Face Detection and Verification in Unconstrained Videos: Challenges, Detection, and Benchmark Evaluation

Mahek Shah

IIIT-D-MTech-CS-GEN-13-106
July 16, 2015

Indraprastha Institute of Information Technology, Delhi

Thesis Advisors

Dr. Mayank Vatsa
Dr. Richa Singh

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science

Keywords: face recognition, face detection, face verification

Certificate

This is to certify that the thesis titled **"Face Detection and Verification in Unconstrained Videos: Challenges, Detection, and Benchmark Evaluation"** submitted by **Mahek Shah** in partial fulfillment of the requirements for the degree of *Master of Technology* in *Computer Science & Engineering* is a record of the bonafide work carried out by him under our guidance and supervision in the Image Analysis and Biometrics group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Mayank Vatsa

Dr. Richa Singh

Indraprastha Institute of Information Technology, Delhi

Abstract

With increasing security concerns, surveillance cameras are playing an important role in the society and face recognition in crowd is gaining more importance than ever. For video face recognition, researchers have primarily focused on controlled environments with a single person in a frame. However, in real world surveillance situations, the environment is unconstrained and the videos are likely to record multiple people within the field of view. Surveillance videos encompass multiple challenges for face detection and face recognition. For instance, detection algorithms may be affected due to size of a face image, occlusion, pose, illumination, and background while recognition algorithms may be affected due to low resolution, occlusion, pose, illumination, and blurriness. State-of-the-art approaches for both face detection and face recognition in such challenging scenarios are currently in nascent stages. Moreover, due to the unavailability of such databases, it is difficult for researchers to pursue this important challenge. This thesis attempts to fill the gap in unconstrained face recognition in two ways: (1) develop a large unconstrained video face database, and (2) create a benchmark protocol and perform baseline experiments for both face detection and recognition. As the first contribution of this thesis, a large video database of 384 videos consisting of 258 subjects is prepared. Each video generally contains multiple subjects in unconstrained settings. Further, ground truth for face and landmark (eye and mouth) detection is manually annotated. As the second contribution of this thesis, we design a benchmark protocol for face detection and recognition evaluation. Using the protocols, we evaluate existing face detection and face recognition approaches, including commercial systems. Poor face detection and verification results showcase the challenging nature of the problem and the database.

Acknowledgments

I would like to thank my advisors Dr. Mayank Vatsa and Dr. Richa Singh for their stimulating suggestions and encouragement which helped me in researching and writing this thesis. Their excellent guidance and motivation was a source of inspiration for me whenever I got stuck in the work. Besides my advisors, I would like to thank Mr Tejas Dhamecha for his continuous guidance, encouragement, and support during my thesis. It would have been impossible to complete the work without his help. Next, I would like to thank my fellow mate Priyanka Verma for her help and cooperation. I am thankful to the entire academic department, specially Mr. Ashutosh Brahmna, for their support and endless help.

And last but not the least, I would like to extend special thanks to my family for their patient love which enabled me to complete this work.

This research is partially funded by Department of Electronics and Information Technology, Government Of India.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There have been several research directions in automated face recognition [1]–[4]. Most of the research focuses on constrained environment with limited variations in pose, expressions, and illumination. On the other hand, real world applications require algorithms to handle variations due to low resolution, noise, multiple subjects in a frame, along with large variations in pose, expressions, and illumination. While early research in face recognition has focused on still images, recent research threads are utilizing videos for improving recognition performance. Video face recognition is also applicable in surveillance applications and can provide abundant information for extracting meaningful features. However, when a video is captured with multiple subjects in a frame in unconstrained settings (without user cooperation), video face recognition becomes an equally or perhaps, more difficult problem. The problem is exacerbated when both gallery and probe are videos captured in surveillance condition. As shown in Figure 1.1, real world applications also require matching a video with another video, obtained from different sensors, to determine the movement of a suspect.

In recent literature, several video face recognition algorithms have been proposed. As shown in Table 1.1, 100% accuracy has been achieved on databases such as Honda/UCSD [5]. On the other hand, challenging databases such as YouTube [6] and Point and Shoot Challenge [7] are used to enhance the capabilities of modern algorithms. However, it is to be noted that none of these databases capture unconstrained videos of crowd, i.e., two or more subjects in each video.

Figure 1.1: A law enforcement application scenario where subjects are matched using surveillance footage only. Top row of the figure shows four frames/images from the Boston bombing case. The suspects (the subject in black hat and the subject in white hat) can be seen walking along with other subjects. The bottom row show the face regions of the suspects.

Therefore, it is challenging to visualize the current capabilities of face recognition algorithm on unconstrained videos with multiple subjects. Next, we present a summary of related databases followed by summarizing our contributions.

1.1 Details of Existing Datasets

There exist few video face database that are used in evaluating the performance of recognition algorithm.

1. Face-In-Action (FIA) [13]:

   FIA database is specially created for border-security-passport-checking application so far, so it requires user cooperation. FIA includes 6,470 videos covering total of 180 different subjects. However, in this dataset there is only one subject per video.

Author	Algorithm	Database	Accuracy
Kim et al. [8]	Visual constraints using generative and discriminative model	Honda/UCSD [5]	100% rank 1 identification accuracy accuracy
Bhatt et al. [9]	Clustering based re-ranking and fusion	YTF [6]	80.7% verification accuracy at 19.4% EER (Equal Error Rate)
Goswami et al. [10]	Memorability based frame selection with deep learning	PaSC [7]	93.4% verification accuracy at 1% FAR (False Accept Rate)
		YTF [6]	61.5% verification accuracy at 1% FAR
Taigman et al. [11]	Convolution Neural Network	YTF [6]	91.4% rank 1 identification accuracy accuracy
Huang et al. [12]	Projection Metric Learning, Grassmannian Graph-embedding Discriminant Analysis	YTF [6]	70.4% verification accuracy
		PaSC [7]	43.63% verification accuracy at 0.01% FAR

Table 1.1: Recent publications on video face recognition

2. Honda/UCSD [5]:

   Honda/UCSD dataset [5] serves the dual purpose of face tracking as well as face recognition. The dataset has been created in constrained manner and with user acquaintance. For the predefined protocol, the baseline results are up to 93% and the best performance reported [8] is 100%.

3. ChokePoint [14]:

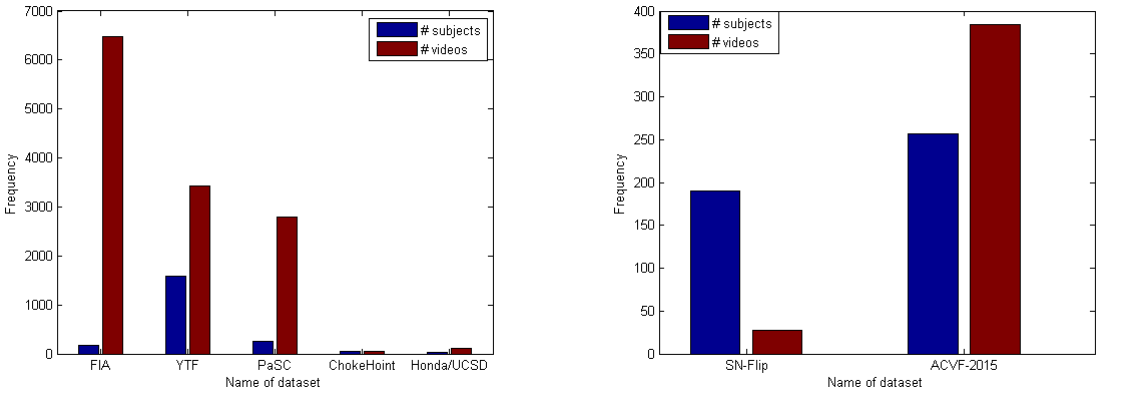
   ChokePoint [14] database is designed to deal with person identification/verification under real-world surveillance conditions using prevailing technologies. It has 48 videos pertaining to 54 subjects.

4. YouTube Faces (YTF) [6]:

   The YTF dataset has been created by Wolf et. al. [6] with the primary purpose of studying face recognition in unconstrained environment. The YTF dataset contains 3,425 videos of 1,595 subjects. The dataset is composed of celebrity videos which are collected from YouTube with the constraint of consisting of one subject per video. It provides predefined protocol sets and current state-of-art results report over 90% accuracy with approximately 9% equal error rate (EER) [11].

5. Point and Shoot Challenge (PaSC) [7]:

   The PaSC database contains videos captured using hand-held and high definition devices. PaSC dataset encompasses 2,802 videos of 265 subjects. For the predefined protocol,



(a) Dataset with single subject in a video (b) Dataset with multiple subject in a video

Figure 1.2: Comparison between existing video datasets

the baseline results are up to 49% verification accuracy at 1% False Accept Rate (FAR) whereas, 93.4% performance has been reported by Goswami et al. [10].

6. SN-Flip [15]:

   SN-Flip database was created by Barr et al. [15], with the requirement of having multiple subjects in one video sequence. It includes 28 videos of 190 subjects.

As shown in Figure 1.2, Honda/USSD, FIA, and YTF databases are confined to have only one subject per video. In real world unconstrained environment, this constraint is difficult to attain. SN-Flip [15] database has multiple subjects in every video; however, all subjects are almost still, thus it may not be very useful for evaluating crowd video matching scenario, i.e., multiple subjects performing diverse actions in a video.

1.2 Research Contributions

It is our assertion that there is significant scope for improving face recognition performance in unconstrained environment, especially the crowd video based scenario. To encourage research in this domain, we have prepared a dataset, which consists of 384 videos with a total of 257 subjects with multiple subjects in each video and unconstrained environment.The key contributions of this research are :

1. 2015 Annotated Crowd Video Face (ACVF-2015) dataset that includes total of 384 videos along with face landmark points for every frame which has one or more face images in it. Along with the videos and landmark points, a set of protocols and end-to-end MATLAB software package are designed to evaluate the performance of face recognition algorithms on this dataset.

2. Face detection baseline is provided by comparing the results of manual annotation and three publicly available codes: 1) Viola Jones [16] face detector (MATLAB open source), 2) HOG descriptor based C++ open source library dlib [17], and 3) face detection aided by fiducial points [18]

3. To establish the face recognition baseline, results are reported with OpenBR [19] and a commercial-off-the-shelf system, FaceVacs.

Chapter 2

Annotated Crowd Video Face Dataset - 2015

The proposed ACVF-2015 dataset contains 384 videos (50,139 frames) of 257 subjects, captured at various locations and each video contains up to 14 subjects. Consent for collecting these videos is taken from all the subjects. Some sample frames are shown in Figure 2.1. Typically, in all the videos, subjects appear in groups. Therefore, in almost all the video frames there are more than one subjects. The recordings are made using handheld devices without mounting on any tripod or similar structure. The dataset details are described below and a summary is provided in Table 2.1.

2.1 Device Details

In this research, data is collected with the help three portable handheld devices having different resolutions. Details of the devices are given below:

1. Nikon Coolpix S570: resolution 640×480
2. Sony handycam DCR-DVD910E: resolution 2304×1296
3. Iphone (4s and 5c) resolution 1920×1080

Device (Resolution)	# Videos	# Frames	# Subjects	# Subjects/Video			# faces
				Min	Max	Avg	Ground Truth
Device I (640×480)	158	21,671	177	1	14	2.5	27,896
Device II (2304×1296)	150	20,007	170	1	10	2.1	23,506
Device III (1920×1080)	76	8,461	106	1	8	2	10,557
Total	384	50,139	257	1	14	2.2	61,595

Table 2.1: Details of the Annotated Video Crowd Dataset - 2015

# video	1	2	3	4	≥ 5
#subjects	73	56	38	29	61

Table 2.2: Number of videos per subject in the ACVF-2015 dataset. For example there are 29 subjects appearing in exactly 4 videos

In this report, these three devices are referred to as Device I, Device II and Device III, respectively. The different devices lead to varying quality in captured videos. The selection of these devices also introduces cross-sensor and cross-resolution covariates in the database. Hence, one can use this dataset for cross resolution face matching problem.

2.2 Manual Annotation, Face Detection and Normalization

Manual annotation has been performed for every visible face present in the frames of every video. Four POIs *(Points Of Interest)* of frontal faces are annotated: Centers of eyes, nose tip, and center of lips, along with subject IDs of that face. This procedure is followed for each video and each frame. If any of the points are not visible due to occlusion or pose variation, the remaining points are annotated. Canonical face frame is obtained from these points. We utilize the publicly available library which uses face detection aided with fiducial points [18], for face detection and cropped faces of size 125×160 are obtained. To check the performance of any system, if a manually annotated points falls into the detected face frame then it is counted as correctly detected face and remaining faces are treated as incorrect detections for that system. As mentioned in Table 2.1, total 61,595 face images are detected from 50,139 frames in 384 videos. Each registered output face image is named using the following convention.

Figure 2.1: Sample frames from the Annotated Video Crowd Face Dataset - 2015. Multiple people appear together in each video along with subjects appearing in indoor unconstrained environment. The videos are captured using three different devices with different sensors and resolutions. The videos are captured while subjects are walking through a passage or passing through doors.

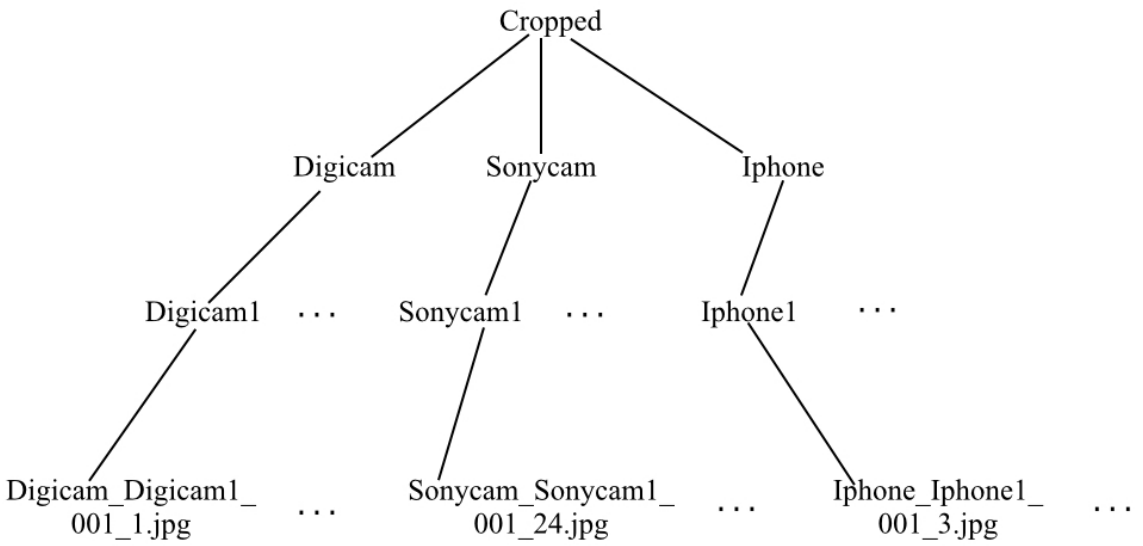


Figure 2.2: Directory structure of the cropped face images provided as part of dataset package

DeviceName_VideoID_FrameNo_SubjectID.jpg

Moreover, registered face images are provided in the /Cropped/DeviceName/VideoID directory for easier access. A sample directory structure is shown in Figure 2.2 and some example annotations with detection are shown in Figure 2.3.

2.3 Application Scenarios for ACVF-2015 Dataset

As mentioned earlier, the ACVF-2015 dataset focuses on unconstrained face recognition (to be precise, face verification) with multiple subjects in a video or image. In this scenario, subjects from one video are matched with the subjects of another video. The gallery set is defined in terms of a *set of videos*. The gallery and probe sets consist of different subjects. With respect to real world applications, this scenario can have three types of evaluations: frame to frame matching, video to frame or frame to video matching, and video to video matching.

- Frame-to-Frame matching:

  In this evaluation setting, every frame of gallery set is matched against every frame of probe set to get score matrix. If gallery set has total m face images match with total n face image of probe set then total size of score matrix will be $m \times n$.

Figure 2.3: The annotation and face detection on an example frame. Points Of Interest(POIs) marked for three faces, whereas the face detection algorithm detects two faces. POIs that are surrounded by each face box are used to assign ground-truth subject IDs with each extracted face. Also, there are some failures in detection cases

- Video-to-Frame Matching:

  In this, face frame of probe set is compared against every video in the gallery set. A set of scores is obtained by comparing each probe face image with all the face frames in gallery video. If gallery video has total q subjects then the set of scores is divided into q subsets. Aggregation is performed for each score subset to get a match score between a probe face image (frame) and a gallery subject. Therefore, comparison of a probe video consisting of m face images (frames) against a gallery video consisting of q subjects, results in $m \times q$ match scores.

- Video-to-Video Matching:

  In this evaluation technique, a set of probe videos is compared with the set of gallery videos. Each probe face image is compared with all the face images in a gallery video and for every match pair, scores are aggregated. Let probe video have m subjects and gallery video have n subjects then $m \times n$ scores will be obtained.

2.4 Evaluation Protocol and Package

The package is designed to make the overall evaluation process as easy as possible. Independent packages are created for face detection and recognition.

2.4.1 Face Detection Evaluation Package

The evaluation package for face detection consists of end-to-end MATLAB code to evaluate the accuracy of face detection algorithms.
To use this code, the user has to provide the file name and rectangle co-ordinates in a file to the package. The format of the input file is shown below:

DeviceName_VideoID_FrameNo_SubjectID.jpg X_Coordinate Y_Coordinate Width Height

The evaluation code provides the accuracy for the corresponding system.

2.4.2 Face Recognition Evaluation Package and Protocol

- Total 10 different disjoint training and testing sets have been created in which average number of videos in training set is 100 with around 115 subjects. The same protocol is followed for testing sets, which has average 240 videos with around 135 subjects in it.
- Each testing set is further divided into gallery and probe sets without any overlapping subject IDs in the test sets.
- The details of the video IDs and subject IDs of each set are provided in the evaluation package.
- MATLAB code for end-to-end evaluation has been created and is provided in the package. This code requires score matrices for each testing set as input and performs all necessary calculations and gives a combined ROC curve which is used for comparing the results.

Chapter 3

Unconstrained Face Detection and Recognition

3.1 Motivation

Face detection and recognition encompass areas such as computer vision, image analysis, law enforcement, surveillance, biometrics, and security. Some face detection and recognition applications are:

- Biometrics: immigration, passport fraud, image search.
- Consumer product: phone unlocking system, desktop login.
- Law enforcement: video surveillance for suspect identification and tracking their movement.
- Smart card applications: passport, voter ID, driving licence.

The problem of face detection is similar to object detection in which we need to find objects and segment them from the image. Face detection is the first phase of the face recognition pipeline; therefore, it is necessary to achieve high performance in this phase. In constrained environment, if image quality is good and only frontal faces are present in the image, it is easy to detect the

Figure 3.1: Challenges for face recognition and detection including pose variation, occlusion, lighting condition, and poor image quality.

face using existing algorithms. However, in unconstrained environments, face detection is still a challenging task. Similarly, face recognition in both videos and images have achieved acceptable level of performance in constrained environments. However, in unconstrained environment it is challenging due to uncontrolled behavior of the person and environment; for instance, varying illumination condition, occlusion, and pose variation affect the appearance of a person. Unlike still images, videos provide abundant information where both spatial as well as temporal information can be used for recognition. This chapter, therefore, focuses on unconstrained face detection and recognition in video.

3.2 Challenges in Unconstrained Face Detection and Recognition in Videos

While videos provide ample amount of information that can be used for face recognition , they also pose unique challenges. Figure 3.1 shows some of the challenges which make face detection and recognition task difficult specifically.

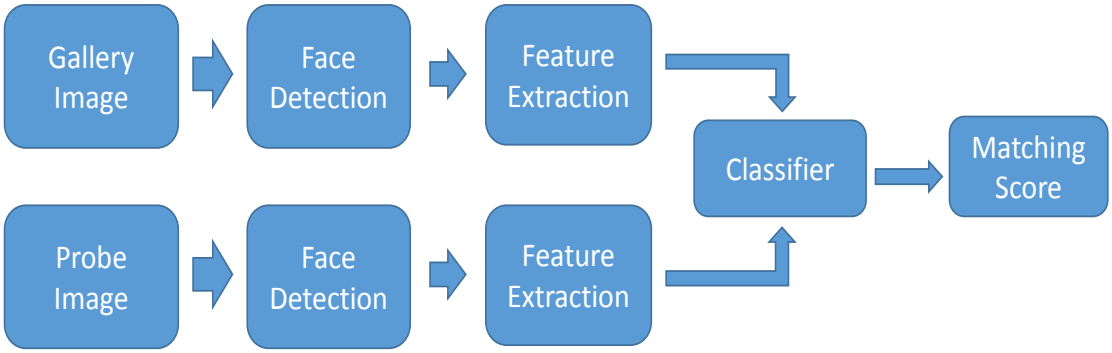


Figure 3.2: Face recognition pipeline

- In unconstrained environment, it is not easy to get acceptable level of performance in face detection due to variation in illumination, pose, and occlusion.
- In low quality videos, it is difficult to differentiate between the face and the background; therefore, both face detection and recognition are challenging.
- Activities performed by different persons in a video can cause occluded face images. Such occlusions make detection and recognition tasks difficult for automated processing.
- At times, it is possible that the subject is at a distance from the sensor and therefore, face area can be small. Such variations make detection as well as recognition very difficult.

3.3 Literature Review

As shown in Figure 3.2, a conventional face verification pipeline includes face detection, feature extraction, and matching with a decision. In this section, some of the existing face detection and recognition approaches are briefly described.

3.3.1 Face Detection Approaches

Face detection is the first phase of face recognition pipeline; hence, it is mandatory to achieve high performance in this phase. Yang et al. [11] classify existing techniques into main four categories [20]:

1. Skin color based technique [21]

   For constrained environment and background, skin color based techniques can be used to find the face segment in the image. Conventional skin color based techniques, first transform the color image into HSI or YCbCr color channels. Then, skin region with the help of suitable threshold value and remove other body parts by using morphological image processing. Existing research has shown that face detection by skin color is affected by several parameter including lighting conditions ethnicity, variations.

2. Template matching technique [22], [23]

   Template matching techniques are used for finding a small part of image which is similar to the template image. In these approaches template face image is utilized to detect a portion of the face in the image. Face location is then determined based on correlation values. This technique is not robust to out-of-plane rotation, and different facial expression and poses variations. The size of face and different lighting conditions are also challenging issues in this technique.

3. Appearance based technique [24]

   The basic idea of appearance based technique is to collect face and non face images to train a classifier and in testing phase, detect a face based on the classifier. Eigenfaces [25] and Fisherfaces [20] are appearance based techniques; however, illumination and pose variations are two important challenges in appearance based techniques.

4. Feature invariant based technique

   Feature invariant algorithms use local features to represent a face. There are different kinds of features for different techniques: 1) Haar cascade features are used by Viola and Jones [16], which is one of the most popular technique and is currently used in OpenCV and MATLAB face detectors. 2) Scale Invariant Feature Transform (SIFT) proposed by David Lowe [26] is a robust local feature descriptor. It helps to find different face features over different rotations and scales. SIFT feature based techniques are also able to provide accurate detection over illumination changes and certain ranges of affine transformation.

3) Histogram of Oriented Gradient(HOG) [27] descriptors are also used for face detection as they are robust to illumination, rotation, and pose variations. Dlib open source library [17] for object detection uses HOG as the integral feature.

3.3.2 Face Recognition Approaches

Many algorithms have been proposed for face recognition. Zhao et al. [3] have divided existing approaches into three parts:

1. *Subspace methods*: These approaches use the entire face image as raw input. Some of the popular methods are: Principal Component Analysis (PCA) [24], Independent Component Analysis (ICA) [28], and Linear Discriminant Analysis (LDA) [29].

2. *Feature based matching methods*: Local features of face components are extracted and their statistical information with the location is provided as input. Local Feature Analysis (LFA) [30], and Gabor Filter (GF) [31] are popular feature based techniques.

3. *Texture Methods*: These methods use texture information that faces exhibit for feature extraction and matching. Some of the popular texture based methods are: Histogram of Oriented Gradients (HOG) [27], and Local Binary Patterns (LBP) [32].

Video based face recognition approaches typically use temporal information in combination with traditional features. The survey on video based face recognition by Barr et al. [33] categorizes existing approaches as set based and sequence based approaches. Further, with the advent of deep learning and dictionary learning approaches, face recognition algorithm also utilize these approaches. Therefore we categorize video face algorithms into three classes:

1. *Set based approaches*: These approaches treat videos as unordered collection of images and take advantage of the multitude of observations. These methods differ in terms of information fusion over observation before and after matching. Set based approaches do not perform well when facial expression changes. Manifold-manifold distance [34], linear dynamic modeling [35], manifold density divergence [36], sparse approximated nearest

Figure 3.3: (a) Examples of accurately detected faces corresponding to each of the three devices. (b) Examples of inaccurate face detection due to partial face and presence of extra non-face/background regions, and (c) sample images demonstrating falsely detected faces which are discarded based on POI annotations

point [37], set-set similarity [6], and image set alignment [38] are some techniques based on set based approach.

2. *Sequence based approaches*: These approaches use spatial and temporal information of a video to increase the recognition rate in bad viewing conditions. These methods use both appearance and motion information from the video to make recognition decision. Sequence based methods can be used for face tracking and it gives better results in recognition in degraded condition where face image is deformed or occluded. The following are some sequence based approaches: sequential importance sampling [39], visual constraints using generative and discriminative models [8], and adaptive Hidden Markov Model [40].

3. *Deep learning and Dictionary based approaches*: In generative approach for deep learning and dictionary learning based face recognition in videos, sequence-specific atoms are created and used for feature extraction and matching. Distance measure is used for recognition. Clustering based re-ranking and fusion [9], video dictionaries [38], rank aggregation [41] are dictionary based techniques, whereas DeepFace [11] and MDLface [10] are examples of deep learning approaches.

3.4 Face Detection and Recognition: Approaches Used

In order to perform baseline evaluation on the ACVF-2015 dataset, we have used several approaches for face detection and recognition. This section explains the approaches used in the evaluation.

3.4.1 Face Detection

Face detection in ACVF-2015 has been performed manually and results are compared with publicly available systems: HOG descriptor based Dlib open source library, Haar feature based Viola Jones [16] face detector, and face detection aided by fiducial points [18]. Manually annotated information is presented in the previous chapter. Figure 3.3 shows some samples of correct detection, inaccurate detection, and wrong detection, for different algorithms.

3.4.1.1 Viola Jones Face Detector [16]

Viola Jones object detector is capable of processing images rapidly while achieving high detection rate. It is robust as it has higher detection rate and low false positive rate. Viola Jones algorithm has four stages: Haar feature selection, creation of an integral image, Adaboost training, and cascaded classification. Viola Jones algorithm uses a cascade of weak classifiers to make a strong classifier:

$$h\left(x\right) = sign(\sum_{j=1}^{M} \alpha_j h_j(x)) \tag{3.1}$$

where, $h(x)$ is a strong classifier obtained from the set of weak classifiers $h_j(x)$ and M is the total number of possible features in an image sub-window. Each weak classifier is a threshold function based on the feature f_j and,

$$h_j(x) = \begin{cases} -s_j & if\ f_j < \theta_j \\ s_j & otherwise \end{cases} \tag{3.2}$$

The threshold value θ_j, the polarity $s_j \in \pm 1$, and coefficients α_j, are determined during training.

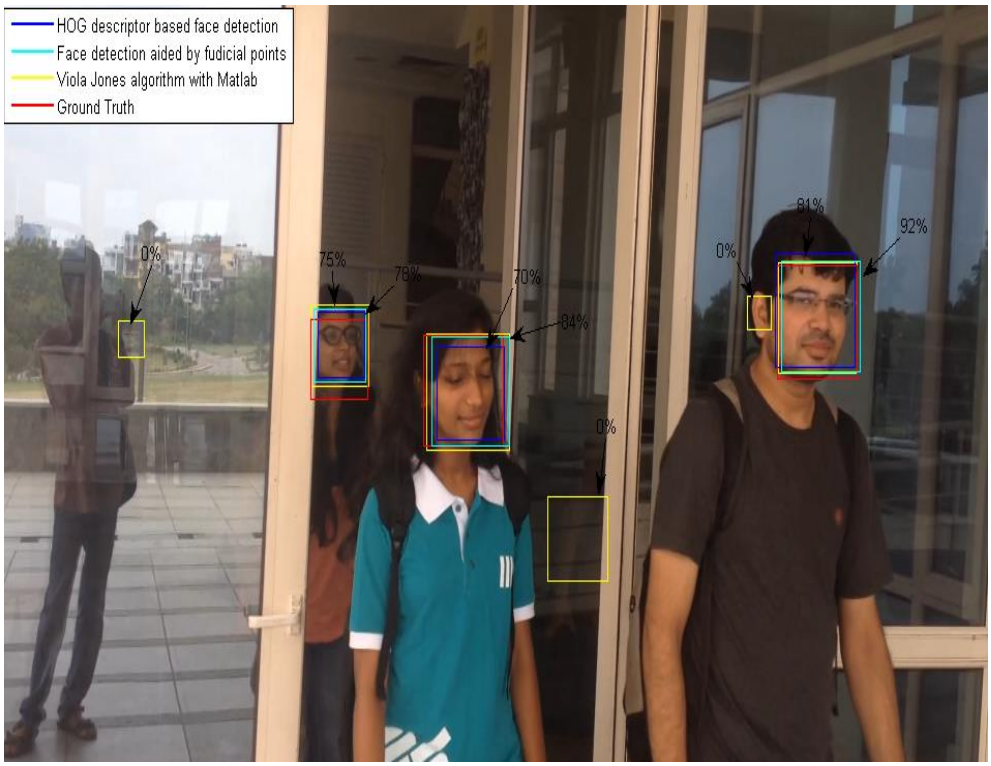


Figure 3.4: Comparison among ground truth, Viola Jones face detector, HOG feature based face detection, and face detection aided by fiducial points with overlapping percentage

3.4.1.2 Face detection aided by fiducial points [18]

Face detection aided by fiducial points [18] uses frontal face detector [16] and a variant of the probabilistic parts-based pictorial structure model which is used to model the joint position and appearance of facial features. It uses a conservative threshold value to achieve low false positive rates. As an output, it gives a face box with approximate location of nine landmark points including two eye corners, nose corners, nose point, and lips corners. As described in [18], to locate features, a generative model of the feature positions combined with a discriminative model of the feature appearance is applied. It is assumed that the appearance of each feature is independent and is modelled differently by training a classifier which uses a variation of the AdaBoost algorithm.

3.4.1.3 Face Detection based on Histogram of Oriented Gradient (HOG)

We use the publicly available open source C++ based library Dlib, which has a module for object detection. It is a Histogram of Oriented Gradient (HOG) [27] based object detector. Dlib's tool makes training HOG detectors very fast and easy. To train Dlib for object detection, all we need to do is just provide set of input images with the bounding box of the object, which we want to detect and it train the model within few seconds. This approach primarily uses HOG descriptor as an input feature. In HOG descriptors the occurrences of gradient in localized portion of image and contours [27] are counted. The algorithm consists of the following steps:

1. *Gradient Computation*: Apply filter mask on the image. The most common method is to apply the 1-D centered, point discrete derivative mask in one or both of the horizontal and vertical directions:

$$D_x = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \; and \; D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{3.3}$$

2. *Orientation Binning*: Create cell histogram. Each pixel in the cell casts a weighted vote for the orientation based histogram channel.

3. *Descriptor Blocks*: Group the cells into larger blocks.

4. *Block Normalization*: Normalize the blocks. Normalization can be performed using one of the following methods.

$$L1 - norm : f = \frac{\nu}{\|v\|_1 + e} \tag{3.4}$$

$$L2 - norm : f = \frac{\nu}{\sqrt{\|v\|_2^2 + e^2}} \tag{3.5}$$

$$L1 - sqrt : f = \sqrt{\frac{\nu}{\|v\|_1 + e}} \tag{3.6}$$

5. *Similarity Measure*: Similarity between vectors is computed using euclidean metric or cosine similarity. If P = (p_1,p_2,...,p_n) and Q= (q_1,q_2,...,q_n) are two vectors then the

euclidean distance (d) is computed as follows:

$$d = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} \tag{3.7}$$

The cosine similarity (C) is computed as follows:

$$C = \frac{P.Q}{\|P\| \, . \, \|Q\|} \tag{3.8}$$

Figure 3.4 shows output rectangle of each approach on the example frame. The percentage above the rectangle shows the intersection percentage with the manually annotated rectangle. The procedure for obtaining the intersecting percentage is:

- Obtain the coordinates of the rectangle for manual annotation and automatic detection.
- Find the intersecting of the two detected detections.
- Calculate the intersection area percentage with the following equation

$$Percentage\, of\, overlap = \frac{Area\ of\ ground\ truth\ rectangle\ \cap Area\ of\ detected\ rectangle}{Area\ of\ ground\ truth\ rectangle\ \cup Area\ of\ detected\ rectangle} \times 100\% \tag{3.9}$$

3.4.2 Face Recognition

Baseline face recognition and evaluation on the ACVF-2015 dataset has been performed using Open Source Biometrics Recognition (OpenBR) [19] and FaceVacs.

3.4.2.1 Open Source Biometrics Recognition (OpenBR)

OpenBR is an open source platform for face recognition in which new grammar/language is built to support interfacing of new algorithms. For each word in a language, there is a specific plugin that performs a specific image transformation task. A combination of these words is used for template enrollment and comparison. The default face recognition algorithm in OpenBR is based on the Spectrally Sampled Structural Subspaces Features (4SF) [42] algorithm which is

a statistical learning based algorithm. The statistical learning of 4SF allows OpenBR to train on specific matching problems. OpenBR uses OpenCV Viola Jones [16] object detector; for eye detection, a custom C++ port of Average of Synthetic Exact Filters (ASEF) [43] approach is used. Rotation and scaling on the detected face images is performed with the help of affine transformations depending on detected eye locations. For face representation, both LBP [32] and SIFT [26] are used. In LBP, histograms are extracted in an 8×8 sliding window and for SIFT descriptor, 10×10 grid is used. Further L_1^{byte} distance is introduced which achieves state of the art matching speed with negligible impact on matching accuracy. The script used for obtaining matching score by matching two images is given below.

$ br algorithm FaceRecognition compare one.jpg two.jpg

where, one.jpg and two.jpg are two images to be matched.

3.4.2.2 FaceVacs (Commercial System)

FaceVacs is one of the best commercial-off-the-shelf face recognition system. Although it can be used only as a black box, evaluation with it provides insights into the challenging nature of the problem.

The above described methods are implemented for creating baseline results for our dataset. The experimental protocols and results obtained by using these approaches are discussed in the next chapter.

Chapter 4

Experimental Results

Baseline evaluations are performed on the set of predefined protocols discussed in Chapter 2. For face detection and recognition, individual results are reported on the ACVF-2015 dataset by using approaches explained in the previous chapter.

4.1 Face Detection

Benchmarking has been performed for face detection by using three existing techniques: Haar feature based face detector [16], HOG [27] descriptor based dlib C++ open source library [17] and face detection with fiducial points [18]. We compared the result of each technique with manually annotated detection results. Each of the three detectors gives a rectangle around the face region detected by them. In our experiments, face image is considered as detected only if the automatically detected face rectangle intersects with the manually annotated face rectangle. If the intersection is empty then we consider the image as not detected. Table 4.1 shows the comparison and detection rate for each detector with the percentage of matching. Figure 4.1 shows the comparative number of detection in stacked bar format. Figure 4.2 shows the graphical representation of face detection results. In HOG feature based open source library, approximately 10,000 faces are detected, for which the percentage of overlap is 0-10%. Similarly, in face detection with feducial points and Viola Jones face detector, detected number of faces

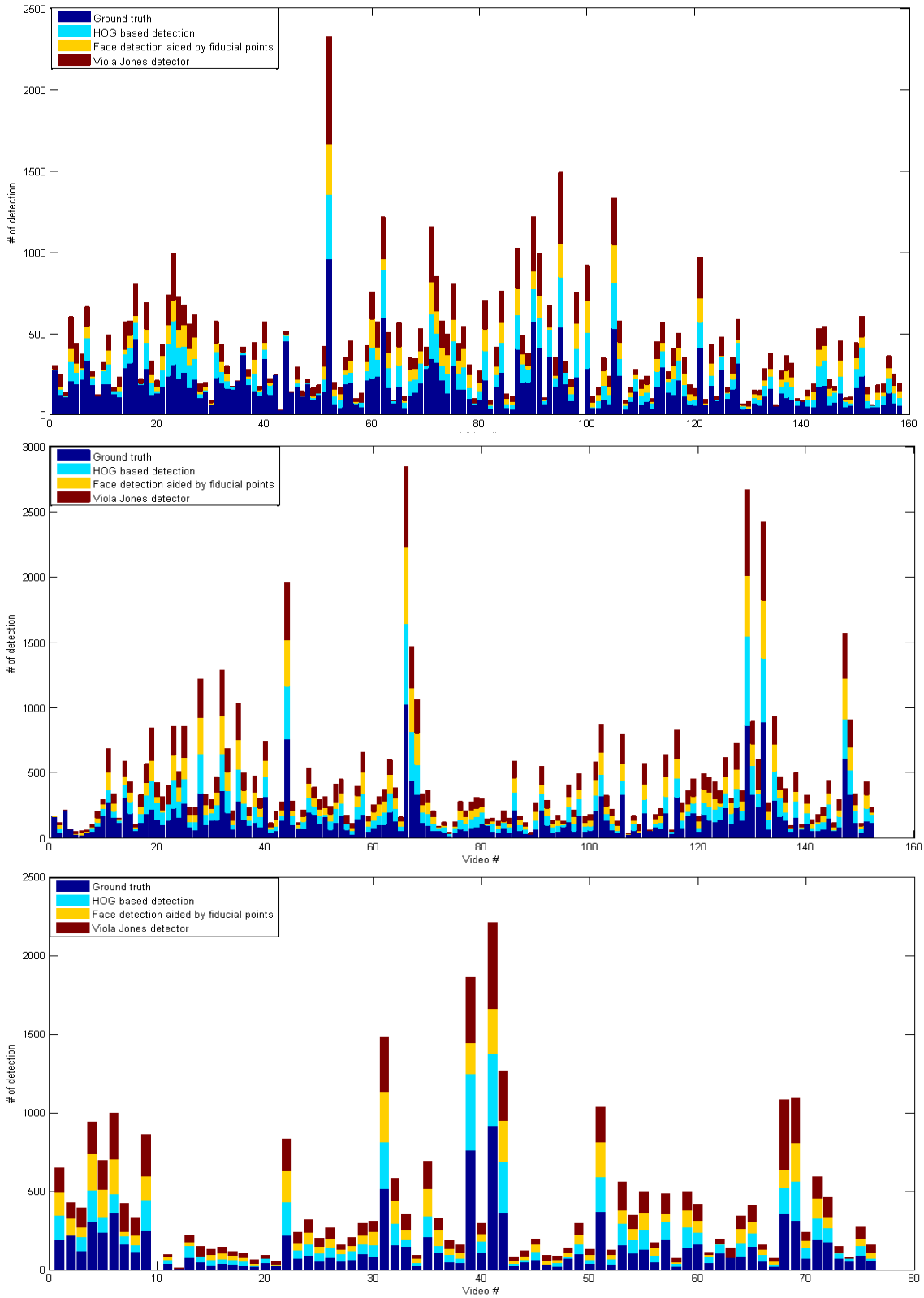


Figure 4.1: Representing the results of face detection. Each stacked bar represents an individual video number for specific device. Different colors of stacked bars represent the number of ground truth faces and the number of detected faces for different systems.

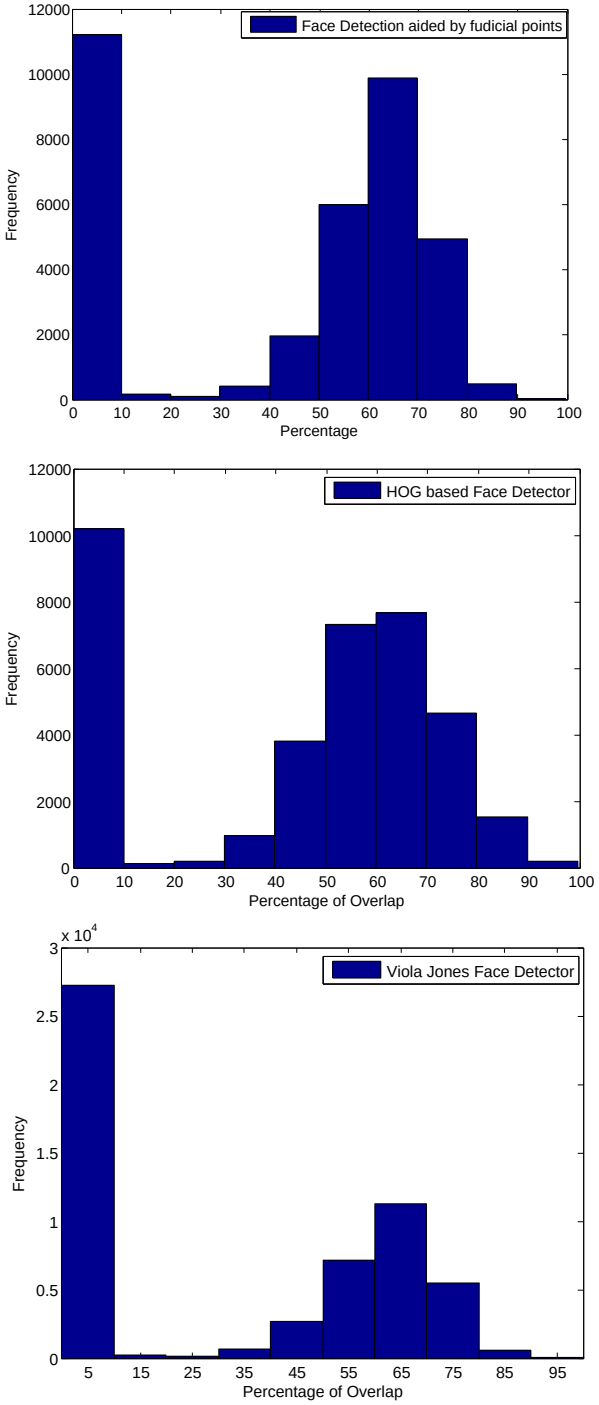


Figure 4.2: Face detection results for individual system. Number of correct detection by the system with the percentage of overlapping faces with the ground truth

Table 4.1: Detection results for three individual system

Detection results for Viola Jones detector

Algorithm	Matching (%)	Detected	Correctly Detected	Falsely Detected	Not Detected
Viola Jones	$\geq 25\%$	55,442	27,846 (45.57%)	10,017 (16.39%)	23,256 (38.05%)
	$\geq 50\%$		24,473 (40.04%)	13,390 (21.91%)	
	$\geq 75\%$		2,280 (3.73%)	35,583 (58.22%)	

Detection results for Face detection aided by fiducial points

Algorithm	Matching (%)	Detected	Correctly Detected	Falsely Detected	Not Detected
Face detection aided with fiducial points	$\geq 25\%$	35,059	26,258 (42.96%)	2,301 (3.76%)	32,560 (53.27%)
	$\geq 50\%$		21,088 (34.50%)	7,471 (12.22%)	
	$\geq 75\%$		3,299 (5.40%)	25,260 (41.33%)	

Detection results for HOG descriptor based detector

Algorithm	Matching (%)	Detected	Correctly Detected	Falsely Detected	Not Detected
HOG descriptor	$\geq 25\%$	36,611	23,651 (38.70%)	9,204 (15.06%)	28,264 (46.24%)
	$\geq 50\%$		21,231 (34.73%)	11,624(19.02%)	
	$\geq 75\%$		1,883 (3.09%)	30,972 (50.67%)	

are 11,000 and 27,500 respectively, with 0 to 10% overlap.

Figure 4.3 shows the combined results of the three approaches with the number of faces not detected by them. These results show that with $\geq 50\%$ overlaps with manual annotation, all the face detectors yield face detection rate of approximately 0.4. Among these detectors, Viola Jones face detector has high detection and false detection rates whereas lowest false detection rate is obtained by face detection aided with fiducial points [18].

4.2 Face Recognition

Face recognition on ACVF-2015 dataset has been performed using Open Source Biometrics Recognition (OpenBR) [19] and a commercial-off-the-shelf system, FaceVacs, with a set of defined evaluation protocols. The verification performance is reported in terms of receiver operating characteristic (ROC) curve. The ROC curves obtained for the gallery-probe sets are combined into one curve using vertical averaging. ROC curves for each of the system for manually annotated faces are also reported. The key observations from the ROC curves shown in Figure 4.4 are as follows:

- FaceVacs appears to outperform OpenBR. However, at higher FARs, the performance

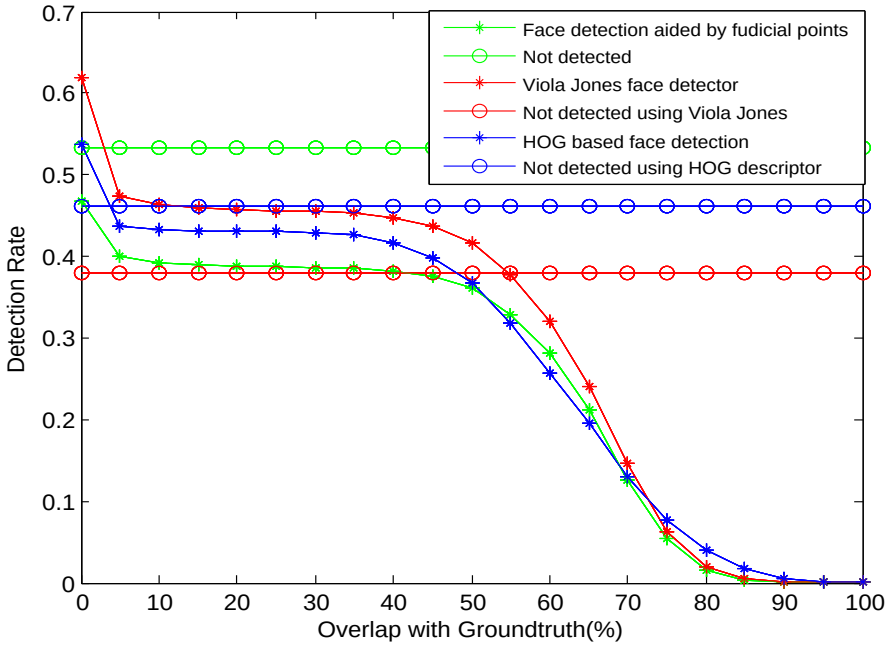


Figure 4.3: Combined face detection result which shows detection rate versus percentage of overlap.

FAR	**GAR**			
	FaceVacs		**OpenBR**	
	Automatic	**Manual**	**Automatic**	**Manual**
0.001	0.027	0.023	0.00	0.00
0.01	0.089	0.059	0.016	0.01
0.1	0.316	0.192	0.21	0.11

Table 4.2: Face verification performance for OpenBR and FaceVacs at various FAR

difference is not significant.

- Score aggregation for video-to-video and video-to-frame matching is performed using two strategies: mean and max. Since both the systems provide similarity scores, the max strategy translates to selecting the scores corresponding to the best match. Both the systems suffer significantly in video-to-video matching using mean aggregation strategy and the best performance is observed with video-to-video matching with max aggregation strategy. This result underlines the importance of frame selection [10]. Table 4.2 shows the performance of all scenarios at different FARs. In all four scenarios, at 0.01 FAR, the best verification rate achieved is only 0.08. This poor performance indicates the complexity of

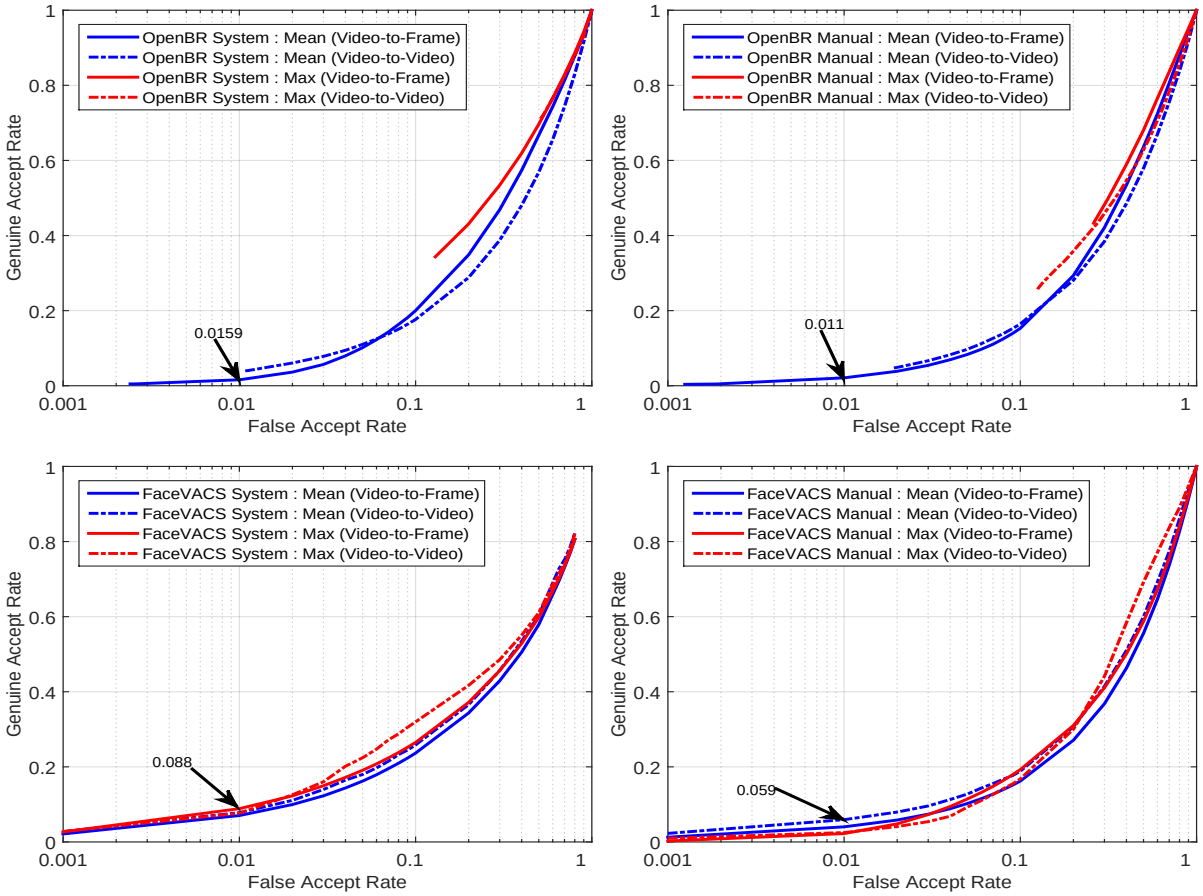


Figure 4.4: Baseline results for face recognition

the problem as well as limitations of current systems.

Chapter 5

Conclusion

Face recognition from video in unconstrained environment has attracted a lot of research interest due to its various applications. Multiple frames in a video provide temporal and intra-class variations that can be leveraged for efficient face recognition. Due to lack of a dataset which has multiple face images in a single frame, we proposed a new dataset termed as ACVF-2015 and performed baseline experiments with existing systems. Results for face detection and recognition on the ACVF-2015 dataset show that popular commercial and open source systems do not perform well on crowd videos in uncontrolled settings. We also provide a platform to evaluate and benchmark face detection and face recognition algorithms in challenging crowd scenarios. This will help in advancing the state-of-art for both detection and recognition of faces.

Bibliography

[1] P. Phillips, P. Grother, and R. Micheals, "Evaluation methods in face recognition," in *Handbook of Face Recognition, second edition*, Springer New York, 2005, pp. 329–348.

[2] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2007.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[4] H. S. Bhatt, "Emerging covariates of face recognition," PhD thesis, IIIT-Delhi, 2014.

[5] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Journal of Computer Vision Image Understanding*, vol. 99, no. 3, pp. 303–331, Sep. 2005.

[6] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[7] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[8] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[9] H. S. Bhatt, R. Singh, and M. Vatsa, "On recognizing faces in videos using clustering-based re-ranking and fusion," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1056–1068, Jul. 2014.

[10] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa, "MDLFace : Memorability augmented deep learning for video face recognition," in *IEEE/IAPR International Joint Conference on Biometrics*, 2014, pp. 1–7.

[11] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[12] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015.

[13] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU face in action (FIA) database," in *Analysis and Modelling of Faces and Gestures*, Springer, 2005, pp. 255–263.

[14] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 81–88.

[15] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn, "Active clustering with ensembles for social structure extraction," in *IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 969–976.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[17] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[18] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Journal of Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009.

[19] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge, "Open source biometric recognition," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[20] P. N. Belhumeur, P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transaction Pattern Analysis Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[21] S. Singh, D. S. Chauhan, M. Vatsa, and R. Singh, "A robust skin color based face detection algorithm," *Tamkang Journal of Science and Engineering*, vol. 6, no. 4, pp. 227–234, 2003.

[22] J. Wang and H. Yang, "Face detection based on template matching and 2DPCA algorithm," in *Congress on Image and Signal Processing*, vol. 4, 2008, pp. 575–579.

[23] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[24] K. Delac, M. Grgic, and P. Liatsis, "Appearance-based statistical methods for face recognition," in *International Symposium ELMAR*, 2005.

[25] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[27] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[28] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by Independent Component Analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.

[29] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," in *International Conference on Audio and Video Based Biometric Person Authentication*, 1997, pp. 127–142.

[30] S. Arca, P. Campadelli, and R. Lanzarotti, "A face recognition system based on local feature analysis," in *International Conference on Audio and Video based Biometric Person Authentication*, 2003, pp. 182–189.

[31] H. Wu, Y. Yoshida, and T. Shioyama, "Optimal gabor filters for high speed face identification," in *IEEE International Conference on Pattern Recognition*, vol. 1, 2002, pp. 107–110.

[32] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[33] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 5, 2012.

[34] R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao, "Manifold–manifold distance with application to face recognition based on image set," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[35] G. Aggarwal, A. Chowdhury, and R. Chellappa, "A system identification approach for video-based face recognition," in *IEEE International Conference on Pattern Recognition*, 2004, pp. 175–178.

[36] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 581–588.

[37] Y. Hu, S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 121–128.

[38] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2626–2633.

[39] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," in *International Journal of Computer Vision and Image Understanding*, vol. 91, 2003, pp. 214–245.

[40] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, pp. 340–345.

[41] H. Bhatt, R. Singh, and M. Vatsa, "On rank aggregation for face recognition from videos," in *IEEE International Conference on Image Processing*, 2013, pp. 2993–2997.

[42] B. Klare, "Spectrally sampled structural subspace features (4SF)," Department of Computer Science, Michigan State University, Tech. Rep., Sep. 2011.

[43] S. H. Park and J. H. Yoo, "A new implementation method of ASEF for eye detection," in *International Conference on Computing and Convergence Technology*, Dec. 2012, pp. 1034–1037.