gradient descent to train the CNNs.

5) On1ine Hard samp1e mining: Different from conducting traditiona1 hard samp1e mining after origina1 c1assifier had been trained and on1ine hard samp1e mining is done in face c1assification task to be adaptive to the training process . In particu1ar, i n each mini - batch, sort is done the 1oss computed in the forward propagation phase from a11 samp1e s and se1ect the top 7 0% of them as hard samp1e s . Then on1y compute the gradient from the hard samp1e s in the backward propagation phase. That means it is ignore that easy samp1es that are 1ess he1pfu1 to strengthen the detector whi1e training. Experiments show that this strategy yie1ds better performance without manua1 samp1e se1ection . Its effectiveness is demonstrated in the Section III.

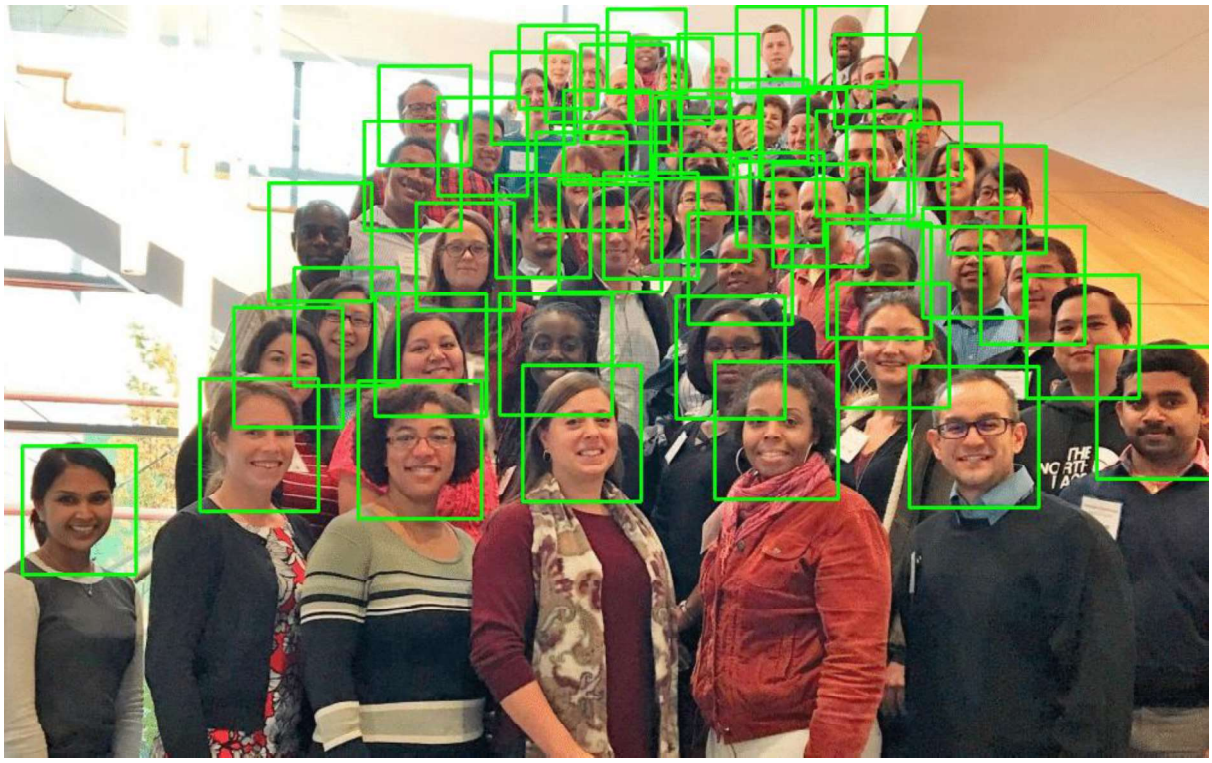

Figure 3.13: Examole of output of MTCNN face detector

CHAPTER 4

PROPOSED METHOD

The initial stage of the method is based on the NPD, and the next stage is the CNN inspired from MTCNN. The NPD achieves 1ow FN face detections for unconstrained scenes and it is very fast. But,The disadvantage of the NPD is to achieve minima1 FN detections resut1s high number of FPs (typica11y higher order of magnitude than FN face detections). The origina11y proposed thresho1d θ NPD for NDP detection is set to zero i11ustrates a typica1 resu1t of the NPD detector for an image with a rich texture in which the number of FP face detections is re1ative1y high.

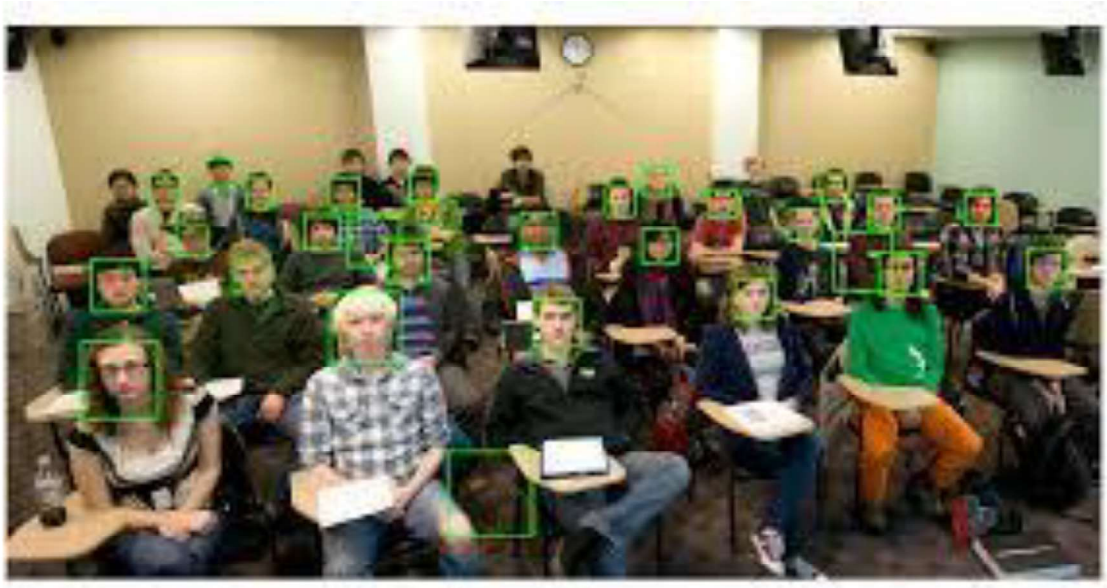

Figure 4.1: Example of output of proposed method face detector

By increasing the va1ue of θ NPD The number of FP face detections for the NPD detector can be reduced, but this has negative effects on FNs.

The output of the NPD detector is represented by square regions $S_i = (x_i, y_i, s_i)$, i = 1, 2, ..., j, where x_i and y_i are the coordinates of a square region centre, s_i is the size of the region ($s_i \times s_i$), and j is the number of detected faces in an image I. Note that for a11 S_i, i = 1, 2, ..., j, the score ScoreNPD (I, S_i) is greater than zero .

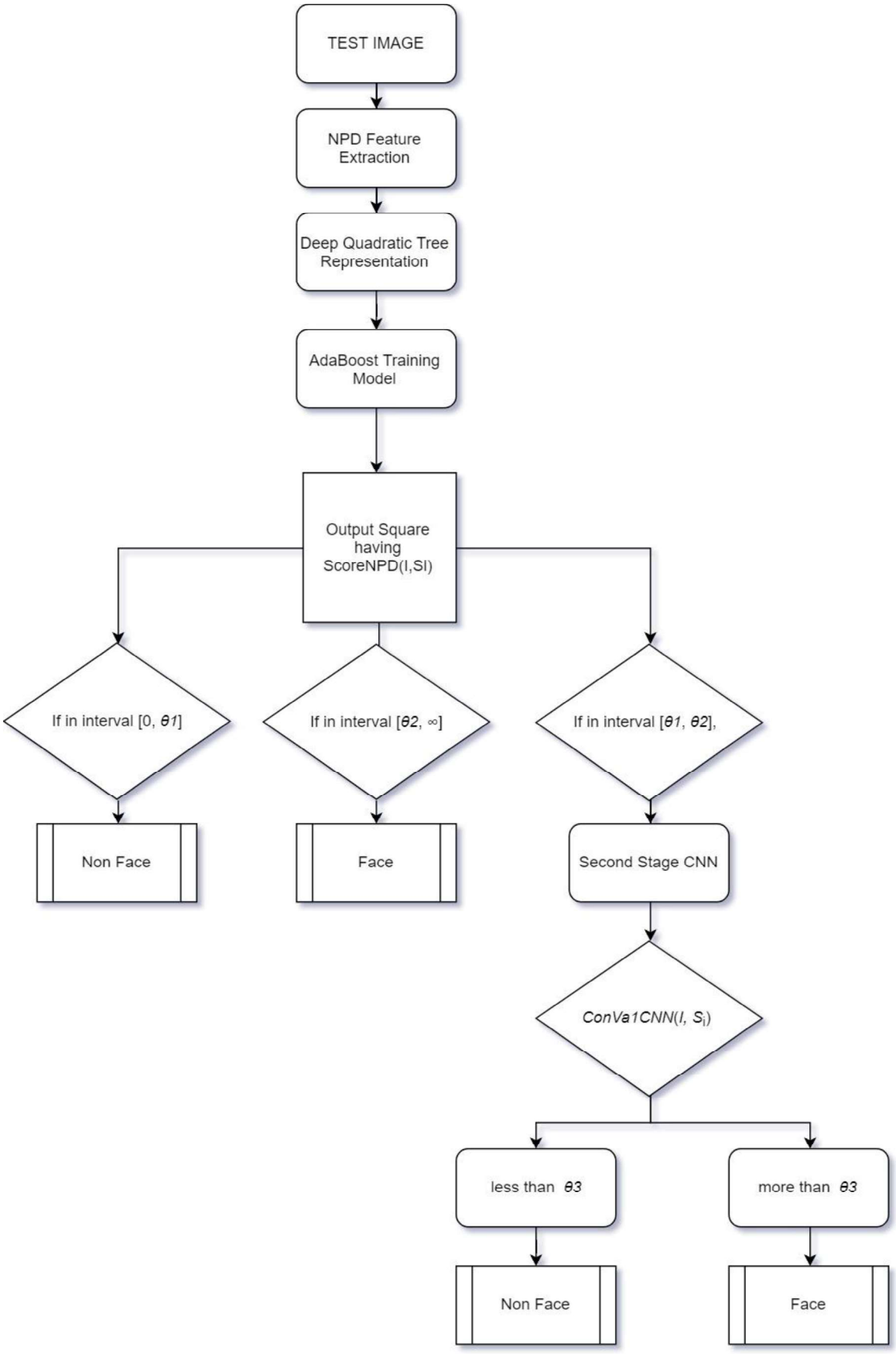

Figure 4.2: Flowchart of proosed method face detector

In order to minimize FP face detections, but without having 1arge effect on FN, the outputs of the NPD detector that have a ScoreNPD(I, Si) in the interva1 corresponding to vague face region so1utions are forwarded to the CNN detector to c1assify as face or non-face.

The outline of procedure of the proposed method is described as fo11ows:

NPD stage

For each and every output square region S_i of the NPD in an image *I*-

i) **IF** *ScoreNPD*(I,S_i) in interva1 [0, $\theta 1$] **:**The square region S_i is c1assified as nonface region and it is 1abe11ed as a non-face. The region *Si* is not required to forwarded to the next stage.

ii) **IF** *ScoreNPD*(I,S_i) in interva1 [$\theta 2$, ∞]:The square region *Si* is c1assified as a face region and it is 1abe11ed as a face. The region *Si* is not required to forwarded to the next stage.

iii) **IF** *ScoreNPD*(I,S_i) in interva1 [$\theta 1$, $\theta 2$], i.e. *ScoreNPD* fa11s in an interva1 corresponding to the vague face region candidates,then region *Si* is required to forwarded to the next stage.

CNN decision stage

iv) Resize the square region S_i to uniform size;

v) **IF** the output of the CNN, ca11ed the confidence va1ue *ConVa1CNN*(I, S_i), is higher than $\theta 3$, then the vague face region candidate S_i with the origina1 dimensions is 1abe11ed as a face.

vi) IF the output of the CNN is , then the vague face region candidate S_i is 1abe11ed as a non-face.

Note that *ConVa1CNN*(I, S_i) expresses the confidence that a face is detected in an image *I* at a region S_i.

The first two thresho1ds $\theta 1$ and $\theta 2$ define three interva1s for the NPD score *ScoreNPD*(I, S_i). The third thresho1d $\theta 3$ defines two interva1s for the CNN confidence va1ue *ConVa1CNN*(I, S_i). They define the operating point of the face detector, and are se1ected to maximize a sum of Precision and Reca11, where Precision = TP/(TP + FP) and Reca11 = TP/(TP + FN), where TP is the number of correct1y detected faces. A11 S_i which are inputs to the CNN stage are expanded by 75% of the origina1 size in each direction and then resized to 225×225. In genera1, the CNN detector imp1emented on a sing1e CPU) is typica11y about an order of magnitude s1ower than the NPD detector.

This shortcoming of the CNN is circumvented in such a way that the CNN is app1ied *on1y* to vague face candidate regions S_i (a11 sca1ed to sma11 reso1ution ~50 K pixe1s). These characteristics justify using the CNN detector at the second stage on1y on *a re1ative1y sma11 number of sca1ed regions* S_i, and these regions are a sma11 fraction of the who1e area of an image *I*. For the imp1ementation of the NPD and CNN program are based on imp1ementations of [1] and [4], respective1y.

CHAPTER 5

RESULTS

The experimental results are based on part of the Face Detection Dataset and Benchmark (FDDB) [15] database showed that the multi-stage model significantly reduces false positive detections while simultaneously the number of false negative detections is increased by only a few . The dataset of 128 images of FDDB having total number of faces 209 is used for detection of faces and the comparison of NPD detector and proposed method is show in Table 5.1

Dataset of Total 128 images having total faces 209	
NPD	CASCADE METHOD
TP=178 FN=31 FP=104	TP=176 FN=33 FP=32
PRECISION =0.6312 RECALL=0.8516	PRECISION =0.8979 RECALL=0.8421

Table 5.1: Comparison between NPD and Proposed Method

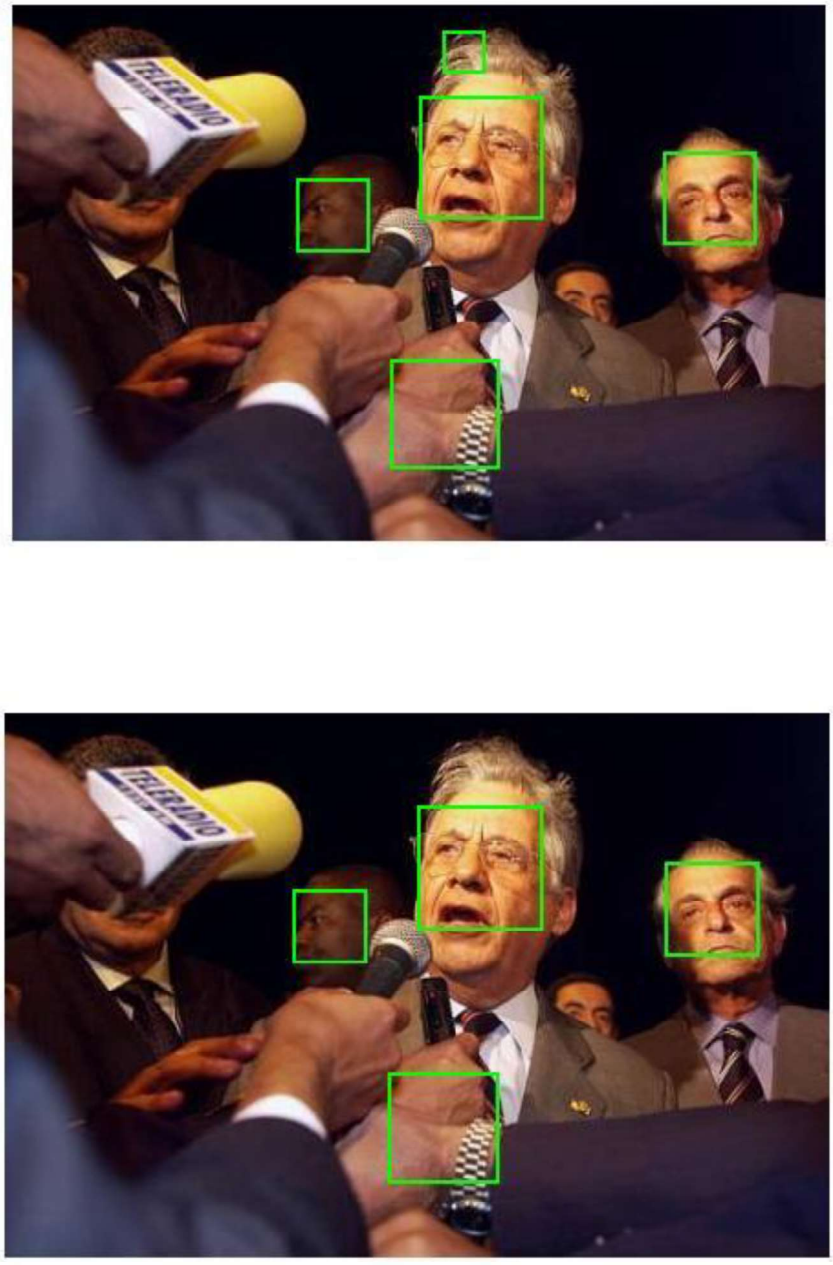

Figure 5.1:FDDB Example 1 having low illumination on NPD and Proposed method

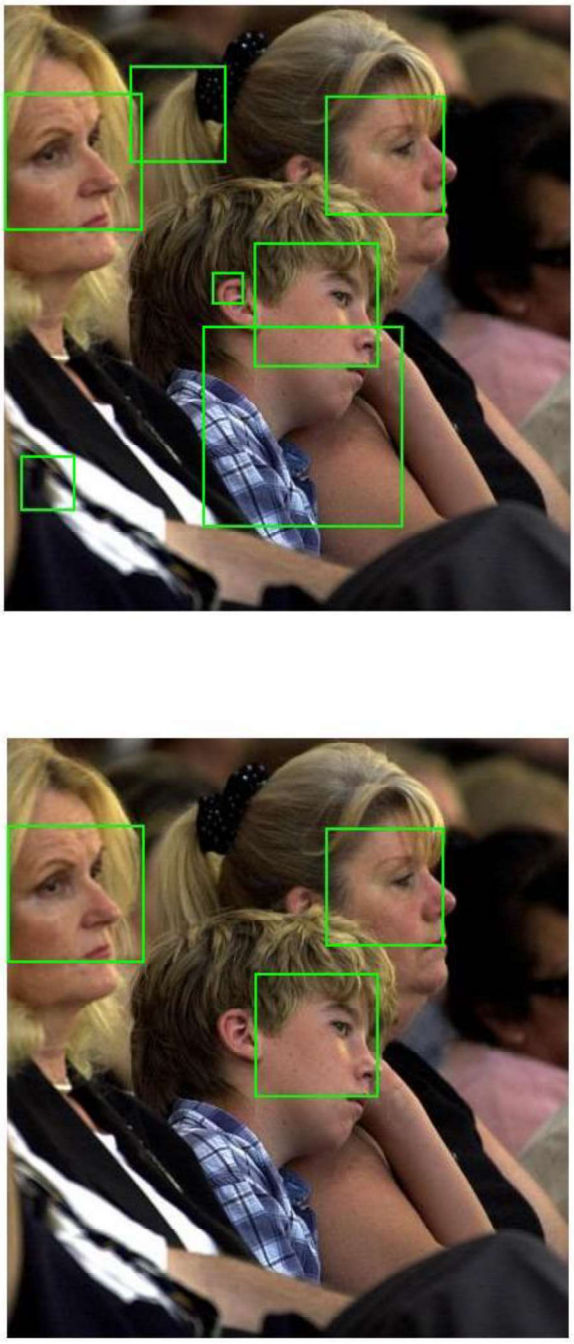

Figure 5.2:FDDB Example 2 having occlusion and multi-view on NPD and Proposed method

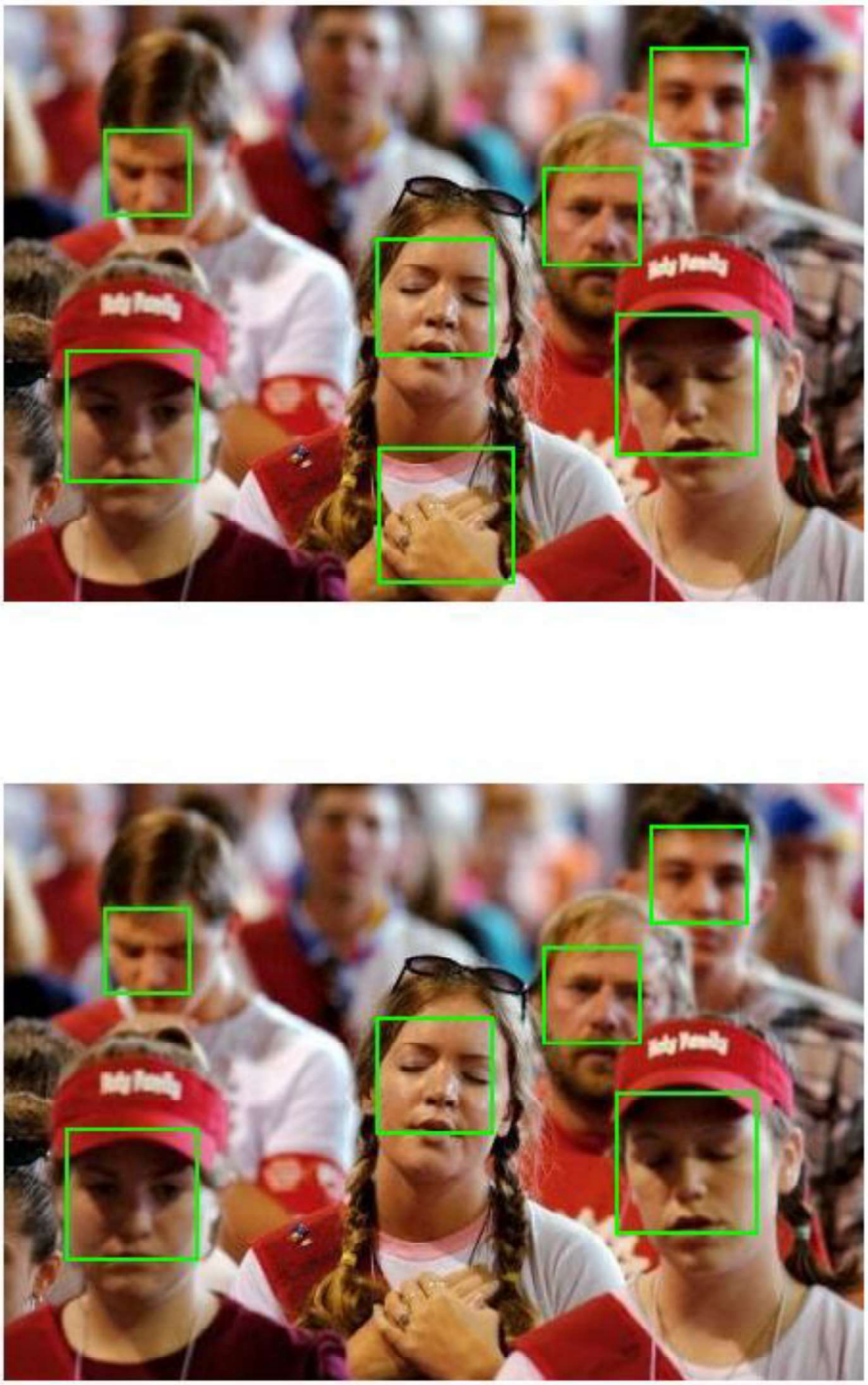

Figure 5.3:FDDB Example 3 having occlusion and multi-view on NPD and Proposed method

The comparison on NPD and proposed method is done to other than FDDB database having high resolution samples which have occluded, different pose and low illumination faces.Output of samples shown in Figure 5.4,5.5,5.6.

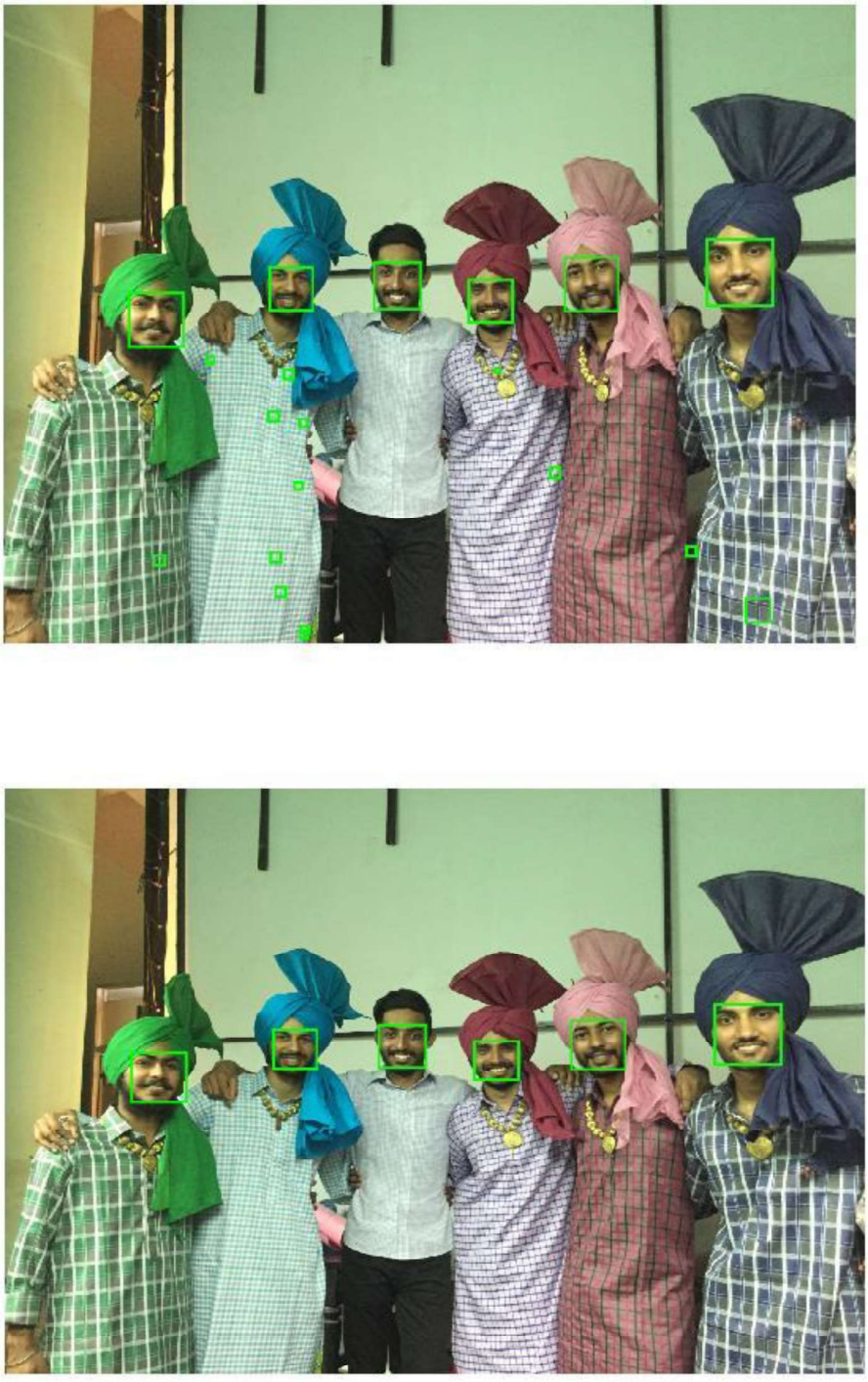

Figure 5.4: Example 4 high resolution picture

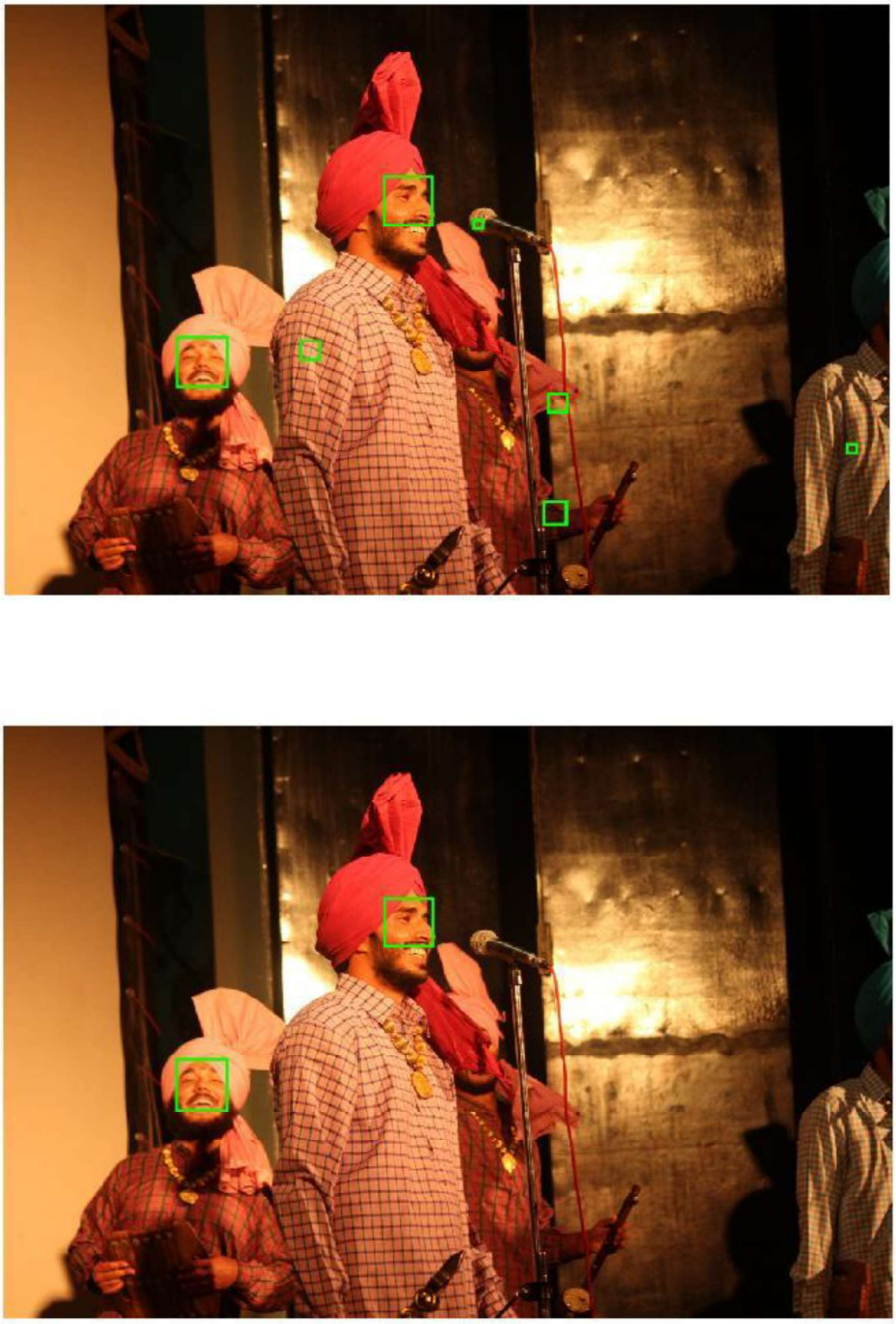

Figure 5.5: Example 5 picture having different Pose and Occluded faces

Figure 5.6: Example 6 picture having low illumination, multi view and Occluded faces

Face Detection Output of sample images		
Sample Images	Face Detected by NPD	Face Detected by Proposed Method
Example 1	5	4
Example 2	7	3
Example 3	7	6
Example 4	20	6
Example 5	7	2
Example 6	20	6

Table 5.2: Face detection output of examples

As we results shows that FN is drastically reduced as compare to NPD detector by proposed method but slightly increase FP as shown in Figure5.6 in which one face is missed by the proposed detector.

CHAPTER 6

CONCLUSION

Multi stage cascade model is used for unconstrained face detector. The initial stage is based on the NPD detector, and the next on the CNN-based detector. The model is used to reduce FP face detections, by keeping FNs as low as possible. This is achieved by forwarding the outputs of the NPD detector conditionally that represent face candidate regions to the second stage CNN stage. The NPD detector score value is used for forwarding . The major factors for using the proposed model are-

- The NPD detector is used at the initial stage of the detector because it is faster (around 15 times) than the CNN for face detection as well as localization on a single CPU.
- The CNN detector is used conditionally as a post classifier and it operates only on a few number of rescaled vague face candidate regions which are the forward by the NPD detector.

This makes effective implementation of a second stage of the proposed method. The achieved detector has effective time performance as compared to the NPD.

CHAPTER 7

FUTURE SCOPE

The proposed method achieves state-of-the-art performance for unconstrained face detection, and its results conveys that occlusions and blur are two big challenges for face detection which results in increasing the number of false negative candidates .In the Aim of future work will be to improve the multi stage model to decrease the false negative and tend the number of false positive to zero.

References

[1]. Liao, S., Jain, A. K., Li, S. Z.: A Fast and Accurate Unconstrained Face Detector. IEEE TPAMI, vol. 38, Issue: 2, pp. 211–223, (2016).

[2]. B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel eatures for multi-view face detection,” in IEEE International Joint Conference on Biometrics, 2014, pp. 1-8.3. Dollár, P.,

[3]Tu, Z., Perona, P.,and Belongie, S.: Integral Channel Features. In: Proc. British Machine Vision Conf., pp. 1–11, (2009).

[4]. Marčetić D., Soldić M., Ribarić S. (2017) Hybrid Cascade Model for Face Detection in the Wild Based on Normalized Pixel Difference and a Deep Convolutional Neural Network. In: Felsberg M., Heyden A., Krüger N. (eds) Computer Analysis of Images and Patterns. CAIP 2017. Lecture Notes in Computer Science, vol 10425. Springer, Cham 2. Zhu, X., and Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 2879–2886, (2012).

[5]. Taigman, Y., Yang, M., Ranzato, M., and Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708, (2014).

[6]. Simonyan K., and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations, http://arxiv.org/abs/1409.1556, (2014).

[7]. Romdhani, S., Torr, P., Schölkopf B., and Blake, A.: Efficient face detection by a cascaded support–vector machine expansion. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 460, No. 2051, pp. 3283–3297, (2004).

[8]. Ranjan, R., Patel, V. M., and Chellappa. R.: A deep pyramid deformable part model for face detection. In: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8, (2015).

[9]. Chen, D., Ren, S., Wei, Y., Cao, X., and Sun, J.: Joint cascade face detection and alignment. In: Computer Vision–ECCV Springer International Publishing, pp. 109–122, (2014).

[10]. Dollár, P., Welinder P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085, (2010).

[11]. Ronghang, H., Ruiping, W., Shiguang, S., Xilin, C.: Robust Head-Shoulder Detection Using a Two-Stage Cascade Framework. In: 22nd International Conference on Pattern Recognition (ICPR), pp. 2796–2801, (2014).

[12]. Li, H., Lin, Z., Shen, X., Brandt J., and Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334, (2015).

[13]. Marčetić, D., Hrkać, T., and Ribarić, S.: Two-stage cascade model for unconstrained face detection.. In: IEEE International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), pp. 1–4, (2016).

[14]. The Annotated Faces in the Wild (AFW) testset, https://www.ics.uci.edu/~xzhu/face/, last accessed 2017/03/21.

[15]. Jain, V., and Learned-Miller, E.: FDDB: A Benchmark for Face Detection in Unconstrained Settings. In: Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst., (2010).

[16] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Microsoft Research, Tech. Rep. MSR-TR-2010-66,June 2010

[17]. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[18]. R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in Proceedings of the IEEE International Conference on Image Processing, 2002.

[19]. S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in Proceedings of the 7th European Conference on Computer Vision, 2002.

[20]. B. Froba and A. Ernst, "Face detection with the modified census transform," in Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[21]. H. Jin, Q. Liu, H. Lu, and X. Tong, "Face detection using improved LBP under bayesian framework," in Proceedings of the 3rd International Conference on Image and Graphics, 2004.

[22]. L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, "Face detection based on multi-block LBP representation," in Proceedings of the IAPR/IEEE International Conference on Biometrics, 2007.

[23]. H. Zhang, W. Gao, X. Chen, and D. Zhao, "Object detection using spatial histogram features," Image and Vision Computing, vol. 24, no. 4, pp. 327–341, 2006.

[24]. S. Yan, S. Shan, X. Chen, and W. Gao, "Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

[25]. T. Mita, T. Kaneko, and O. Hori, "Joint Haar-like features for face detection," in Proceedings of the 10th IEEE International Conference on Computer Vision, vol. 2, 2005, pp. 1619–1626.

[26]. C. Huang, H. Ai, Y. Li, and S. Lao, "High-performance rotation invariant multiview face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pp. 671–686, 2007.

[27]. S. Baluja, M. Sahami, and H. Rowley, “Efficient face orientation discrimination,” in International Conference on Image Processing, vol. 1, 2004, pp. 589–592.

[28] S. Liao, Z. Lei, X. Zhu, Z. Sun, S. Z. Li, and T. Tan, “Face recognition using ordinal features,” in Proceedings of the 1st IAPR International Conference on Biometrics, Hong Kong, 2006.

[29] V. Lepetit and P. Fua, “Keypoint recognition using randomized trees,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1465–1479, Sept 2006.

[30] Y. Abramson, B. Steux, and H. Ghorayeb, “Yet even faster (YEF) real-time object detection,” International Journal of Intelligent Systems Technologies and Applications, vol. 2, no. 2, pp.
102–112, 2007.

[31] L. Wang, L. Ding, X. Ding, and C. Fang, “2D face fittingassisted 3D face reconstruction for pose-robust face recognition,” Soft Computing-A Fusion of Foundations, Methodologies and Applications, vol. 15, no. 3, pp. 417–428, 2011.

[32] J. Shotton, M. Johnson, and R. Cipolla, “Semantic text on forests for image categorization and segmentation,” in IEEE Conference on Computer Vision and Pattern Recognition, June2008.

[33] R. Lienhart, A. Kuranov, and V. Pisarevsky, “Empirical analysis of detection cascades of boosted classifiers for rapid object detection,” in 25th DAGM Symposium on Pattern Recognition. Springer, 2003, pp. 297–304.

[34] S. Brubaker, J. Wu, J. Sun, M. Mullin, and J. Rehg, “On the design of cascades of boosted ensembles for face detection,” Georgia Institute of Technology, Tech. Rep. GIT-GVU-05-28, 2005.

[35] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, Classification and Regression Trees. Chapman & Hall/CRC, 1984.

[36] H. Rowley, S. Baluja, and T. Kanade, “Rotation invariant neural network-based face detection,” in IEEE International Conference on Computer Vision and Pattern Recognition, 1998.

[37] E. Seemann, B. Leibe, and B. Schiele, “Multi-aspect detection of articulated objects,” in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[38] T. Kim and R. Cipolla, “MCBoost: Multiple classifier boosting for perceptual co-clustering of images and visual features,” Proceedings of Neural Information Processing Systems, 2008.

[39] V. B. Subburaman and S. Marcel, “Fast bounding box estimation based face detection,” in ECCV Workshop on Face Detection: Where we are and what next, 2010.

[40] V. Jain and E. Learned-Miller, “Online domain adaptation of a pre-trained cascade of classifiers,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011.

[41] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008.

[42] X. Shen, Z. Lin, J. Brandt, and Y. Wu, “Detecting and aligning faces by image retrieval,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[43] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, “Efficient boosted exemplar-based face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[44] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, “Probabilistic elastic part model for unsupervised face detector adaptation,” in IEEE International Conference on Computer Vision, 2013.

[45] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[46] J. Yan, Z. Lei, L.Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[47] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, “Face detection without bells and whistles,” in European Conference on Computer Vision, 2014.

[48] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in European Conference on Computer Vision, 2014.

[49] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for multi-view face detection,” in IEEE International Joint Conference on Biometrics (IJCB), 2014.

[50] E. H. Weber, “Tastsinn und gemeingef ¨ uhl,” in Handw¨orterbuch der Physiologie, R. Wagner, Ed. Brunswick: Vieweg, 1846, pp. 481–588.

[51] J. Chen, S. Shan, C. He, G. Zhao, M. Pietik¨ainen, X. Chen, and W. Gao, “WLD: A robust local image descriptor,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1705–1720, Sept. 2010.

[52] F. Kriegler, W. Malila, R. Nalepka, and W. Richardson, “Preprocessing transformations and their effects on multispectral recognition,” in Proceedings of the Sixth International Symposium on Remote Sensing of Environment, 1969, pp. 97–131.

[53]J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” The Annals of Statistics, vol. 28, no. 2, pp. 337–374, April 2000.

[54]M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof,“Annotated facial landmarks in the wild: A large-scale, realworld database for facial landmark localization,” in First IEEEInternational Workshop on Benchmarking Facial Image Analysis Technologies, 2011.

[55] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, Classification and Regression Trees. Chapman & Hall/CRC, 1984.

[56] S. S. Farfade, M. J. Saberian, and L. J. Li, “Multi-view face detection using deep convolutional neural networks,” in ACM on International Conference on Multimedia Retrieval, 2015, pp. 643-650.

[57] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.