Detect Saliency in Crowds Using Face Features

By

Waqas Ali

NUST201464100MSEECS61314F

Supervisor

Dr. Anis ur Rahman

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MS-CS)

In

NUST School of Electrical Engineering and Computer Science (SEECS)

National University of Science and Technology (NUST), Islamabad, Pakistan.

(2016)

Certificate

Certified that the contents of thesis document titled "Detect Saliency in Crowds using face Features" submitted by Mr. Waqas Ali have been found satisfactory for the requirement of degree.

Advisor: _____

Dr. Anis ur Rahman

Committee Member1: _____

Dr. Asad Anwar Butt

Committee Member2: _____

Dr. Mian Muhammad Hamayun

Committee Member3: _____

Dr. Muhammad Imran Malik

Abstract

This thesis addresses the problem of saliency detection in crowded scenes. Crowded scenes are those which have irregular scene density. Crowded scenes are more frequent in real world with applications in public management, security, population monitoring and urban planning. However, the task is highly challenging due to the large individual competing for attention. While regular scene density saliency models have been promising, they are still limited in their capability to detect salient regions in crowded scenes. Recent researches have shown the important of faces in human scenes. Faces are most important body part in human body and psychology studies has shown talking faces are more salient. Deep learning has emerged as a powerful learning mechanism, able to learn higher-level deep features when provided with a relatively large amount of labeled training data. Such networks have shown state-of-the-art object and scene recognition results on the ImageNet dataset. Here, we used crowd features to cluster crowd scenes into different crowd levels. Then, using low-level features combined with different face attributes, a deep fully connected neural network for each crowd level was proposed. We benchmark our results with previous model on crowded scenes, and present superior results on the Crowd dataset.

Certificate of Originality

I hereby declare that the research paper titled "Detect Saliency in Crowds using Face Features" is my own work and to the best of my knowledge. It contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST or any other education institute, except where due acknowledgment, is made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS-NUST or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic is acknowledged. I also verified the originality of contents through plagiarism software.

Signature: _____

Author Name: Waqas Ali

Acknowledgements

I bow before Almighty Allah to express my gratitude for He is the only one who brightens my mind and heart when I feel standing alone in the darkness. He is the only one who helps me when I fall and who is the only hope when I feel broken. I am nothing but His benevolence makes me what I am today. He blessed me even more than I deserve. Thank you Allah!

My words are not enough to pay special credit and appreciation to my supervisor, Dr. Anis ur Rahman for their continued support and encouragement. I wish to convey my sincere thanks to him for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor than him, for my research.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Asad Anwar Butt, Dr Mian Hamayun and Dr. Imran Malik for their insightful comments and encouragement. I offer my sincere appreciation for the learning opportunities provided by my committee.

I place on record, my sincere thank you to my Institution and the Department for providing me a platform where I can learn not only about course books but also about how to excel in every field of life. My Institution makes me feel proud.

I would also like to thank my teachers, my classmates, my friends and my colleagues for being the best human beings I have ever met and to those who forwarded positive criticism to me.

I express my heartfelt thanks to my wife for all she had done for me and to my Siblings. This whole success is due to my siblings. I couldn't have come to this milestone without their help and support, and above all their love and care. Their encouragement when the times got rough is much appreciated and duly noted. It was a great comfort and relief to know that they are with me.

Last but not the least, no words in this world are enough to show gratitude and say thanks to my Parents for supporting me spiritually throughout and for all of the sacrifices they have made for me. Their prayers for me were the only thing that sustained me thus far. They were always there for me to support in the moments when there was no one to answer my queries.

My deepest gratitude!

I truly dedicate my endeavor to my Parents and Wife

For their endless love, support, encouragement and sacrifices for me.

Table of Contents

List of Figures

List of Tables

Chapter 1:

1. INTRODUCTION

The term crowd is referred as group of people gathered or gathering of people of some cause, such as political rally, sports event, public festivals. It is also defined as other group nouns for collection of humans or animals, such as aggregation, audience, concert, group, mass, mob, public etc. According to an opinion researcher Vincent Price crowd and masses are compared as: Crowds are defined by shared emotions and experiences while masses are defined by their interpersonal isolation. In human sociology, the term mobbed means extremely crowded while in terms of animals and birds, many individuals of one species combine together to drive away the larger individual of another species. Social aspect refers to the formation, management and control of crowd on the basis of their point of views. For example in political rallies crowds are controlled to become align to avoid indiscipline or to prevent damage. Political crowds are controlled by law enforcement but for huge and dangerous crowds military forces are used.

Psychological aspects are considered with the crowd psychology. Crowd psychology is referred to the behavior and thought of individual crowd members and as well as whole crowd. Due to safety issues of crowd members, this topic is receiving much attention of agencies and Government. How crowds respond to different situations? A report of Year 2009 explains many behaviors of crowds. It explains that if only a few members of crowd know about some decision then whole crowd becomes united on that decision regarding their directions and speed of movement. The position of informed members of crowd affects the behavior of crowd.

Crowd psychology is also known as mob psychology. Many theories have been developed by social psychologists that explain the ways in which psychology of crowds differs or interact with that of individuals in it. This field relates to the behavior and thoughts of the individuals of the crowd and of whole crowd as well. Crowd behavior is greatly influenced by the loss of

responsibility and behavior of individual of crowd, which increase with the size of crowd. There is not so vast research regarding types of crowd.

Two Scholars Momboisse (1967) and Berlonghi (1995) work on the purpose of formation of crowds to differentiate among them. Momboisse defined four types of crowds: casual, conventional, expressive and aggressive. Berlonghi classified crowd as spectator, demonstrator or escaping to define purpose of gathering. Herbert Bulmer classified crowd in terms of emotional intensity system. He explained four types of crowds: casual, conventional, expressive and acting. This system is of dynamic nature according to which a crowd changes its level of emotional intensity over time and can be classified in any of four types. Crowds can be active or passive. Active crowds are known as mob and passive as audience. There are four types of active crowds: aggressive, escapist, acquisitive orr expressive mobs. Aggressive mobs are violent. Football riots and L.A riots are examples of aggressive crowds. Escapist mobs are created because of panicked people who want to get rid of dangerous situation. When a large number of people start fighting for limited resources, acquisitive mobs are created. Example of such a mob is crowd of Hurricane Katrina who looted after incident in 2005. The mob which is formed by the gathering of people for some active purpose is known as expressive mob. Examples are civil disobedience, concerts, religious events and political matters etc. Due to increase in world population crowd scene are more frequent in real world. It attracts researcher from different field to studies the nature of crowd and solve different problem in real world like crowd analysis, crowd behavior, crowd management etc.

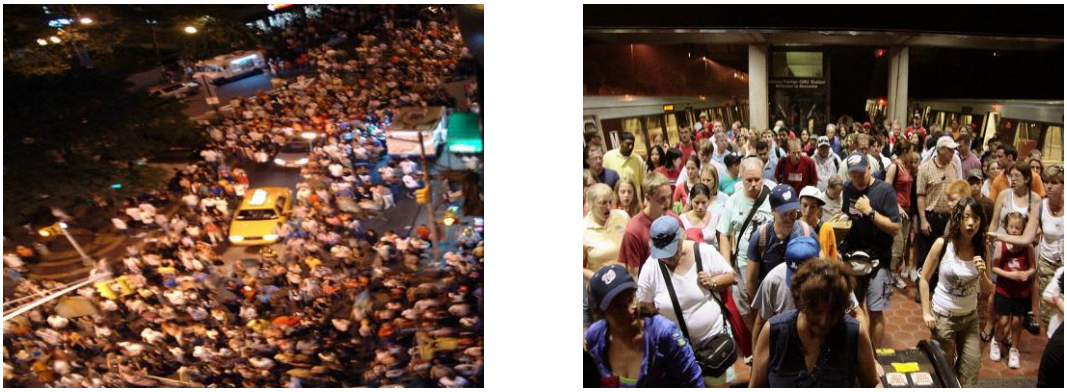

Figure 1.1: Crowd in real world

An animal brain has very limited neurons to process visual information which is passed through eyes. Processing of this information is computation extensive work like detecting objects. The solution of this problem is to recognize small areas and few objects at time in a visual scene. Later on, many objects and areas can be processed in that scene one after the other. To fully process all locations at a time, visual attention may be a solution but this creates a problem. The problem is that it is difficult to select target of attention. Visual salience helps our brains to make a better selection. It provides a mechanism to select the most relevant visual scene while eliminating the data in background. Visual saliency is based on the characteristic of visual pattern such as green line in red lines which attract our attention rapidly. Using these characteristic various computation model have been developed for detection of salient regions. The goal of these saliency models is to predict the most informative region or where people look in the visual scene.

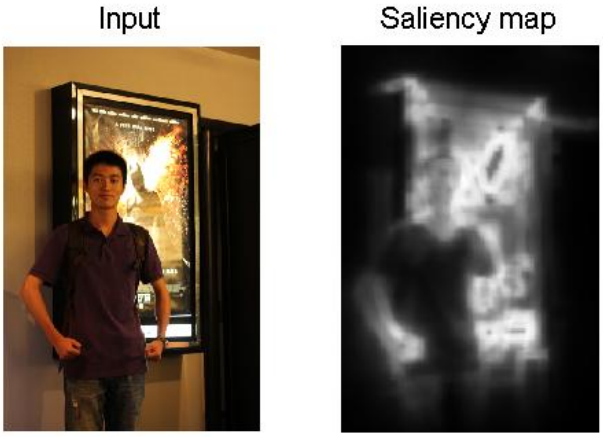

Figure 1.2: Saliency Map

Visual saliency is computed either using bottom-up or top-down cues. In bottom-up visual saliency, context of the scene is used and is computed in a pre-attentive manner. Like in moving objects if motion or direction of an object is different from its surrounding objects it will be perceive as salient. Other bottom-up cues in an image could be color, orientation and intensity. In top-down visual saliency external cues are used and saliency will be task-driven

or task-dependent. Top-down cues are external cue and totally depend on the task. These cues vary from one problem to another like searching in crowds or objects etc.

Figure 1.3: Saliency Maps

1.1 Motivation

Crowded scenes are more frequent in real world with applications in public management, visual surveillance, population monitoring and urban planning. It has attracted many researchers in computer vision to solve many problems. Visual saliency plays an important role in solving these problems. It extracts the salient region in the crowded scenes like for anomaly detection it extracts the region which is different from its surrounding etc. Following are the fields in which visual saliency plays important role in crowds:

Anomaly detection in Crowds: One of the typical applications in crowded scenes is to detect abnormal events. Human has very limited capabilities to process information in a video scene due to which automated anomaly detection is needed to provide a vigilant surveillance. Abnormal events are subjective and context sensitive. For example running on a track is normal whereas running in a crowded and public place will be considered as abnormal event.

Abnormal event or anomaly is detected by either tracking or without tracking. In tracking category speed and direction are used for detection whereas in non-tracking category motion and texture features are used for detection. Every event which is anomaly also will be salient region in the scene.

Human tracking: The number of public security camera is increasing in public area for crowd management and monitoring. One of the application of video surveillance is human/pedestrian tracking. Human operators have limited capability to observe crowds for tracking. Therefore an automated tracking system is required for this purpose. In automated tracking system density of crowd has significant impact. Tracking in high crowd density such as at airports and train station is very difficult because only few parts of the body are available due to occlusion. Similarly in low crowd density full body of each pedestrian is available. Body part detection, body shape and background modeling are some techniques that are used for tracking. These techniques are not well suited to track people in high crowd density in which majority people are in motion and human bodies are partially or fully occluded.

Crowd behavior: Closed Circuit Television (CCTV) is used in surveillance system through which varies objects and their behavior can be supervised. One of the applications of crowd surveillance is to monitor the crowd behavior. Crowd behavior analysis is used in different fields like computer vision social studies, crowd management and crowd simulation etc. In computer vision, crowd behavior analysis has solved many problems like detection of riots or chaotic acts in crowds, Identify crowd type etc.

1.2 Problem Statement

Detection of saliency in crowd scenes is a difficult task. Crowd scene has high density and large number of human complete for attention. To detect salient regions in crowd scenes an effective mechanism is needed that rectify the image to exclude un-salient region. Face is most salient part in human body and can play an important role in saliency detection in crowds.

1.3 Challenges

Crowded scenes are complex, exhibiting both dynamics and psychological characteristics. These characteristics make identification of salient regions in crowds a challenging task because large number of individuals competing for attention. Conventional saliency models are not suitable to predict human fixations in crowded scenes as the models do not consider crowd features.

1.4 Contribution

In this paper, we proposed a simple yet effective model to compute saliency in crowded scenes. A deep fully connected neural network for saliency that uses low-level features (color, intensity, orientation) and mid-level features (faces, crowd density). The proposed model was evaluated using an available dataset comprising static crowded scenes with eye-tracker recordings. Using different criteria, we showed significant improvement against existing visual saliency models for crowds.

The main contributions of this paper are as follows:

- A new method was proposed to compute face density. Face density is further used to compute crowd density.
- Crowd features such as number of faces and crowd density were used to categorize crowds into low, mid and high level. It is natural to think that crowd of different density should receive a different level of attention.
- A deep fully connected neural network was proposed to compute saliency in crowded scenes which use both lower level features and faces attributes and taking into account crowd level.

1.5 Thesis Outline

The thesis is organized as follows. Chapter 2 presents the literature review on crowd analysis, visual saliency, and importance of faces and gives a description of the crowd database that was used during all experiments for saliency detection. Chapter 3 presents the methods used to extract image features, creating saliency map based on which the regions are classified and use a deep fully connected neural network. Experimental evaluation and results are presented in Chapter 4. Thesis concludes with a summary and suggestions for further research in Chapter 5.

Chapter 2

2. LITERATURE REVIEW

2.1 Overview

In this chapter, we have discussed the background of visual saliency and crowds. Literature view is divided into 3 parts: (1) crowds analysis (2) visual saliency (3) visual saliency in crowd. In first part we have provided the literature on crowd analysis. In second part we have discuss visual saliency and in third part we have provided literature on visual saliency in crowds.

2.2 Crowd Analysis

There is an abundant literature for crowd analysis. Different models are developed for anomaly detection, crowd tracking, crowd behavior, crowd counting and motion pattern segmentation. Vijay Mahadevan [1] provided an anomaly detection model based on temporal and spatial features in video sequences. The model first focuses on the video representation so that anomaly detection can be performed on crowded scenes. After that temporal and spatial features are identify using low-probability of dynamic texture and discriminant saliency detector respectively. Louis Kratz [2] proposed anomaly detection model using relationship between dense local motion pattern and spatio-temporal features. Using this local spatio-temporal pattern crowd behavior is identifying which in turn gives unusual events in video sequences. Shandong Wu [3] model consist of three aspects. First particle trajectories are used for crowd modeling which is further used for crowd flow representation. Secondly chaotic invariant features are regulated to form chaotic dynamics in crowd context. After that a probabilistic model is proposed to detect anomaly detection and localization. The model is based on maximum likelihood which identifies the abnormal scene. The localization purpose is to locate the position and size of anomaly. Vikas Reddy [4] model for anomaly detection is based on the segmentation of foreground objects and background dynamics. After that optical flow of only foreground objects are computed as motion feature. Other features such as size and texture are also considered to analyze anomaly detection. Multiple classifiers are used for each feature type and results are combined to detect anomaly. Ke Ma [5] model is based on the

dense trajectory of motion point. These dense trajectories are used to compute social interactive forces and optical flow in crowded scenes. A single class Support vector machine algorithm is developed for the computation purpose.

Aniket Bera [6] presented a human tracking model for dense crowd which is based on particle filters. Particle filters is a well-known algorithm for object tracking. A multi-agent motion model is proposed for the computation which takes into account pedestrians behavior and confidence metric. Saad Ali [7] proposed model for human tracking in high density crowded scenes. In his research three floor fields are computed for tracking an individual, Static floor field identifies the area in scene which has high probability of attractiveness, Dynamic floor field is used for crowd behavior and Boundary floor field specifies the influence of the other objects in scene like obstacle, wall and humans. Using all three floor fields and scene layout the model is developed for human tracking in crowded scenes. Jialue Fan [8] proposed a convolutional neural network to tracking human in crowd scene. In this model temporal and spatial features are extracted from parametric feature pool for specific object. A shift-variant CNN is designed which take into account previous image frame to predict the future position in next image frame. Ran Eshel [9] proposed model overcome the occlusion problem in crowd scene by considering the head for human tracking. In his model multiple cameras are used for head detection and tracking is performed by considering the motion direction and velocity of human. Irshad Ali [10] proposed an automated detection and tracking of human in high density crowded scenes. AdaBoost detection cascade model is used for tracking purpose which was based on particle filter and for appearance modeling color histograms is used. Similar to Ran Eshel occlusion problem is covered by only consider head for tracking purpose. The error in head detection is reduced by the relation between head plane and ground plane.

Daniel Roggen [11] proposed a framework to recognize crowd behavior using mobile devices sensor. Traditionally cameras are used for analysis of crowd behavior which has disadvantage that it is geographically restricted. In proposed model mobile sensors are used for movement, orientation and GPS information. Using these data points activities of individuals are analyzed and these activities are grouped to form crowd behavior. Ramin Mehran [12] introduces abnormal behavior detection method in crowded scene. Unlike object tracking or

segmentation model is based on social force which estimated the interaction force on each individual in crowd. Social forces model is mainly used to describe the crowd behavior which is derived for the individual interaction. Interaction force of each pixel is calculated to find out the force flow which is then classify as abnormal behavior or normal behavior. S. R. Musse [13] take account into the relationship between of group of individual and the behavior originated from each individual. In his model each human is treated and individual object which react in the presence of other objects. In these interactions, parameters, goal, emotional status, interest, domination value etc, of each individual are changed accordingly. Davies [14] proposed a model for detection of crowd velocity (direction and magnitude). In this model he used discrete Fourier transform with linear area transform for segmentation of static and dynamic region in the image. In his approach motion feature are calculate on each pixel which are then aggregated to compute crowd motion behavior.

Weina Ge [15] model is based on conditional random field to detect and count people in crowded scenes. In his model he provided a Bayesian probability model which combines the spatial stochastic and conditional mark process to identify the shapes in crowded scenes. Conditional mark process identifies the correlations between size, orientation and image location. David Ryan [16] proposed method use local features for crowd size counting. Unlike holistic feature extraction for crowd counting local features are specific to individual or small region in the image. Each individual or group is identified using the foreground segmentation technique. Area, perimeter, area-to-perimeter ratio, edges and edge angle histogram are the main features for each group. Sum of crowd in each group represent the total crowd size. Calculating crowd size in each group also gives the crowd density at different regions in the image. In this way local density in each region/group can also be used to find out the crowd density in whole image. Hou [38] developed a methodology to estimate number of people in crowd and locate their location. In proposed model background is subtracted from foreground using Gaussian Mixture Model [17,18] After that foreground pixels are obtained by binaries the foreground image based on a threshold. Threshold represents the intensity difference between background image and foreground image. A neural network model is learned to count the number of people in crowded scene which take foreground pixel as feature set. Expectation Maximum algorithm is developed using estimated crowd size to identity the

location of each induvial. For detection purpose KLT [19] model is used which is good for corner detection and in tracking. After that Cluster model is performed on KLT features so that similar pixel can be grouped to locate individual. Dan Kong [20] describes a method to count people in crowded scene which is based on viewpoint invariant learning. In proposed method foreground is subtracted from background and unlike previous research blob size histogram is used in training along with edge orientation. Density of each blob in calculated and along with orientation scale it is used for feature normalization in each blob. A single hidden layer neural network is developed, which find out the relationship between number of people and feature set.

2.3 Visual Saliency

When human visual system perceives a scene it does not process whole scene. It extracts the region or objects that are most attracting and informative. These objects or regions are called salient. Visual saliency has gained much attention in last four decades and has applications in many domains like classification [21], image segmentation [22] and recognition in computer vision. Saliency models are based on the notion of detecting regions or objects that stand out from their neighbors and in some applications salient regions are considered as foreground regions. Different models are composed for visual saliency some of them are based on bottom up approach and some of them are task driven which use prior knowledge about the scene. In bottom up approach image features are used to extract the salient region like color, intensity, orientation and motion. Some work in saliency detection focused on visual uniqueness and is often deciphered from the image attributes such as gradient and edges. In task driven, objects or regions are detected based on the task and use high level features.

Laurent Itti [23] proposed first bottom-up visual saliency model. It is a simple and feed-forward feature extraction model for human visual attention. Proposed model is based on color, orientation, intensity, size and textual features. These features are extracted from input image and a visual saliency map is computed as output. Cheng [24] proposed bottom-up saliency detection model which is based on global contrast in image. The model is based on following consideration, unlike local contrast global contrast spate large number of objects from its surrounding, global contrast assign comparable saliency value to similar regions and produce highly uniform saliency map, saliency map depends on the nearby contrast regions and distant regions are less important. Based on above consideration, a histogram based contrast model is developed to measure saliency. To improve the result of histogram-based saliency map, a region-based contrast method is applied which incorporate spatial relations in image. In region-based contrast, image is segmented into different regions and color contrast is computed for each region. Goferman [25] proposed a new type of saliency model, context-aware visual saliency, which aims to identify the regions in an image that represent the scene. The model is based on the idea that salient regions are distinctive with respect to both global and local surroundings, thus some parts of background also salient. The model follows four

principle of human attention, local considerations (color, intensity etc), global considerations (frequently occurring features), visual organization rule (center prior) and high-level features (location and object detection).

Ali Borji [26] combines the low-level and high-level features to propose a saliency detection model. In this model low level features are extracted previous bottom-up visual saliency models and high-level features are extracted from top-down visual saliency model. Almost 30 low level features are extracted in which 13 features are from energy of steerable pyramid filters, 3 features are color, intensity and orientation, 3 features are value of red, green and blue color channel, 3 features represents the probability of each channel, 5 features are probability of each channel in 3D color histogram and 3 features are saliency map from bottom-up saliency. High level features are text, faces (which are detected using face detector), people and cars (which are detected using car detector). A SVM is implemented with Adaboost to compute saliency map using these low-level and high-level features. Tilke Judd [27] proposed model predict the human fixation in natural scene. Low, mid and high level features are considered in image semantics. In low level features, color, intensity, orientation, texture and motion features along with the probability of each channel (red, green and blue) are computed. In mid-level features horizon line in the image is detected and in high level features face and human are computed using face and human detector. A linear Support Vector Machine computation model is developed to compute the saliency map to predict human fixation.

Harel [28] use graph algorithm to achieve saliency map of an image. In this model he use varies step to developed saliency map. First an activation map is computed on different feature channels which represent the distinctively of a region to its surrounding. In second step, Normalizing activation map, activation maps are normalized to combine mass on activation maps. In last step Markov chains are defined on each activation map to compute the saliency values. Tie Liu [29] convert salient region detection problem into binary labeling task in which salient object are separated from background. In proposed model both static and dynamic features are used for feature set. From a single image multi-scale contrast, color spatial distribution and center surround histogram is used to describe salient object globally,

regionally and locally. For sequential images motion feature and appearance coherent feature is used to describe salient object. A conditional random field computation model is developed on above features to detect salient objects.

2.4 Visual Saliency in Crowd

Lim and Kok [30] propose a novel framework to detect salient region in crowded scenes. In proposed method low-level features are extracted from crowd motion and are transformed into global similarity structure. Dense optical flow algorithm is used to estimate the crowd motion field. After that low level features such as stability map and phase shift map are extracted from crowd motion field. Both features are transformed into the global similarity structure which gives the intrinsic manifold of motion dynamics. These dynamics gives the ranking on global similarity structure and extrema are identify as salient regions. Lim [31] proposed a model which takes into account flow field of crowd and its temporal variation to detect salient region. Flow field determine the motion dynamics within the given area and proposed model also resolved the issue of bottleneck and occlusions in the crowded scenes. The area having high motion dynamics is consider as salient region and unstable motion area is disregarding as noise. Mancas [32] proposed method used multi-scale approach for feature extraction from optical flow and global rarity. These features are used to compute bottom-up saliency map. In proposed method each video frame in divided into 3/5 pixel wide cell. Motion feature are computed for each cell and compared with neighborhood basis. Features are then discretized into 4 directions and 5 speeds. Spatio-temporal filter is applied on the discretized features channels which separate the space and time dimension. This gives the location of the salient region in the visual scene.

Rachel [33] shows the important of face and head in visual saliency in crowded scenes. In the experiment virtual human are used to stimulate the crowded scene with different orientation, motion variation, size, gender, age and clothing style of the clones. A selective variation model is developed using eye-tracking device which recognize head, top texture and face texture as more salient part in human body. It also be verified that facial features have no effect on visual saliency of humans. Antoine [34] shows the importance of faces and audio in social scene. Conversation is recorded in various contexts and fixation map is computed for ground truth. A statistical model is developed based on the observation in fixation map which maximize the expectation (Expectation-Maximization). The feature set is comprise of static low level saliency map, dynamic low level visual saliency, faces, center bias and audio.

Experiment shows that faces are more salient especially the talking faces and low level features are irrelevant in social scenes.

Moran Cerf [35] proposed method combine the low-level features with face detection to compute visual saliency. The model is best suited for natural scenes with only frontal faces. Ground truth are developed using eye tracking devices in which almost 80% fixation are on faces. In proposed method low-level features, color, intensity and orientation, are computed using graph based visual saliency and faces are detected by using Viola & Jones algorithm. The performance of the model has been improved by introducing the face channel in saliency map detection. Sophie [36] proposed model is based on three saliency maps static, dynamic and face. Static saliency map is computed using orientation and spatial frequency. Dynamic saliency map is computed using object motion amplitude and face saliency map is computed using face detection algorithm. All of these saliency maps are computed for each frame in the video and combined with center bias to compute saliency map. Jiang [37] proposed model statically show the importance of face in crowded scenes. Jiang created dataset for crowded scenes in which faces are manually labeled with rectangle and two other features are calculated, size of the face and post of the face. In this model eye tracking device is used to compute the fixation map and statically shows that faces are more salient in crowded scene.

Chapter 3

3. PROPOSED APPROACH

3.1 Overview

In our proposed solution we used low-level features (color, intensity, orientation) alongside mid-level features (face size, face density, and face pose). We also proposed different features to compute crowd level in the visual scene. The crowd level was determined by considering crowd density and number of faces in each image. K-medoid clustering algorithm was applied on 2d-feature vector (crowd density, number of faces) to compute the crowd level.

The proposed model was evaluated using an available dataset comprising static crowded scenes with eye-tracker recordings. Using different criteria, we showed significant improvement against existing visual saliency models for crowds.
The main contributions of this paper are as follows:

- A new method was proposed to compute face density. Face density is further used to compute crowd density.
- Crowd features such as number of faces and crowd density were used to categorize crowds into low, mid and high level. It is natural to think that crowd of different density should receive a different level of attention.
- A deep fully connected neural network was proposed to compute saliency in crowded scenes which use both lower level features and faces attributes and taking into account crowd level.

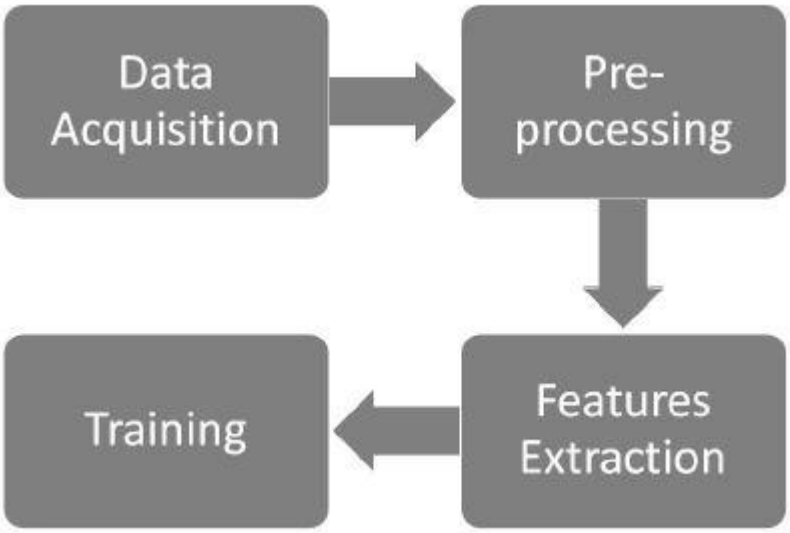

Figure 3.1: Model Diagram

3.2 Dataset

For our model we used data set [38] that is specially created for research on crowds. The data set comprise 500 images with varying crowd density and number of faces. The dataset contain both indoor and outdoor scenes. Each image had fixed resolution 1024 x 768. Human faces were manually labeled as rectangles, while each face had two more attributes: pose and partially occluded. Pose attribute can be frontal, profile or back. If the angle between face view and image plane was less than 45 then it was categorized as frontal. If angle was between 45 and 90 degree then pose was profile otherwise back. Partial occlusion is a binary attribute if face is partially occluded then it will be 1 otherwise it will be 0. Fixation maps of the ground truth were created by sixteen students containing 10 male and 6 females. Eye tracking device is used to record the eye movements of the subjects. 15.79 mean and 0.97 standard derivation is calculated in eye fixations of each test.

Figure 3.2: Dataset Examples

3.3 Methodology

3.3.1 Preprocessing:

Dataset consist of 500 images and 16 subjects are used to compute ground truth of visual saliency. Eye tracking device is used to record fixation point. In preprocessing part we first compute the ground truth using these fixation points. Fixation points are in 2D space and represent the x-coordinate and y-coordinate. A 3rd parameter is the duration of a fixation point. Each image is viewed for 5 sec to each subject and each subject has different number of fixation points. In developing ground truth we first ignore all those fixation points that are outside the image dimension. Each image is composed of 1024x768. A fixation point in the image represent the salient region where as none represent the salient region. A Gaussian filter of eye radius size is applied on the ground truth and after that image is normalized using max-min normalization.

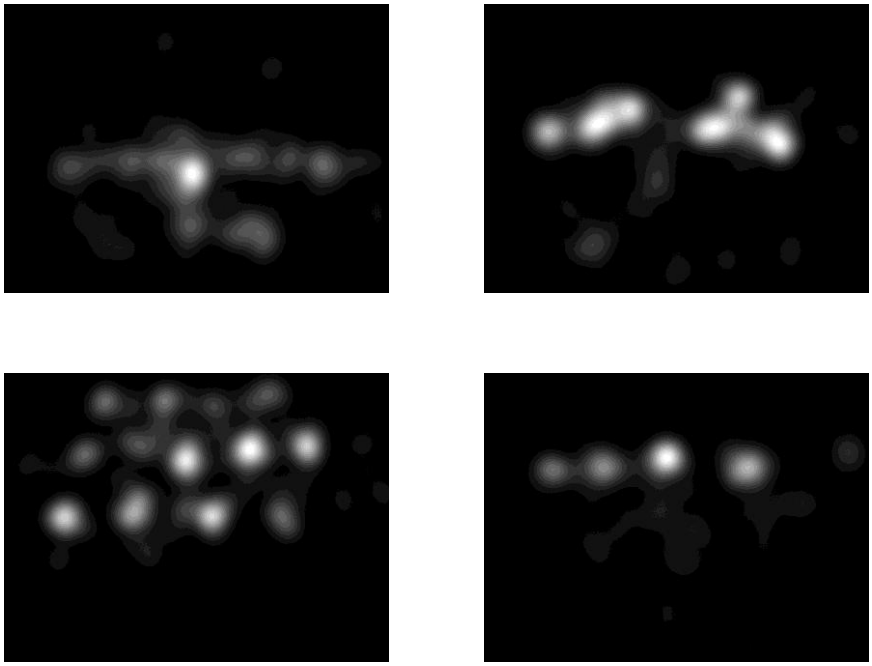

Figure 3.3: Fixation Map of dataset

Dataset contain varies number of faces in images ranging from 3 faces to up to 268 faces. Each face is manually labeled and following attributes are computed

- Top-coordinates
- Left-coordinates
- Width
- Height
- Center
- Pose

There is possibility that actual number of faces in high crowded scene is greater than labelled faces.

Following are other attributes that are computed for each face and image.

Local Face Density:

Density of each face is calculated with respect to other face. Density represents the attractiveness of region with respect to other region. If value of density is less then that region is higher attractive. If an image contains large number of faces then every face will have high value of density and there will be high competition for attention. The density of each face is locally calculated with respect to faces within the image. To compute face density we first compute the distance between two faces. Higher the distance lowers the face density. If two faces are apart a lot then there will be less competition for attention. Each face represents an oval in the image so we have to compute the distance between two ovals. In our computation we take necessary measurement for distance calculation. Rather than computing distance between centers of oval we compute the distance between the boundaries of the ovals. To compute the distance between boundaries we first need to compute radius of each oval so that we can subtract that radius in computation of radius. Radius of each oval is calculated using width, height and at the angle between center and the point at which distance will be calculated. The point is calculate using below formula

$$P = face1.center - face2.center$$

That point lies on the distance line of ovals. The angle is computed between the point and surface line using below formula

$$angle = \text{atan}(\frac{P.x}{P.y})$$

Once angle has been computed, we compute the radius of oval using below formula.

$$radius = (w * h)/(sqrt(\ (w * w) * (\sin(angle)^2) + (h * h) * (\cos(angle)^2))$$

After calculated radius we calculate distance between both oval using distance formulas below

$$distance = (face1.x - face2.x)^2 + (face1.y - face2.y)^2$$

To calculate distance between boundaries of ovals we subtract the radius of both ovals from above distance. Distance between the boundaries of oval is used to compute to face density. Remember face density is computed with respect to other face so both faces will have face density with respect to each other. Below is the formula to compute face density

$$density = \exp(\frac{distance}{2 * sigma^2})$$

Density of a face is computed with respect to each face and after that each density is summed to compute the overall density of a face with respect to each face. This density represents the accumulated effect other faces on a single face.

Local Face Size:

Face size represent the size of a face in the image. Faces are more salient region in a human body and in crowded scene large faces are attractive to attention. Statically it have proven that if a scene is much crowded then faces are not important feature this is because in highly crowded scene each face will have same size and will not take part in saliency computation. Face size is computed using bellow formula

$$face\ size = face.width * fac.height$$

Crowd Density:

Crowd density plays an important role in visual saliency in crowded scene. Crowd density represents the crowdedness in scene. The density of a crowded scene depends on the number of human in the scene. Research has shown the different crowd density receive different level of attention. In some research crowd density comprise of region density which use to solve many problem like crowd behavior. In our computation we use faces to compute crowd density. As we know crowd density is directly related to number of human in the scene so to calculate crowd density we use face density. Crowd density is the mean of face density in the image. It is some kind of global density which differentiates an image from other and is more dynamic as compare to number of people.

$$Crowd\ Density = mean(facesDensity)$$

The above formula gives the normalized crowd density and a good metric to classify an image sparse or dense scene.

3.3.2 Feature Extraction

We first compute visual saliency map of low-level features (color, intensity, orientation) of the dataset. We used Graph base visual Saliency [28] model for this purpose which had better performance than [23] basic model. After that we compute visual saliency map of face attributes; size, density, frontal pose and profile pose.

Low-Level Saliency Map:

Color, intensity and orientation are the textual feature of an image which plays an important role in visual saliency. Itti proposed his model to compute these features set and developed a model for visual saliency on low-level features set. In our model we used graph based visual saliency which is the improved version of itti model. Using graph based visual saliency we compute saliency map of each low level feature color, intensity and orientation. Color features represent the variation in the color of an image. Intensity represents the heat map of an image whereas orientation represents the direction of different objects in the saliency map. Below are the visual saliency maps of each low-level feature: color, intensity and orientation

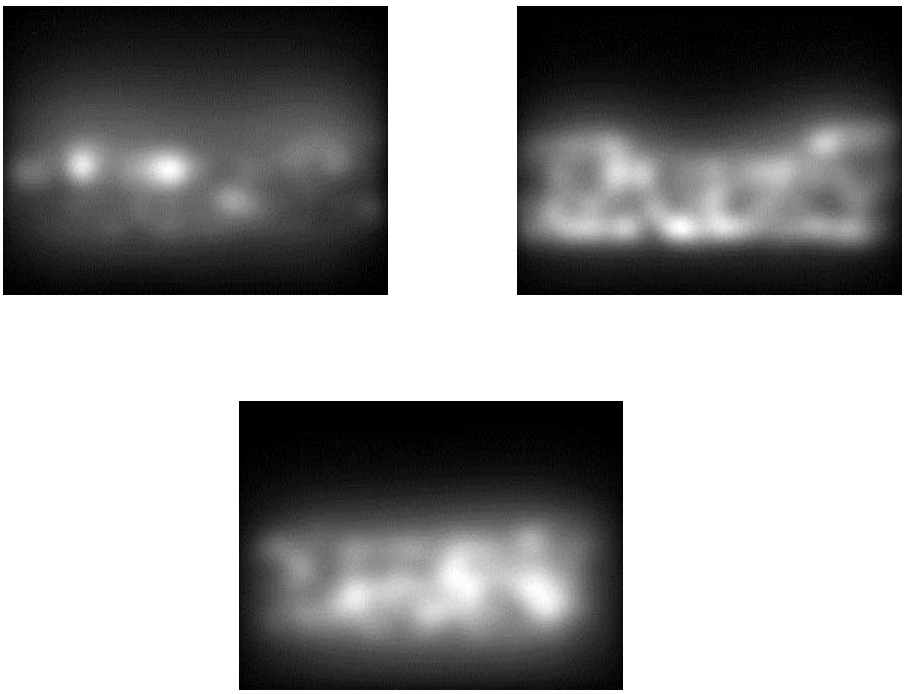

Figure 3.4: Saliency Map using low color, intensity and orientation

Face Attributes

We have also computed other attributes like face size, face density, frontal pose and profile pose. We have also created visual saliency map for each these attributes. These visual saliency maps are feature set in the computation model. To compute visual saliency map for each face attribute, we set the center of face with the value of each attribute. Thus we have 4 visual saliency maps for an image where each visual saliency map represents one single attribute. The saliency map is resized to 256 width and 192 height so that salient region become more promient. Gaussian filter is applied on the visual saliency map of each attribute and after than each image is normalized. Below are the visual saliency maps of each feature set: density, frontal, profile and size.

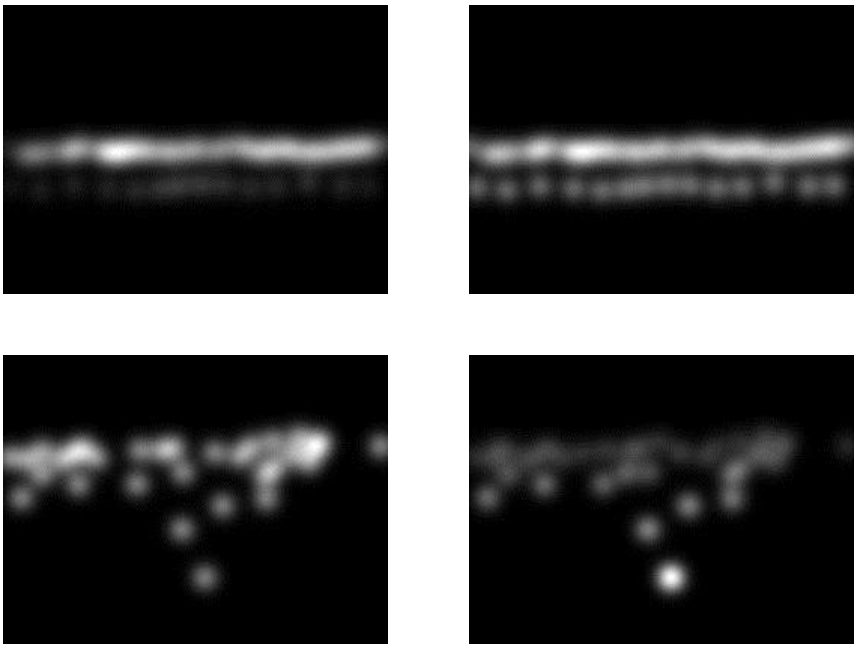

Figure 3.5: Saliency map of density, frontal, profile and size of face

3.3.3 Computation Model

The saliency maps that created in feature extraction process act as features set for our computation model. Before applying our computational model, we first categorized the data set into different crowd levels (low, mid, high) and developed model for each crowd level. It is natural to think that different crowd level attracts different kind of attention.

Data Partitioning:

In previous step we have calculated the density of each crowd scene. We used the local density of each face to compute the crowd density of the scene. In our computation take into account crowd density and propose a method to compute different crowd level. Crowd density along with number of faces in each image was used to calculate the crowd level. We first normalized both feature vector between 0 and 1 and then applied K-Medoid clustering algorithm on these two features set to compute crowd level. Crowd data was then categorized into three levels: low, mid and high.

Data Sampling:

Data is partition into different crowd level to proposed computation model for each crowd level. In computation we divide the data in training and testing data set and perform 10-cross validation for the evaluation of our models. Thus in 500 images 450 images are used for training and 50 images are used for testing purposed. In those 450 images these are divided into different crowd level. For each crowd level a model is train. In our computation model we have feature set of size 7 (color, intensity, orientation, size, density, frontal, profile). For sampling purpose we rearrange each feature saliency map into one dimension vector such that each value in vector represents single pixel. Thus for training purpose we have feature set of

$$Feature\ Set = [training\ images * Height * Width][Feature\ set\ Size(7)]$$

The data is imbalance in the above feature set because there will be very less number of salient region as compare to salient region. To overcome this problem we have selected positive point in the top 5% of salient region and negative point from bottom 20% non-salient region. We have selected 10 data points from each positive and negative data sample. Thus 20 data set per image. Thus total training dataset will be

$$Training\ Set = [Number\ of\ training\ images * 20][Feature\ set\ Size(7)]$$

3.3.4 Model Training

We proposed deep fully connected neural network to predict saliency map for each crowd level. Our model has one input layer, three hidden layer and one output layer. Scaled Conjugate Gradient was used as training algorithm on each layer whereas log-sigmoid was used as transfer function between the layers. Weights on each layer was calculated separately and then combined all of layers to create connected neural network. First, hidden layer calculated the weights using input size of seven and produced output of 200 features set for second hidden layer. Second, hidden layer computed weights using input size of 200 features set and produced output of 200 features set for third hidden layer. Third, hidden layer computed weights using input size of 200 features and produced output of two features set for output layer.

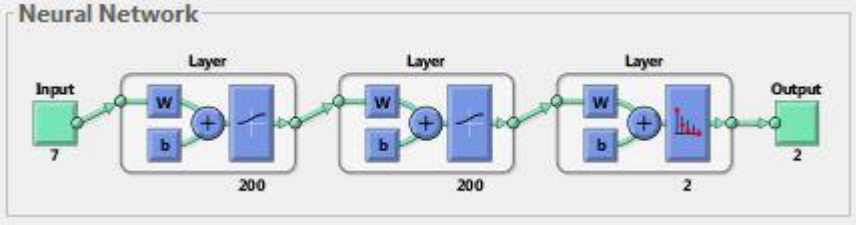

Figure 3.6: Neural networking with fully connected 2-hidden layers

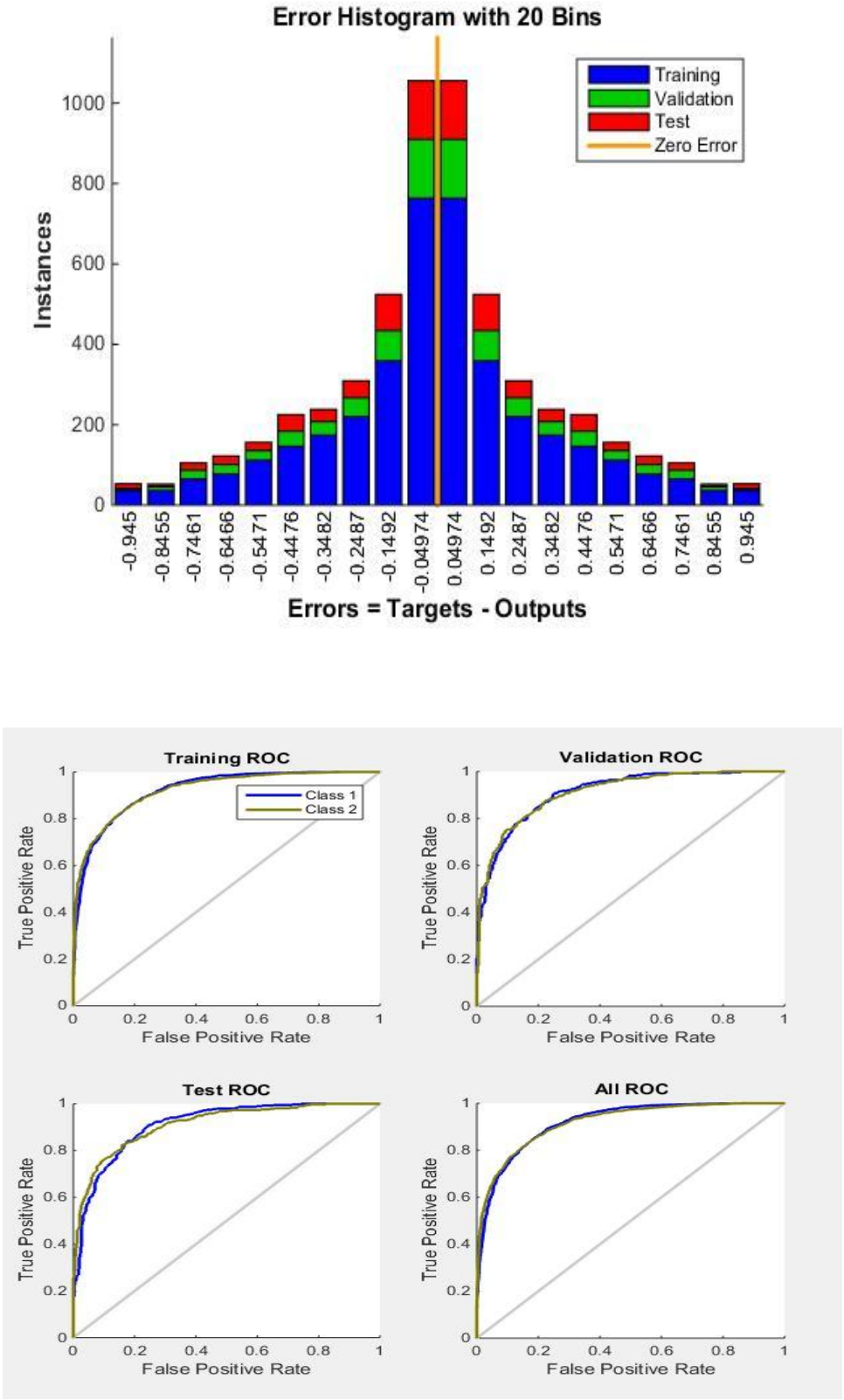

Figure 3.7: Histogram and ROC diagram of computation model

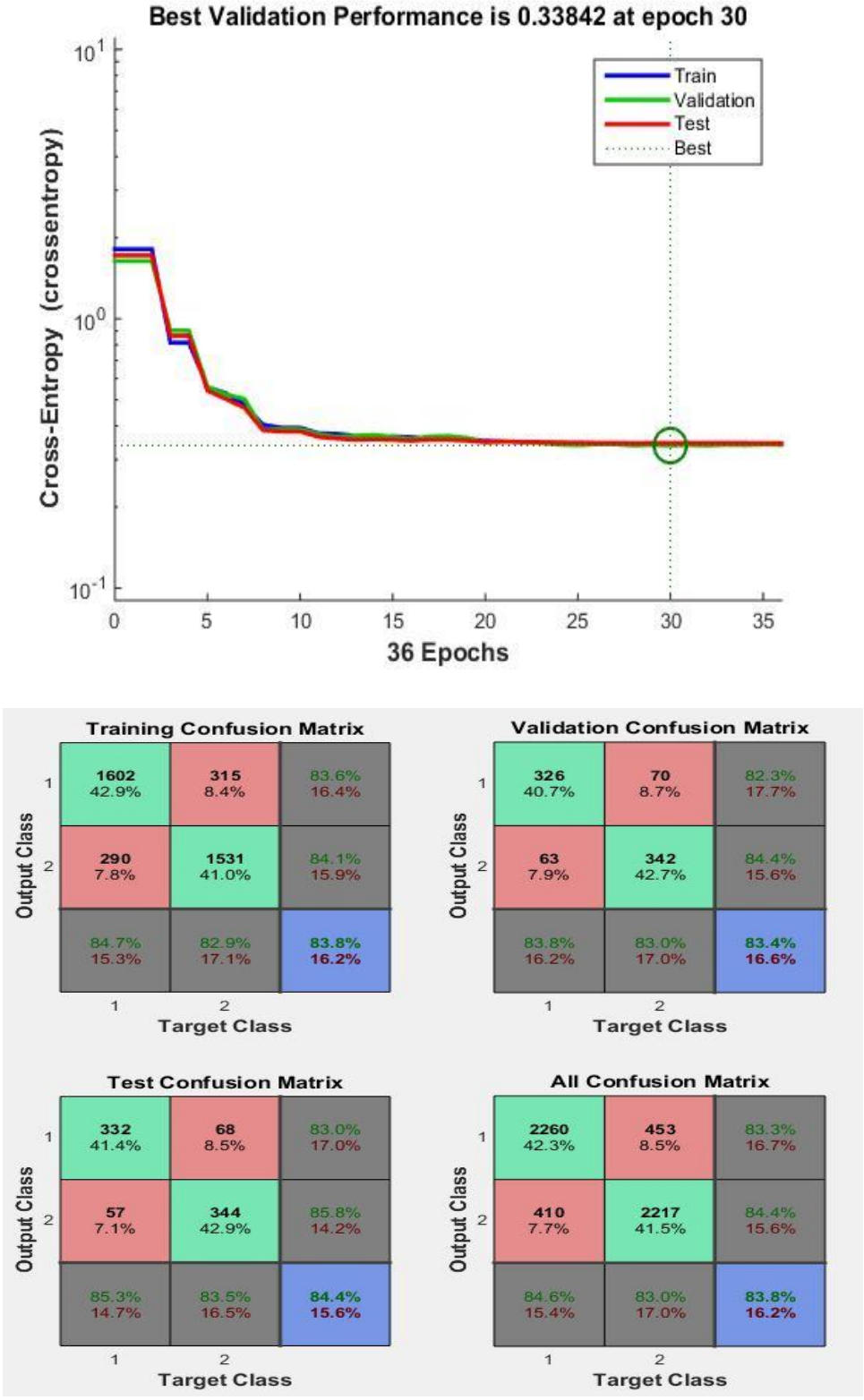

Figure 3.8: Error Minimization and Confusion Matrix of computation model

Chapter 4

4. RESULT AND EVALUATION

4.1 Overview

We performed 10-cross validation for the evaluation of our model. 450 images were used for training and 50 images were used for testing. From each images 10 positive and 10 negative samples were taken from top 10% salient region and below 20% non-salient region respectively. Each set contained seven features (color, intensity, orientation, face size, density, frontal face, profile face) and each value was normalized to have zero mean and unit variance. Once model has been developed using the training data set, we compute the visual saliency map for the test dataset. We have used 10-cross validation so in each iteration 50 visual saliency map is created and after 10^{th} iteration we have saliency map of each image in dataset. We have compared visual saliency map created by our model with the fixation map created using eye tracking device. We used Area under the Curve (AUC), Normalized Scanpath Saliency (NSS) and Correlation Coefficient (CC) to measure the statistical relationship between the ground truth and the saliency predictions of the model. In case of AUC, we used two different implementations, Judd AUC and Shuffled AUC.

4.2 Model and Results

We have developed different models with respect to different criteria below are the models and their results.

4.2.1 Single Deep Neural Network Binary Class Problem

In this model we didn't take into account crowd density and developed one single model for all dataset. We treated saliency computation as 2 class problem (salient or non-salient). We have use 10 positive sample and 10 negative sample so feature set is of size 9000 and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.7929	**0.6508**	**1.2411**	**0.5514**

Table 4.1: Result of Single Deep Neural Network Binary Class Problem

In our evaluation we have deep down into the result and found that 138 images has shuffle AUC value greater than 0.7, 241 images have shuffle AUC value between 0.6 and 0.7 and 121 images has shuffle AUC value less than 0.6.

4.2.2 Single Deep Neural Network Multi Classes Problem

In this model we didn't take into account crowd density and developed one single model for all dataset. Unlike previous model we treated saliency as multiclass problem. In which we just classify a region as salient or non-salient but also identify how much a particular region is salient. A pixel can have value between 0 and 255 so we have 256 classes. We have use 10 positive sample and 10 negative sample so feature set is of size 9000 and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.7728	**0.6659**	**1.3980**	**0.5689**

Table 4.2: Result of Single Deep Neural Network Multi Classes Problem

In our evaluation we have deep down into the result and found that 150 images has shuffle AUC value greater than 0.7, 222 images have shuffle AUC value between 0.6 and 0.7 and 128 images has shuffle AUC value less than 0.6.

4.2.3 Four Model Clustering based on Number of Faces

In this model we take into account number of faces and cluster images based on number of faces. Number of cluster represents the number of crowd level in the given image and we developed model for each crowd level. We have use 10 positive samples and 10 negative samples so feature set is different for each model based on number of images in the cluster and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.7926	**0.6501**	**1.2276**	**0.5468**

Table 4.3: Result of Four Model Clustering based on Number of Faces

In our evaluation we have deep down into the result and found that 138 images has shuffle AUC value greater than 0.7, 236 images have shuffle AUC value between 0.6 and 0.7 and 126 images has shuffle AUC value less than 0.6.

4.2.4 Four Model Clustering based on Number of Faces and Density

In this model we take into account number of faces and density of the image. We cluster images based on number of faces and density. Number of cluster represents the number of crowd level in the given image and we developed model for each crowd level. We have use 10 positive samples and 10 negative samples so feature set is different for each model based on number of images in the cluster and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.7781	**0.6601**	**1.3550**	**0.5511**

Table 4.4: Result of Four Model Clustering based on Number of Faces and Density

In our evaluation we have deep down into the result and found that 173 images has shuffle AUC value greater than 0.7, 211 images have shuffle AUC value between 0.6 and 0.7 and 116 images has shuffle AUC value less than 0.6.

4.2.5 Five Model Clustering based on Number of Faces and Density

In this model we take into account number of faces and density of the image. We cluster images based on number of faces and density. Number of cluster represents the number of crowd level in the given image and we developed model for each crowd level. We have use 10 positive samples and 10 negative samples so feature set is different for each model based on number of images in the cluster and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.7741	**0.6622**	**1.3392**	**0.5503**

Table 4.5: Result of Five Model Clustering based on Number of Faces and Density

In our evaluation we have deep down into the result and found that 178 images has shuffle AUC value greater than 0.7, 221 images have shuffle AUC value between 0.6 and 0.7 and 101 images has shuffle AUC value less than 0.6.

4.2.6 Four Model Clustering based on Density

In this model we only take into account density of the image. We cluster images based on number density. We set the cluster size to 4 and number of cluster represents the number of crowd level in the given image and we developed model for each crowd level. We have use 50 positive samples and 50 negative samples so feature set is different for each model based on number of images in the cluster and 7 features are used for computation. Below is the table showing the result of this model:

Judd AUC	Shuffle AUC	NSS	CC
0.8184	**0.6781**	**1.3234**	**0.5890**

Table 4.6: Result of Four Model Clustering based on Density

In our evaluation we have deep down into the result and found that 213 images has shuffle AUC value greater than 0.7, 210 images have shuffle AUC value between 0.6 and 0.7 and 79 images has shuffle AUC value less than 0.6.

We have also developed some other model in which we use variation in crowd level like 3, 4 and 5. We have also changed input feature set for clustering. We also used face size along with faces and density to compute cluster size. None of these models have better accuracy then clustering based on density. If we change cluster size from 4 to 5 then there is very little change in the value of above model. We have compared our models with existing state of art model on given dataset and verified that our each model has better accuracy than existing state of art model The results show that NN-F model performed better as compared to the state-of-the-art model for saliency in crowd. Similarly, the NN-CF and NN-CFD models using dynamic categorizations also out performed previous models. Furthermore, to verify the importance of crowd features, in NN-S model, we used only one level of crowd. The results show the importance of crowd features and clustering into low, mid and high levels to improve saliency prediction in crowded scenes.

4.3 Evaluation

We have compared the result of our models with existing state of art saliency models. Below are the results of different models on crowd datasets.

4.3.1 Itti Model

We have run Itti[26] model on the crowd data set which is based on low level features (color, intensity and orientation). Below are the results of Itti model, Baseline of Itti model is 0.53 Shuffle AUC whereas in given crowd dataset model has perform little well. The reason is that in highly crowded scene high level and midlevel features don't have much impact.

Judd AUC	Shuffle AUC	NSS	CC
0.61	**0.55**	**0.81**	**0.39**

Table 4.7: Result of Itti model on crowded dataset

4.3.2 Graph Based Visual Saliency

Graph based visual saliency [28] is the improvement of Itti model and has better result than Itti model. Below are the results of GBVS model, Baseline of GBVS model is 0.63 Shuffle AUC whereas in given crowd dataset it has reduced.

Judd AUC	Shuffle AUC	NSS	CC
0.70	**0.58**	**0.96**	**0.44**

Table 4.8: Result of GBVS model on crowded dataset

4.3.3 Boolean Map Saliency

Boolean Map Saliency Model [39] is based on bottom-up features of the image. In this model images are characterized by set of binary images. We run the model on crowd dataset and below are the result of above model

Judd AUC	Shuffle AUC	NSS	CC
0.78	**0.65**	**1.41**	**0.55**

Table 4.9: Result of BMS model on crowded dataset

4.3.4 Covariance Based Saliency

Covariance mean based saliency model [40] is developed using multiple kernal learning. We have run the following model on the crowded dataset below is the result of the covariance based saliency.

Judd AUC	Shuffle AUC	NSS	CC
0.77	**0.59**	**1.22**	**0.45**

Table 4.10: Result of Cov-Sal model on crowded dataset

4.3.5 Ensembles of Deep Network

A deep ensemble neural network [41] model is developed which is data driven and extract the features from dataset. We have run the trained model on the crowd dataset. Below are the results of the model on crowd dataset.

Judd AUC	Shuffle AUC	NSS	CC
0.81	**0.62**	**1.14**	**0.45**

Table 4.11: Result of Deep Network model on crowded dataset

4.3.6 Multi-Level Network

A deep multi-layer convolution network [42] model is developed for saliency detection. The model has 0.85 Judd AUC for MIT300 [43] dataset. We run the above model with the given weights on the crowd dataset and found that the value of AUC decreases. Below are the results of ML-Net model

Judd AUC	Shuffle AUC	NSS	CC
0.82	**0.68**	**0.60**	**1.78**

Table 4.12: Result of ML-Net model on crowded dataset

Chapter 5

5. CONCLUSION AND FUTURE WORK

Visual Saliency plays an important role in many fields like automated CCTV camera monitoring, anomaly detection, human detection etc. Various models have been developed to detect visual saliency in videos and natural scene. Due to complex nature and irregular density of crowd these models are unable to perform better in visual saliency. Most of the models are developed on low level features and ignore the high or mid-level features in crowded scenes. Also these models don't consider the specialize features of crowds like crowd density and crowd behavior. Crowded scenes are more frequent in real world, like in political campaigns, matches, concerts etc, and have many applications in different fields like in computer vision, crowd management, crow psychology etc.

We proposed a deep fully connected neural network model for computation of saliency maps in crowded scenes. In our model we used low level features along with mid-level features to build feature set. For mid-level features we take into account faces. Many researches have shown that faces are most salient region in human body. We compute different face attributes like face size, face density, frontal face and profile face. We provided an accurate measurement of face density. Beside all these local features that are related to faces, we compute crowd density which is related to images. Crowd density represents the crowdedness in an image and a most important feature in crowded scene. Crowd density can be used to classify crowd into different levels like low, mid and high etc. It is natural to think that different crowd levels attract different level of attention.

All of these features (color, intensity, orientation, face size, face density, frontal face and profile face) along with crowd density are used to develop a computation model which classifies a region into salient or non-salient region. We have also changed the binary class problem into multi class problem in which we just not only provide which region is salient but also compute the exact saliency value for that region.

In our computation model we have found that if we use multi class problem then our model is very slow as compare to binary class problem. This is because high computation and more weights need to be calculating if there are high numbers of classes.

Our model is only limited to static images and can't be performed on videos. In future work we can developed a model for videos. In videos we can many other features like optical flow, motion, direction and speed we can also take into account these features for computation. Our model is based on faces and in dataset faces are manually labelled with width and height. It is mandatory we have faces detected already for our computation. In highly crowded scenes faces have no significant importance for saliency detection thus in that case only low level features take part in saliency computation. Motion features are also restricted in highly crowded scenes and will have to impact on saliency detection.

Bibliography

[1] Mahadevan, Vijay, et al. "Anomaly detection in crowded scenes." *CVPR*. Vol. 249. 2010.

[2] Kratz, Louis, and Ko Nishino. "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[3] Wu, Shandong, Brian E. Moore, and Mubarak Shah. "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes."*Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[4] Reddy, Vikas, Conrad Sanderson, and Brian C. Lovell. "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture." *CVPR 2011 WORKSHOPS*. IEEE, 2011.

[5] Ma, Ke, Michael Doescher, and Christopher Bodden. "Anomaly Detection In Crowded Scenes Using Dense Trajectories."

[6] Bera, Aniket, and Dinesh Manocha. "Realtime multilevel crowd tracking using reciprocal velocity obstacles." *arXiv preprint arXiv:1402.2826* (2014).

[7] Ali, Saad, and Mubarak Shah. "Floor fields for tracking in high density crowd scenes." *European conference on computer vision*. Springer Berlin Heidelberg, 2008.

[8] Fan, Jialue, et al. "Human tracking using convolutional neural networks."*IEEE Transactions on Neural Networks* 21.10 (2010): 1610-1623.

[9] Eshel, Ran, and Yael Moses. "Homography based multiple camera detection and tracking of people in a dense crowd." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[10] Ali, Irshad, and Matthew N. Dailey. "Multiple human tracking in high-density crowds." *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer Berlin Heidelberg, 2009.

[11] Roggen, Daniel, et al. "Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods." *arXiv preprint arXiv:1109.1664* (2011).

[12] Mehran, Ramin, Alexis Oyama, and Mubarak Shah. "Abnormal crowd behavior detection using social force model." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[13] Musse, Soraia Raupp, and Daniel Thalmann. "A model of human crowd behavior: Group inter-relationship and collision detection analysis." *Computer Animation and Simulation'97*. Springer Vienna, 1997. 39-51.

[14] Junior, Julio Silveira Jacques, Soraia Musse, and Claudio Jung. "Crowd analysis using computer vision techniques." *IEEE Signal Processing Magazine* 5.27 (2010): 66-77.

[15] Ge, Weina, and Robert T. Collins. "Marked point processes for crowd counting." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[16] Ryan, David, et al. "Crowd counting using multiple local features." *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*. IEEE, 2009.

[17] Grimson, W. Eric L., et al. "Using adaptive tracking to classify and monitor activities in a site." *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998.

[18] Stauffer, Chris, and W. Eric L. Grimson. "Learning patterns of activity using real-time tracking." *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000): 747-757.

[19] Tomasi, Carlo, and Takeo Kanade. *Detection and tracking of point features*. Pittsburgh: School of Computer Science, Carnegie Mellon Univ., 1991.

[20] Kong, Dan, Douglas Gray, and Hai Tao. "A viewpoint invariant approach for crowd counting." *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3. IEEE, 2006.

[21] Haralick, Robert M., and Karthikeyan Shanmugam. "Textural features for image classification." *IEEE Transactions on systems, man, and cybernetics*6 (1973): 610-621.

[22] Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation."*IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000): 888-905.

[23] Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998): 1254-1259.

[24] Cheng, Ming-Ming, et al. "Global contrast based salient region detection."*IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.3 (2015): 569-582.

[25] Goferman, Stas, Lihi Zelnik-Manor, and Ayellet Tal. "Context-aware saliency detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*34.10 (2012): 1915-1926.

[26] Borji, Ali. "Boosting bottom-up and top-down visual features for saliency estimation." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[27] Judd, Tilke, et al. "Learning to predict where humans look." *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009.

[28] Harel, Jonathan, Christof Koch, and Pietro Perona. "Graph-based visual saliency." *Advances in neural information processing systems*. 2006.

[29] Liu, Tie, et al. "Learning to detect a salient object." *IEEE Transactions on Pattern analysis and machine intelligence* 33.2 (2011): 353-367.

[30] Lim, Mei Kuan, et al. "Crowd saliency detection via global similarity structure." *arXiv preprint arXiv:1410.3756* (2014).

[31] Lim, Mei Kuan, et al. "Detection of salient regions in crowded scenes."*Electronics Letters* 50.5 (2014): 363-365.

[32] Mancas, Matei, et al. "Abnormal motion selection in crowds using bottom-up saliency." *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011.

[33] McDonnell, Rachel, et al. "Eye-catching crowds: saliency based selective variation." *ACM Transactions on Graphics (TOG)*. Vol. 28. No. 3. ACM, 2009.

[34] Coutrot, Antoine, and Nathalie Guyader. "How saliency, faces, and sound influence gaze in dynamic social scenesShort Title??." *Journal of vision* 14.8 (2014): 5-5.

[35] Cerf, Moran, et al. "Predicting human gaze using low-level saliency combined with face detection." *Advances in neural information processing systems*. 2008.

[36] Marat, Sophie, et al. "Improving visual saliency by adding ‘face feature map’and ‘center bias’." *Cognitive Computation* 5.1 (2013): 63-75.

[37] Jiang, Ming, Juan Xu, and Qi Zhao. "Saliency in crowd." *European Conference on Computer Vision*. Springer International Publishing, 2014.

[38] Hou, Ya-Li, and Grantham KH Pang. "People counting and human detection in a challenging situation." *IEEE transactions on systems, man, and cybernetics-part a: systems and humans* 41.1 (2011): 24-33.

[39] Zhang, Jianming, and Stan Sclaroff. "Saliency detection: A boolean map approach." *Proceedings of the IEEE International Conference on Computer Vision*. 2013.

[40] Erdem, Erkut, and Aykut Erdem. "Visual saliency estimation by nonlinearly integrating features using region covariances." *Journal of vision* 13.4 (2013): 11-11.

[41] Pan, Junting, et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." *arXiv preprint arXiv:1603.00845* (2016).

[42] Cornia, Marcella, et al. "A Deep Multi-Level Network for Saliency Prediction."*arXiv preprint arXiv:1609.01064* (2016).

[43] http://saliency.mit.edu/