Face and Gender Classification in Crowd Video

Priyanka Verma

IIIT-D-MTech-CS-GEN-13-100
July 16, 2015

Indraprastha Institute of Information Technology
New Delhi

Thesis Advisors

Dr. Richa Singh
Dr. Mayank Vatsa

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

Keywords : Face Recognition, Gender Classification, Crowd database

Certificate

This is to certify that the thesis titled **"Face and Gender Classification in Crowd Video"** submitted by **Priyanka Verma** for the 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 her under our guidance and supervision at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Richa Singh

Dr. Mayank vatsa

Indraprastha Institute of Information Technology, Delhi

Abstract

Research in face and gender recognition under constrained environment has achieved an acceptable level of performance. There have been advancements in face and gender recognition in unconstrained environment, however, there is significant scope of improvement in surveillance domain. Face and gender recognition in such a setting poses a set of challenges including unreliable face detection, multiple subjects performing different actions, low resolution, and sensor interoperability. Existing video face databases contain one subject in a video sequence. However, real world video sequences are more challenging and generally contain more than one person in a video. This thesis provide the annotated crowd video face database with more than 200 videos pertaining to more than 100 individuals, along with face landmark information and gender annotation to encourage research in this important problem. We provide two distinct use-case scenarios, define their experimental protocols, and report baseline verification results existing on two face recognition systems, OpenBR and FaceVACS. Gender classification is also performed on this database and the results are reported using OpenBR along with a combination of different feature extractors with SVM classification. The results show that both the baseline results do not yield more than 0.16 genuine accept rate at 0.01 false accept rate. A software package is also developed to help researchers evaluate their systems using the defined protocols.

Acknowledgments

Towards the completion of my Masters degree, I would like to pay my heartily tributes to people who contributed in many ways. After expressing gratitude towards God and my loving parents, I would like to thank my advisors Dr. Richa Singh and Dr. Mayank Vatsa for their support and guidance throughout the journey. Their constant guidance and input have helped me prosper towards a more confident and improved personality. They made great efforts in supporting me through all possible ways. Their advice has always served me gain more knowledge and in selecting better options. I would like to specially mention Tejas Dhamecha and Mahek Shah, without whose support this work would not have to be done. I would also like to thank my friends, especially Ajay Malik and Nidhi Agarwal for being there as a constant source of inspiration and motivating me in the worst as well as best times in IIITD. This section can not be complete without a vote of thanks to academic department for their help and never ending support.

The research is partially funded by Department of Electronics and Information Technology, Government of India

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Research Motivation

Research in face recognition has matured enough [12, 21, 34], and now it can be used in actual applications such as face tagging, mobile phone unlocking, and time-attendance. According to the Multiple Biometric Grand Challenge [19] (MBGC) and Point and Shoot Challenge [7] (PaSC) evaluation reports, in controlled environment, state-of-the-art face recognition systems achieve up to 0.997 verification rate at 0.001 false accept rate (FAR) [7,19]. However, in an uncontrolled environment such as in surveillance camera videos, face recognition remains very challenging and state-of-the-art performance reduces significantly. The unconstrained environment would include (but is not limited to) acquisition using low cost devices, varying lighting conditions, minimum user co-operation, and presence of multiple subjects within the field of view. The surveillance cameras normally capture videos at low resolution. So, there are very limited pixels that account for the faces. On these challenges to achieve good recognition accuracy existing algorithms, in general, require an inter eye distance of 90 pixels. Moreover, in surveillance videos there are variations in pose and lighting conditions, making face recognition a more challenging task. Figure 1.1 shows these challenges of face recognition in surveillance cameras.

An efficient system that works in unconstrained environment is likely to be useful in multiple applications. One such important scenario is when both gallery (target) and probe (query) are

Figure 1.1: Challenges of face recogniton, a) represents images with variations in illumination, b) represents variations in pose, c) represents faces with noise variations

obtained without requiring user cooperation. For instance, recently in Delhi, a child kidnapping case occurred [3]. In this case only the CCTV footage of suspect is available and we want to match the footage from one camera against other CCTV footages to identify the suspect's movement. As shown in Figure 1.2, gallery and probe images and videos are typically obtained from a surveillance footage that may contain multiple subjects. In order to facilitate law enforcement agencies, it is critical for face recognition research to attain impressive performance in the aforementioned application scenario [9]. Further, these applications also involve addressing emerging covariates [11] of low image quality, varying resolution, and sensor inter-interoperability, along with traditional covariates of pose, illumination, and expression [34] as well as age and weight variations [30].

In recent years, researchers have been working on designing video-based face recognition algorithms [6, 8, 13, 32] to address some of these challenges. Researchers are also interested in using soft biometrics such as gender and ethnicity for recognition task. In comparison to face recognition, gender recognition has been a relatively poorly explored problem. However, automatic

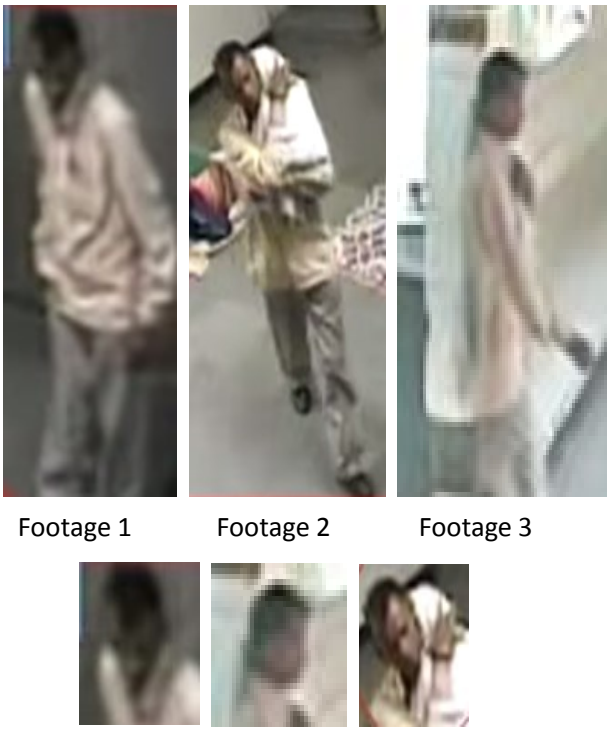


Figure 1.2: A law enforcement application scenario where subjects are matched using surveillance footage only. Top row of the figure shows four frames/images from a child Kidnapping case [3].The bottom row show the face regions of the suspect.

gender recognition is also used in a variety of real world applications, such as human system interactions and in demographic data collection [20]. With respect to face recognition, gender classification can also reduce the search time for face retrieval.

1.2 Literature Review

This thesis focuses on two aspects of unconstrained face recognition: gender classification and face recognition. Below is a brief summary of recent advances in gender classification and face recognition.

1.2.1 Gender Classification

In a biometric pipeline, gender classification is treated as soft information which helps several aspects including indexing, analysis, and utilization as an attribute. In literature, gender classification approaches utilize either pixel intensity values, geometric features, appearance based features or textual features combined with two class classifier (male and female being two classes). Bejios-Calfa *et al.* [5] and Dhamecha *et al.* [14] have shown that linear discriminant approaches can help in achieving over 90% gender classification accuracy. Ng *et al.* [25] summarize a brief literature review of gender classification approaches. The authors suggest that while gender classification approaches have achieved acceptable level of accuracies in controlled environment, uncontrolled settings still pose a significant challenge and a lot of work is required to make the technology applicable for real life applications.

1.2.2 Face Recognition

Algorithms in automatic face recognition can be broadly classified into three categories: subspace based, feature based and texture based approaches. Subspace based approaches such as PCA, LDA, and ICA (and their variants) exploit the observation that face images form low dimensional manifold and this face space can form a weighted representation for recognition purposes. Feature based approaches utilize the geometric properties of faces whereas texture based approaches utilize high and low level frequency texture variations exhibited in face images. Further there are approaches that combine two or more of these paradigms of face recognition and they are known as hybrid approaches. Detailed literature reviews of existing face recognition algorithms are presented in [4, 10, 28]

1.3 Research Contributions

It is our assertion that there is a significant scope for improving face and gender recognition performance in unconstrained environment, particularly in crowd video scenarios where enrollment as well as query videos/images are obtained in unconstrained settings. To encourage research

in this important area, we have prepared a dataset consisting of 201 videos pertaining to 133 subjects which comprises of 96 male and 37 female subjects, where each video contains multiple subjects. The key contributions are:

1. Annotated Crowd Video Face (ACVF) Database-2014 includes videos and frames along with landmarks of faces (two corners of both the eyes, two corners of lips, two corners of nose, and a nose tip) and gender annotation in each frame. 10 times random subsampling based cross validation protocol files and a MATLAB software package for evaluation is also included.

2. To establish the baseline, the results are reported using OpenBR [23] and a commercial-off-the-shelf system, FaceVACS [19]. The results are shown on two different experimental protocols.

3. Gender Classification baseline is established on the ACVF database using multiple classification algorithms and OpenBR [23].

Chapter 2

ACVF-2014 Dataset and Unconstrained Face Recognition

There are many existing datasets containing face images under controlled conditions. However, many real life applications require processing image frames in surveillance videos and therefore, there is a requirement in the face recognition community for a face video dataset that is captured in unconstrained settings. With such a dataset, the benefits of video-based face recognition can be explored. In this research, we created ACVF dataset which consist of videos captured in unconstrained environment with multiple subjects in every frame. Three different devices are selected for data acquisition which introduces cross-sensor and cross-resolution covariates in the database.The proposed ACVF-2014 database contains 201 videos (28,011 image frames) of 133 subjects, captured at various locations, and each video contains up to 14 subjects. Among these subjects there are 96 male subjects and 37 female subjects. On average each subject appears in 2 videos.

2.1 Related Work

Table 2.1 presents a summary of existing datasets, most of which contain videos under constrained environment and lack real world challenges such as variations in pose, illumination, and

Table 2.1: Details of existing video face databases. The proposed database, ACVF-2014, records crowd (multiple subjects) in motion in every video.

Dataset	**Description**	**# Subjects**	**# Videos**
Face In Action [17]	passport checking scenario (constrained), single subject/video	180	6,470
YouTube Faces [32]	unconstrained, celebrity videos, single subject/video	1,595	3,425
PaSC [7]	unconstrained, single subject/video	265	2,802
ChokePoint [33]	unconstrained, fixed camera surveillance, single subject/frame	25	48
SN-Flip [4]	almost still subjects, multiple subjects/frame	190	28
ACVF-2014	unconstrained, hand held devices, multiple subjects/frame	133	201

noise that occur in real world surveillance videos.

1. *Face-In-Action Database*

   FIA [17] was created with focus on a typical border-security-passport-checking scenario, thus expecting user cooperation. In this dataset, videos of 180 participants were collected in indoor as well as outdoor environments. The subjects were guided to mimic a scenario of passport checking. Each video is 20 seconds long and contains only one subject per video.

2. *YouTube Faces Database*

   In 2011, Wolf *et al.* [32] created the YouTube Faces (YTF) database, which focuses on unconstrained face recognition. The database consists of 3425 videos of 1595 celebrities collected from a famous video sharing website YouTube. It provides predefined protocol sets and current state-of-the-art results report around 90% accuracy with approximately 9% equal error rate (EER) [31].

3. *Point and Shoot Challenge Database (PaSC)*

   The Point and Shoot Challenge database [7] contains single subject videos captured using handheld and high definition devices. On the pre-defined protocol, the baseline results are up to 49% verification accuracy at 1% FAR whereas the best performance is 93.4% and is reported by Goswami *et al.* [18].

4. *Chokepoint Database*

   Chokepoint [33] database contains unconstrained videos captured in the surveillance sce-

nario. It consists of 25 subjects and 48 video sequences. Three cameras are fixed at a position to obtain the video sequences, which contains only one subject in every frame.

5. *SN-Flip Database*

   Recently, SN-Flip database was released by Barr *et al.* [4] where each video contains multiple subjects. However, all the subjects in this database are almost still, thus it may not be well suited to evaluate realistic crowd video matching scenarios, i.e., multiple subjects performing some actions.

6. *SCFace-Surveillance Camera Face Database*

   SCFace database [24] consist of static images taken in uncontrolled indoor environment with five fixed cameras. The number of subjects and images in this database are 130 and 4160 respectively. Subjects in the dataset are guided to mimic real world situations.

2.2 ACVF-2014 Dataset: Device Details

The proposed ACVF-2014 dataset is collected using three portable handheld devices having different resolutions. These devices are: Nikon Coolpix S570, Sony handycam DCR-DVD910E, and Apple iPhone (4s and 5c). The three devices are referred to as Device I, Device II, and Device III respectively. The device difference leads to varying quality of captured videos.

To yield more background on the data collection process, images are shown from different devices in Figure 2.1, elucidating the variation in pose, or location, sensor, distance to camera, and other biometric modalities, while detected and cropped faces of two subjects from frames is shown in Figure 2.3. The original size of frames varies according to the device used to capture the video and cropped face is of size 160×125 pixels. These close-ups in detected faces illustrate many aspects of this database that make it more challenging.

In this database no user level intervention is forced. Most of the videos are captured while the students are entering or leaving class rooms. The students are not given any instructions and therefore, the acquisition environment is completely unconstrained. Videos are

Figure 2.1: Difference in quality of frames by different devices

taken in both indoor and outdoor environments. Another motivation for this database is to discourage those cases in which a person looks into the video camera directly for a prolonged time, transforming the task to frontal still image recognition. The videos are recorded using handheld devices without mounting on any tripod or similar structure. The dataset details are described below and a summary is provided in Table 2.2. Typically, in all the videos, subjects appear in groups; therefore, almost all the video frames contain more than one subject (refer to Figure 2.2).

Table 2.2: Details of the Annotated Crowd Video Face Database-2014.

Device (Resolution)	**#Videos**	**#Frames**	**#Subjtects**	**Subjects/Video**			**#Faces**			
				Min	**Max**	**Avg**	**G.Truth**	**Automatic Detection**	**False Detects (Removed)**	**Final Detects (Used)**
Device I (640X480)	115	16,704	120	1	14	2.8	22,635	13,973	4,415	9,558
Device II (2304X1296)	72	9,566	116	1	10	2.3	12,263	10,459	3,071	7,388
Device III (1920X1080)	14	1,741	20	1	4	2.1	2,563	3,309	1,267	2,042
Total	201	28,011	133	1	14	2.6	37,461	27,741	8,753	18,988

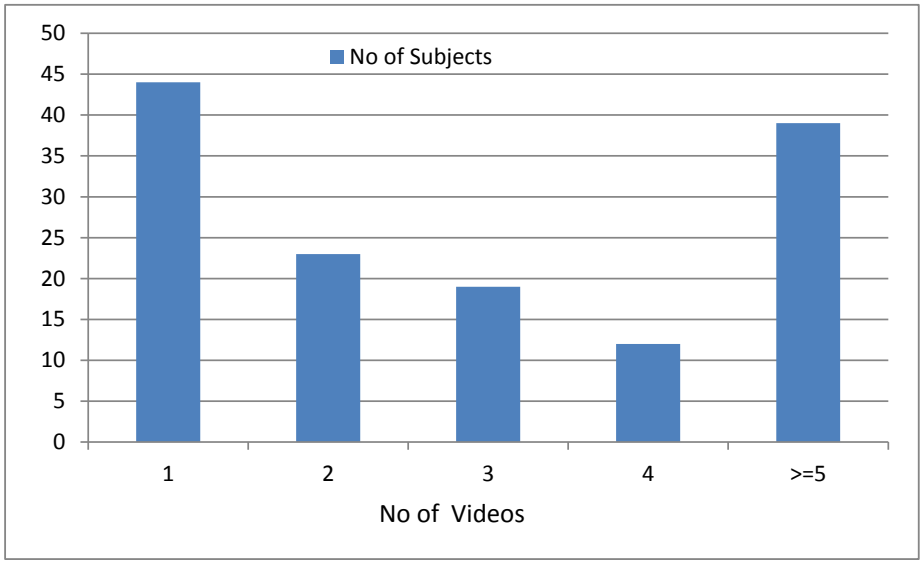


Figure 2.2: Illustrating the number of videos per subject in the ACVF Dataset, for instance there are 44 subjects which appear in exactly one video.

2.3 Annotation, Face Detection, and Registration

Subject IDs along with a *point of interest* (POI) and *gender* of all the faces present in a frame are manually annotated. Point of interest is a manually marked point which is surrounded by the face box. We utilize the publicly available code of Everingham *et al.* [15] for face detection and extract the cropped faces of size 125×160. The face detection algorithm also finds nine landmark points from the face region: two corners of both the eyes, two corners of lips, two corners of nose, and a nose tip. These nine landmark points are utilized to register a detected face with a canonical face frame.

A subject ID is assigned to each extracted face image only if the manually annotated POI lies in the face rectangle of detected face. The procedure is illustrated in Figure 2.3, where in a frame there are two subjects who are annotated with their gender type as 0 or 1 where 0 represents male gender and 1 represents female. If no POI falls within a detected face rectangle, it is considered as incorrect: a case of false face detection. It is possible that after POI based filtering, partial faces and faces with background information may

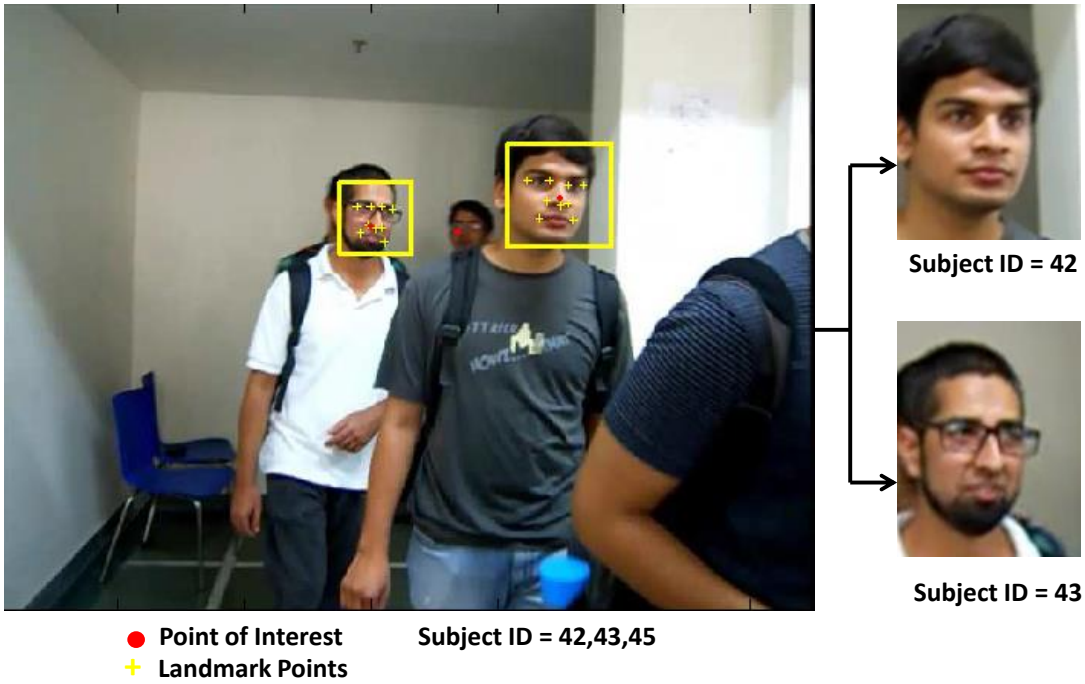


Figure 2.3: The annotation and face detection on an example frame. There are three POIs marked, where as the face detection algorithm detects two faces. POIs that are surround by each face-box are used to assign ground-truth subject IDs with each extracted faces. Also, there are some failures in detection cases, e.g. subject 45 is not detected in this example.

be obtained (see Figure 2.4). Such faces may be considered inaccurate face detections. Figure 2.4 shows samples of detected and registered faces, inaccurately detected faces, and false detections. Due to the presence of covariates such as low resolution, blur, and nonuniform lighting, not all faces are successfully detected. As mentioned in Table 2.2, out of the total manually marked 37,461 faces, only 27,741 faces are detected, out of which 8,753 face images are discarded based on POI annotations (Figure 2.4 shows some failure cases). Thus, the remaining set of 18,988 faces corresponding to 133 unique subjects is utilized in the experiments. Figure 2.6 illustrates the number of detected faces of each subject in each video along with the respective ground truth information.

2.4 Application Scenarios and Experimental Protocols

As mentioned earlier, the ACVF dataset focuses on unconstrained face recognition with multiple subjects in a video or an image. With these variations, there are two application

(a) (b) (c)

Figure 2.4: (a) Examples of accurately detected faces corresponding to each of the three devices. (b) sample of inaccurate face detection such as partial face and presence of extra non-face/background regions, and (c) shows examples of false detections which are discarded based on the POI annotations.

Figure 2.5: Detected and cropped faces of two different subjects by different devices. First row shows Device II, second row shows Device III and Third row shows Device I

scenarios in which this database can be utilized:

(a) **Scenario I: Matching a subject from one video with a subject from another video.**

If we have a set of footages acquired from a set of devices, and we wish to check if a person in footage A appears in footage B, i.e. both the samples in the comparison pair are obtained without expecting user cooperation. In this scenario, gallery set is defined in terms of a *set of videos*. Let the gallery set be defined as $\mathcal{G} = \{I_{v,f,n} | v \in \mathcal{V}\}$; where $\mathcal{V}$ is the set of video IDs selected to be the part of gallery and the n^{th} detected face image from a frame f of video v be denoted as $I_{v,f,n}$. This scenario has the following three different evaluation settings, each associated with a certain real world application:

- **Frame-to-Frame Matching:** Scores are obtained by matching every face image (frame) in the probe set with every face image (frame) in the gallery set. The comparison of a probe video consisting of m face images and a gallery video consisting of n face images results in mn match scores.
- **Video-to-Frame Matching:** The probe face frame is compared against every video in the gallery set. A set of scores is obtained by comparing a probe face frame with all the face frames in the gallery video. If the gallery video consists of q subjects, the set of scores are divided into q subsets, each corresponding to one subject. The scores within each subset are aggregated to obtain 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 mq match scores.
- **Video-to-Video Matching:** The probe video set is compared against the gallery video set. Each of the probe face images (frames) are compared with all the face images (frames) in the gallery video. For every video pair matching, the set of scores are aggregated such that a match score is obtained for every subject-pair comparison. Therefore, comparison of a probe video consisting of p subjects against a gallery video showing q subjects results in pq match scores.

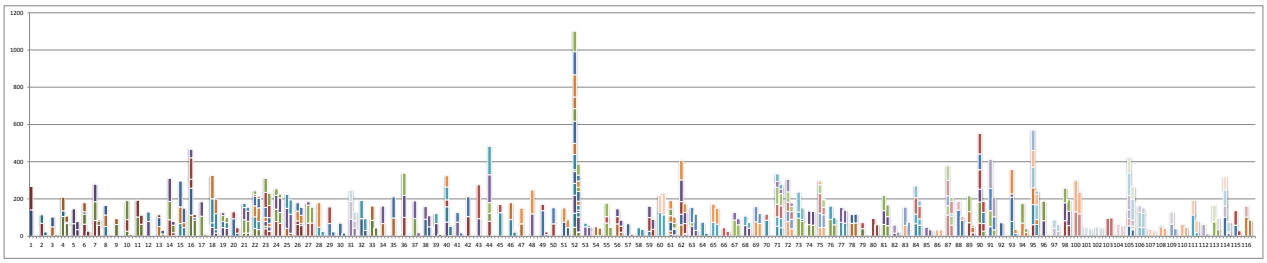
In all the three cases, the scores of one probe video comparison must not affect the scores of another probe video. For Scenario I, the videos for the gallery set are chosen such that every subject is present in at least one of the videos. The process of obtaining gallery-probe split of videos is repeated 10 times to obtain the cross validation sets. The number of videos in the gallery set ranges between 61 to 71. These cross validation sets are included in the evaluation package.

(b) **Scenario II: Matching an image of a subject with an image of a subject.** In Scenario II, the gallery set is defined in terms of a *set of images*, i.e., the video information is not considered, and images are referred using only indices. In this scenario, the gallery set is defined as $\mathcal{G} = \{I_k | k \in \mathcal{K}\}$; where I_k denotes the k^{th} image and $\mathcal{K}$ is the set of image indices selected to be a part of the gallery set. In this protocol, it is possible that both gallery and probe images may be from the same video. This protocol helps to understand the performance of algorithms when face matching is required within a video, at different time stamps. 10 images per person are randomly selected to constitute the gallery set, while all the remaining images constitute the probe set. There are 8 subjects having less than 10 images and therefore, all the images pertaining to these subjects are included in the probe set. Thus, the gallery set contains 1250 (125 subjects, 10 images per subject) images, and the probe set contains 17,738 images.

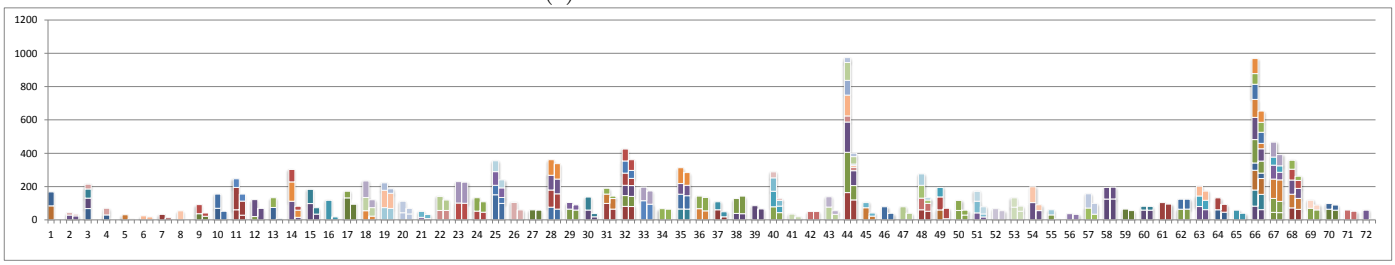
It is important to note that the database is designed to evaluate face recognition systems and therefore, training data is not provided. Researchers may use any data (in any amount) from other sources , not overlapping with from the ACVF-2014 database, to train their algorithms. This makes the evaluation completely non-overlapping and blind, which is the case with real world uncontrolled face recognition applications.

2.5 Baseline Face Recognition Results

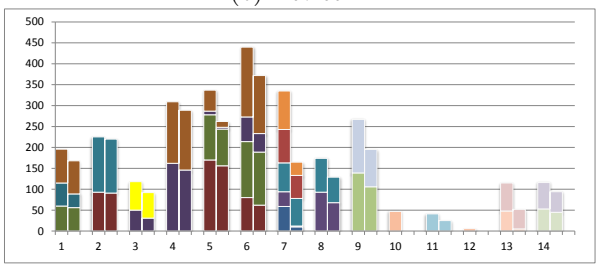
Baseline evaluations have been performed using OpenBR [23] and FaceVACS (which is among the best commercial face recognition systems [19]). The face recognition module



(a) Device I



(b) Device II



(c) Device III

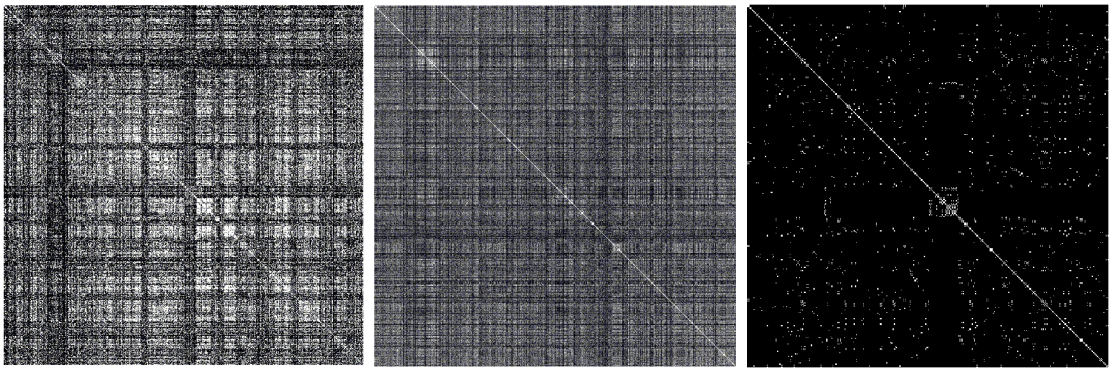
Figure 2.6: Representing the results of face detection. Two stacked bars are shown side-by-side for each video: first stacked bar represents the number of ground truth faces and the second staked bar represents the number of detected faces. The subparts of the bar (shown in different colors) represent each subject in the video. For example, video # 1 from Device III shows that there are three subjects (green, blue and orange) in the video. Note that the presence of more colors in one stacked bar translates to larger crowd (subjects).

of OpenBR is based on Spectrally Sampled Structural Subspaces Features algorithm, also known as 4SF. For OpenBR, a built-in face detection module is used, whereas, for FaceVACS, eye coordinates are provided for each face image to ensure 100% enrollment in gallery. The verification performance is reported in terms of Receiver Operating Characteristic (ROC) curve. The ROCs obtained for each cross-validation split are combined into one curve using vertical averaging [16]. The results for Scenarios I and II are reported in Figures 2.8 and 2.9, respectively. The key observations are:

- In both the scenarios, at 0.01 False Accept Rate(FAR), the best verification rate achieved is only 0.16 Genuine Accept Rate (GAR). Further, many ROC curves start

around 0.05 FAR which is likely to happen when the match score distribution does not have a long tail. This poor performance indicates the complexity of the problem as well as the limitation of the current systems.

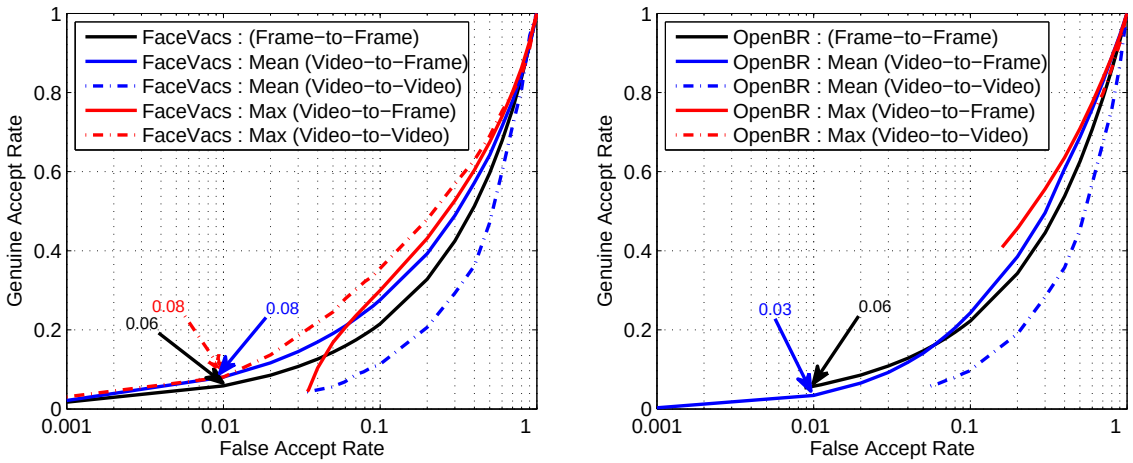
- In both the scenarios, FaceVACS appears to perform slightly better than OpenBR. However, at higher FARs, the difference in the performances is not significant. It should be noted that eye annotation information is provided as an additional input to FaceVACS whereas OpenBR operates on loosely cropped faces from which it has to detect face region on its own.



(a) (b) (c)

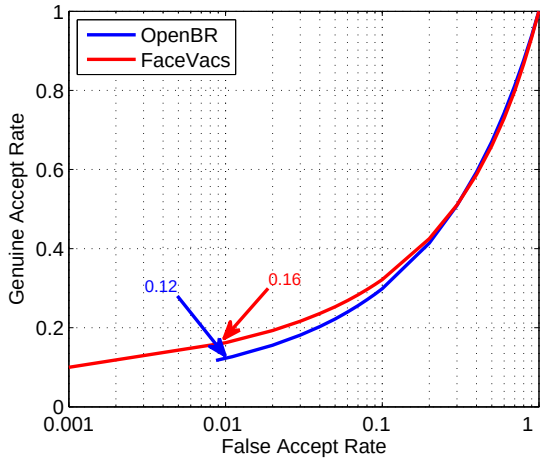
Figure 2.7: Visualization of 18,988×18,988 similarity matrices obtained from (a) OpenBR and (b) FaceVACS. (c) shows the ideal similarity matrix for the given database. Darker pixels represent lower similarity between the corresponding gallery and probe image pair. All the three matrices are symmetric.

- 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 [18].
- At low FAR (≤ 0.01), Scenario II yields slightly better verification rate than Scenario I. In Scenario II, it is possible (and also likely) to have images of a subject from the same video in gallery as well as in probe. Intuitively, they should be easier to match and such scores lead to a minor improvement in performance.
- Figure 2.7 shows the similarity matrices (symmetric) of both the systems obtained



Scenario I

Figure 2.8: On the proposed ACVF-2014 database, ROC curves showcasing the verification performance of FaceVACS (left) and OpenBR (right) for different settings of Scenario I



Scenario II

Figure 2.9: In Scenario II, since no frame or video associations are considered while generating the gallery probe splits, this scenario is close to still-to-still matching.

by comparing all the detected faces with each other. The ideal similarity matrix is also shown, which has value 1 for all the genuine scores and 0 for all the impostor scores. The entropy of this matrix is very low whereas the entropies of the other two matrices are very high. This analysis substantiates the results obtained from ROC curves that a significant effort is required to achieve higher accuracies on the AVCF 2014 database.

2.6 Evaluation Package and Guidelines

The evaluation package provides researchers a platform to perform their experiments. We encourage researchers to apply their own face detection algorithms and use the provided end to end evaluation code to obtain results. Each registered output face image obtained from the face detection algorithm is named using the following convention.

$$DeviceName\_VideoID\_FrameNo\_SubjectID.jpg \tag{2.1}$$

Moreover, the registered face images are provided in the `/Cropped/DeviceName/VideoID` directory of the package for easier access. A sample directory structure is presented in Figure 2.10.

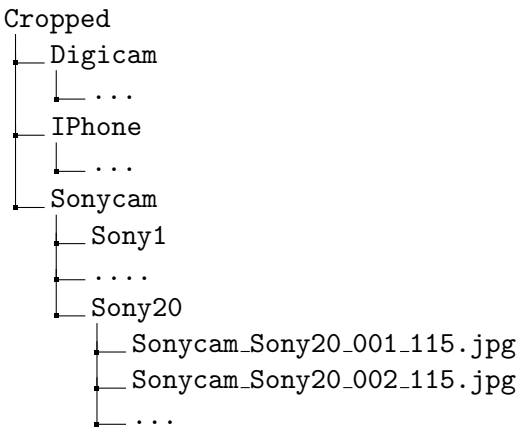


Figure 2.10: Directory structure of the cropped face images provided as part of dataset package.

In the evaluation package we provide:

- Raw videos, detected faces images, and annotation information (POI and face landmark points detected using [15]),
- Text file indicating, the gender of each subject; male gender is represented by 0 and female represented by 1.
- Protocol files and mask matrices, and
- MATLAB code for end-to-end evaluation.

The package is designed to make the overall evaluation process as easy as possible. To carry out the evaluation analysis, a 18,988×18,988 similarity matrix is required as input. Various evaluations can be performed from this similarity matrix and all the protocol and annotation files are provided as part of the package.

Chapter 3

Gender Classification

3.1 Overview

Attributes such as age and gender can be utilized to boost the performance of a face recognition system. Such attributes helps in speeding up the search process as they can be used as a basic indexing approach. Gender classification approaches attempt to classify the given face image into male or female classes. Among all the soft attributes, gender classification is useful not only in biometrics but also in human computer interaction (refer to Figure 3.1), surveillance, and monitoring. In controlled settings, existing algorithms have achieved acceptable level of performance (,i.e., over 90% classification accuracy). In unconstrained environment, which includes low quality noisy images, with pose, expression and illumination variations, postulate a big challenge. In this chapter, a gender classification framework via textual feature, is presented and results are shown on the proposed ACVF-2014 dataset.

3.2 Feature Descriptors

An effective facial representation is required for successful gender classification. For extracting local texture features, Local Binary Patterns (LBP) operator and Histogram of

Figure 3.1: Robot interacting with Humans [1]

Oriented Gradient (HOG) descriptor are widely used. In this research we have used these two texture feature based approaches for feature extraction for gender classification.

3.2.1 LBP: Local Binary Pattern

The local binary pattern operator provides an array or image of integer labels which describe the appearance of the image on a small scale. These labels or histograms are used further for image analysis. Its widely used version is designed for monochrome still images but it is also possible to extend the operator for color images.

(a) *Basic LBP*

Ojala *et al.* [26] introduced the basic local binary pattern by making an assumption that the pattern and its strength are aspects of texture. The original version of LBP works on any image by considering a block of 3×3 pixels. The pixels in a block are thresholded by the center pixel. These obtained values are summed up to get the value (label) of the center pixel. Depending on the gray value of the center pixel and the neighborhood pixels we obtain $2^8 = 256$ different labels, when the neighborhood consist of 8 pixels. Figure 3.2 illustrates the process

(b) *Modified LBP Operator*

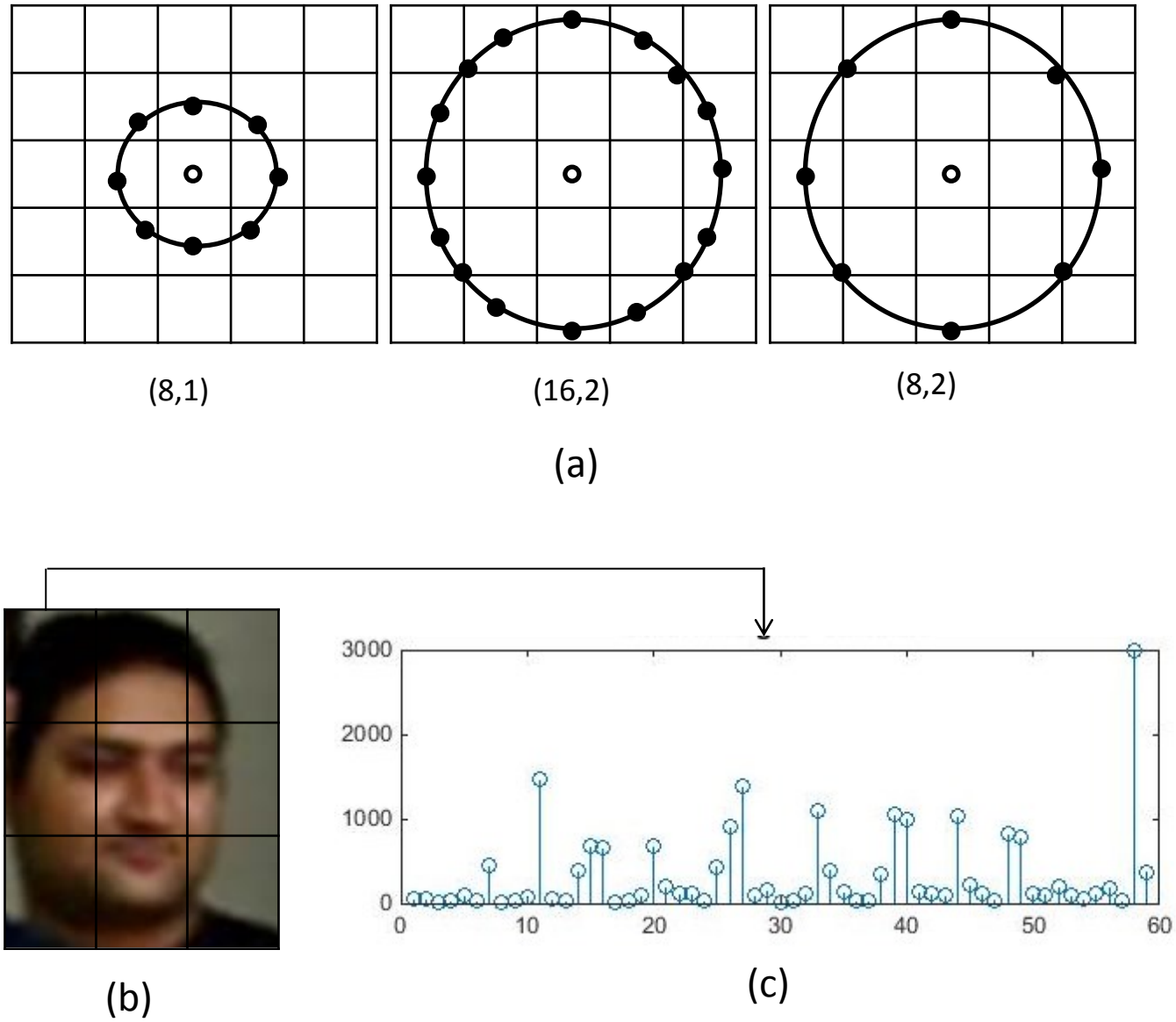


Figure 3.2: Illustration of how LBP descriptor is obtained. a) Different size neighborhoods, b) Image is divided into patches and, c) LBP histogram of a patch

In [29], a slightly modified version is proposed which is used in this research. Consider a gray image $I(x, y)$ and let g_c denote the gray level of a pixel (x, y), i.e., $g_c = I(x, y)$. Moreover, let g_p denote the gray value of a sampling point in a circular neighborhood of P sampling points and radius R around point (x, y):

$$g_p = I(x_p, y_p), p = 0, ..., P1 \tag{3.1}$$

$$x_p = x + Rcos(2p/P), \tag{3.2}$$

$$y_p = y + Rcos(2p/P), \tag{3.3}$$

It is assumed that the local texture of the image $I(x, y)$ is characterized by joint distribution of gray values of $P + 1(P > 0)$ pixels:

$$T = t(g_c, g_0, g_1, ..., g_{p-1}) \tag{3.4}$$

There will be no loss of information if we subtract the center pixel value from the neighborhood:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c). \tag{3.5}$$

Center pixel is assumed to be statistically independent of the difference, so factorization is performed.

$$T = t(g_c) \;\; t(g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{3.6}$$

The first factor is $t(g_c)$ which is the intensity distribution over $I(x, y)$, it contains no useful information whereas joint distribution of differences is used to model the local texture.

$$t(g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{3.7}$$

The solution proposed by Ojala *et al.* [27] for this problem is to apply vector quantization. The dimension of the expression is reduced by using vector quantization with a codebook of 384 codewords, where these codewords correspond to 384 bins in a histogram. Further, the signs corresponding to the differences are considered:

$$t(s(g_0 - g_c), s(g_1 - g_c), ..., s(g_{p-1} - g_c)) \tag{3.8}$$

where $s(z)$ is a step function.

$$s(z) = \begin{cases} 1 & if\ z>=0 \\ 0 & otherwise \end{cases} \tag{3.9}$$

Local binary pattern operator is obtained by summing up these thresholded differences weighted by powers of two. The $LBP_{P,R}$ operator is defined as:

$$LBP_{P,R} = \sum_{p=0}^{p-1} s(g_p - g_c))2^p \tag{3.10}$$

where, g_c corresponds to the gray value of the center pixel (x_c, y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s is a thresholding function.

3.2.2 HOG: Histogram of Oriented Gradients

The basic idea of HOG descriptor is that face appearance and shape can be characterized well by the distribution of edge directions and local intensity gradients. HOG is implemented by dividing the image into small regions, then the histogram of edge directions or gradient directions is obtained over the pixels. The HOG descriptor obtained by combining the entries of histograms of each region. HOG feature of a detected cropped image is shown in Figure 3.3.

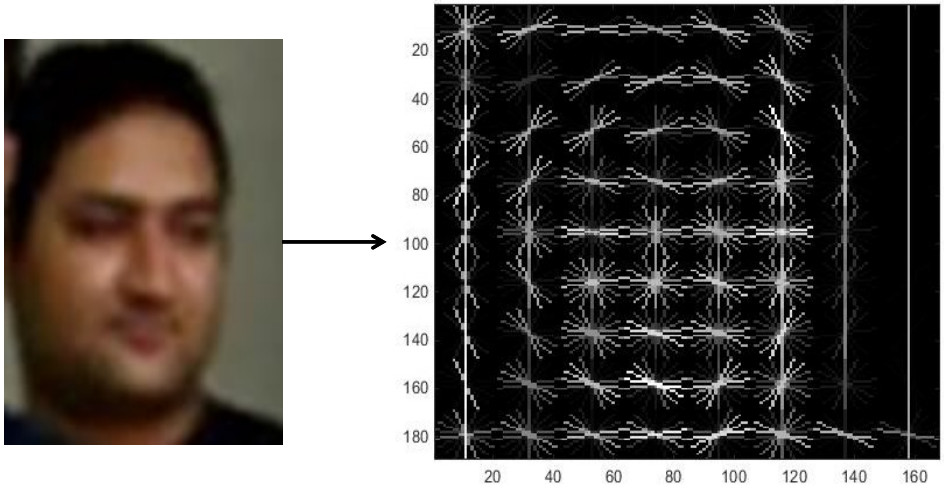


Figure 3.3: HOG descriptor of a detected face from a frame.

3.3 Gender Classification Algorithms

Figure 3.4 illustrates the pipeline for gender classification. The algorithm detects faces from the input frame and texture features are extracted. These features are then given as input to a support vector machine (SVM) classifier. We next present a brief overview of SVM and how it is used in gender classification.

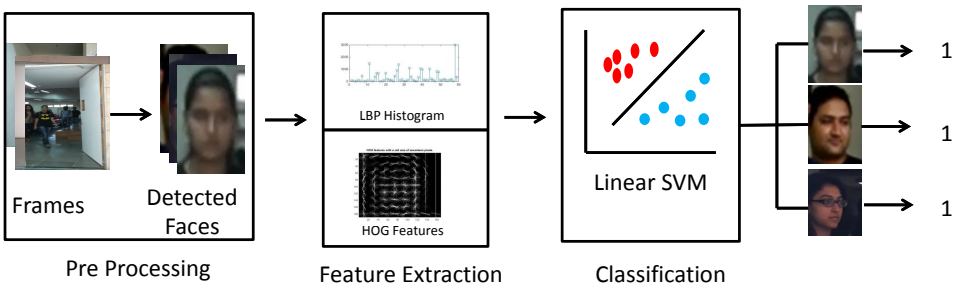


Figure 3.4: Steps involved in a gender classification system,

3.3.1 Support Vector Machine for Gender Classification

SVM is a supervised learning algorithm which classifies the data by providing a hyperplane with the objective of minimizing the mis-classification rate. Using given training set along with the corresponding labels, SVM model is built to classify the test data into one or another class. As shown in Figure 3.5, SVM represents the data points in space such that two classes can be discriminated easily. If the data is not linearly separable, SVM can map the input data to higher dimensions by using the kernel trick.

Let the training data contain N samples, (x_i, y_i) where, x_i is a LBP/HOG feature pertaining to an image in the training set and y is the label, which is either 1 or 0 based on the gender of the image(1female, 0-male). SVM finds the best separating hyperplane given by the equation $w^T x + b$ that maximizes the distance between the two class distributions. For the two classes, equations of the two corresponding hyperplanes are :-

$$w^T x + b = 1 \tag{3.11}$$

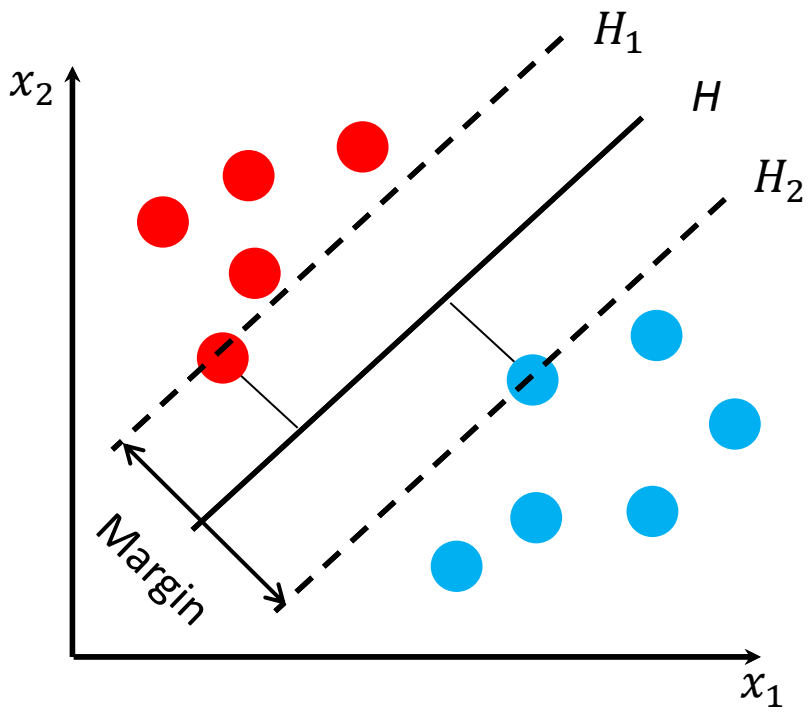


Figure 3.5: Maximum Margin Hyperplanes H1 and H2, Samples on margin hyperplane are support vectors

$$w^T x + b = -1 \tag{3.12}$$

The objective of the optimizing function is to maximize the margin $\frac{2}{w^T w}$ which is defined as the distance between the two hyperplanes. The equivalent optimization function is

$$minimize \ \ \frac{1}{2} w^T w \tag{3.13}$$

subject to the constraint that

$$y_i(x_i w^T + b) >= 1 \quad \forall i \tag{3.14}$$

Further, the formulation is extended to introduce a slack variable ε_i and cost parameter C,

$$minimize \ \ \frac{\|\omega\|^2}{2} + C(\sum_i \varepsilon_i)^k \quad k = 1 \ or \ 2 \tag{3.15}$$

subject to the constraints

$$y_i(x_i w^T + b) >= 1 - \varepsilon_i \quad for \;\; i = 1, ..., N)$$
$$\varepsilon_i >= 0$$

Dual form of the aforementioned primal form is,

$$maximize \;\; \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{3.16}$$

subject to $0 <= \alpha_i <= C$ and $\sum_i \alpha_i y_i \;=\; 0$..

Once trained, the learned model is used for classification in the test data. Specifically, the input probe LBP/HOG features are extracted and the trained SVM model predicts the gender when the two classes are male and female.

3.3.2 OpenBR

Open Source Biometric Recognition (OpenBR) is a framework which provides tools to evaluate and design new biometric algorithms and an interface to incorporate biometric technology into end-user applications. The Algorithms implemented in OpenBR are applicable to face recognition, gender classification, and age prediction. The default algorithm in OpenBR for gender classification is based on the algorithm proposed in [22] where LBP and SIFT features are utilized in combination with PCA and SVM.

3.4 Experimental Protocol

The experimental protocol for gender classification is as follows :

i. The ACVF-2014 videos are divided into test and train set such that the train set consists of exactly one video corresponding to each subject and remaining videos fall under test set.

ii. Ten such randomly sampled train-test partitions are created for cross validation.

3.5 Results

Gender classification is performed using the off the shelf system OpenBR System and SVM classifier combined with LBP and HOG descriptors, to which detected and cropped faces are provided as input. In accordance with the evaluation protocol, results are reported on the test set. The following metrics are used for evaluation: total accuracy, male classification accuracy, and female classification accuracy. The results are summarized in Figure 3.7, where we observe that SVM with HOG descriptor provides the best average accuracies. An example of misclassification can be seen in Figure 3.6 where one male subject is classified as female.

Figure 3.6: Gender Mis classification example

i. OpenBR Results

Default models of OpenBR can perform age and gender classification of Caucasian individuals very well whereas, lower accuracies are observed when it is used on Asian individuals [2]. We also observe similar results; it can be seen from the Figure 3.8 that the average accuracies obtained by OpenBR are around 50%,

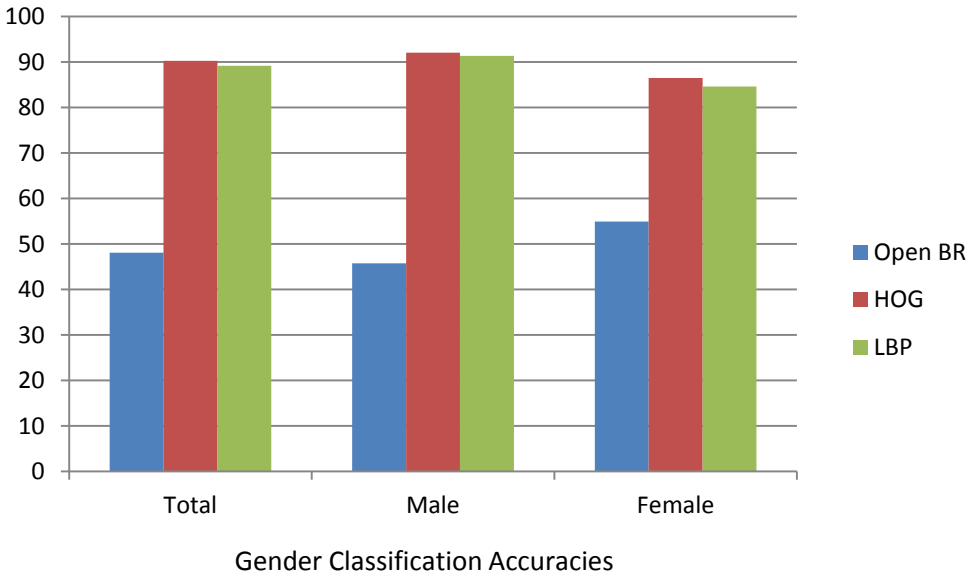


Figure 3.7: Comparison of Open BR, HOG and LBP Results

which can be considered low for real world applications.

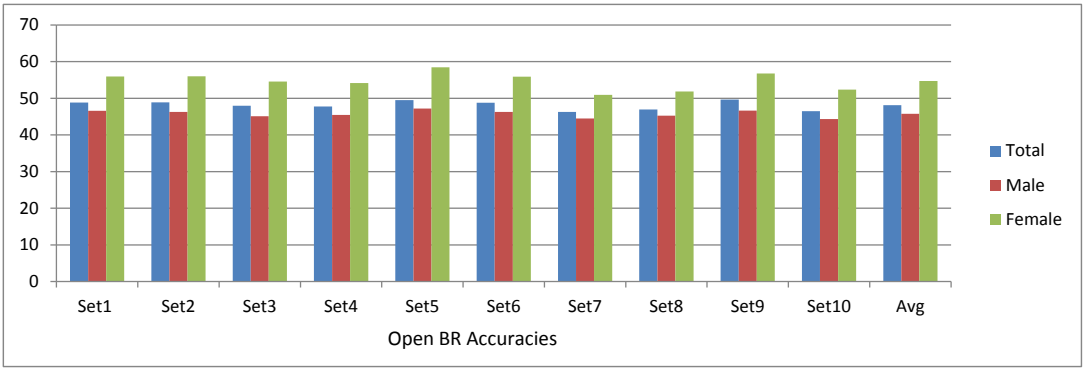


Figure 3.8: OpenBR Results

ii. LBP Results

When LBP features are used with linear SVM, the average gender classification accuracy is 89.16%. It is shown in Figure 3.9 that the male and female classification accuracies are also very high. This indicates that training or representative data can improve the performance.

iii. HOG Results

HOG feature descriptors classified using linear SVM gives the best results with

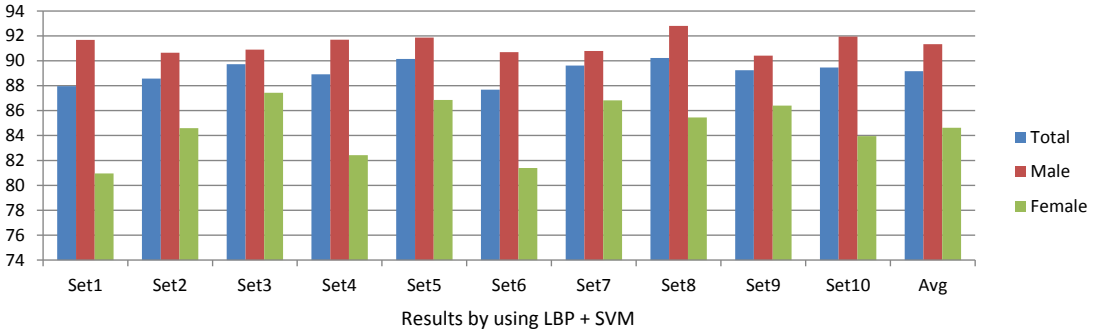


Figure 3.9: LBP + SVM Results

90.25% total accuracy, male accuracy 92.05 %, and female accuracy 86.50 %. As shown in Figure 3.10, in general, higher number of faces are misclassified as male. We have observed the same trends with both OpenBR and LBP approaches and we believe that this is primarily due to the unbalanced nature of the training set.

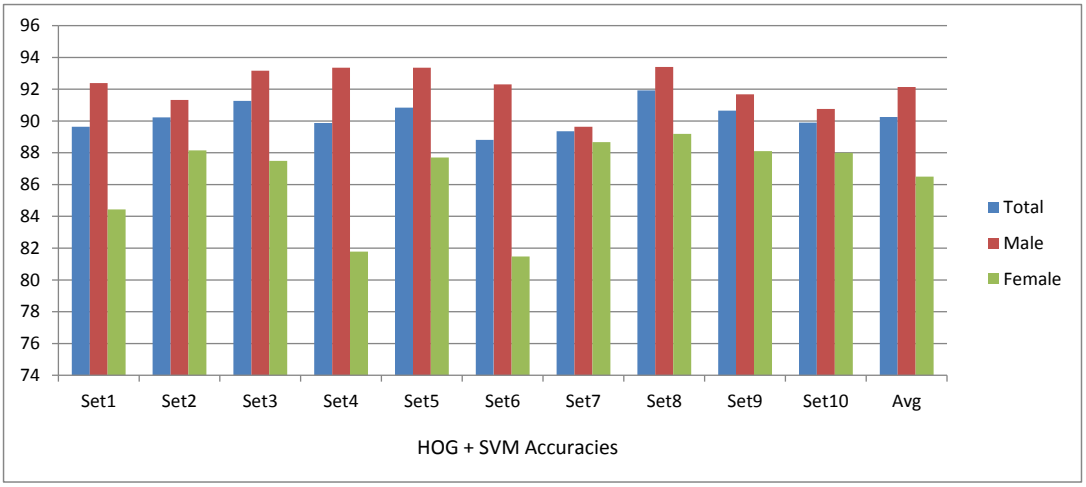


Figure 3.10: HOG + SVM Results

Chapter 4

Conclusion and Future Work

Face and gender recognition in unconstrained surveillance scenario is very difficult. ACVF-2014 database is proposed which contains 201 unconstrained videos (28,011 frames) of 133 subjects captured at various locations, and each video contains up to 14 subjects. To provide researchers a platform for evaluating their own algorithms in the challenging conditions, an evaluation package is provided which consists of protocol files along with MATLAB end to end evaluation code. Baseline results are provided on a commercial system to observe the complexity of this dataset, and results shows that it is a very challenging dataset since neither baseline result yields more than 0.16 genuine accept rate at 0.01 false accept rate. Our future work aims at proposing novel algorithms for face and gender recognition on this challenging dataset.

Bibliography

[1] Bot scene. http://botscene.net/2013/02/11/robovie-goes-to-school/.

[2] How to train openbr for age and gender recognition? https://groups.google.com/forum/#!topic/openbr-dev/8YHvz9hTpNM.

[3] Youtube. https://youtu.be/fUTirK_43ps.

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

[5] Bekios-Calfa, J., Buenaposada, J. M., and Baumela, L. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 4 (2011), 858–864.

[6] Best-Rowden, L., Klare, B., Klontz, J., and Jain, A. K. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (2013), pp. 1–8.

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

[8] Bhatt, H., Singh, R., and Vatsa, M. On recognizing faces in videos using clustering-based re-ranking and fusion. *IEEE Transaction on Information Forensics and Security 9*, 7 (2014), 1056–1068.

[9] Bhatt, H., Singh, R., Vatsa, M., and Ratha, N. Improving cross-resolution face matching using ensemble-based co-transfer learning. *IEEE Transaction on*

Image Processing 23, 12 (2014), 5654–5669.

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

[11] Bhatt, H. S., Singh, R., and Vatsa, M. Covariates of face recognition. Tech. Rep. IIITD-TR-2015-002, IIIT-Delhi, 2015.

[12] Bledsoe, W. W. The model method in facial recognition. *Panoramic Research Inc., Rep. PR1 15* (1966), 47.

[13] Chen, Y.-C., Patel, V. M., Phillips, P. J., and Chellappa, R. Dictionary-based face recognition from video. In *Proceedings of the 12th European Conference on Computer Vision*. Springer, 2012, pp. 766–779.

[14] Dhamecha, T., Sankaran, A., Singh, R., and Vatsa, M. Is gender classification across ethnicity feasible using discriminant functions? *International Joint Conference on Biometrics Compendium (IJCB)* (2011), 1 – 7.

[15] Everingham, M., Sivic, J., and Zisserman, A. Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing 27*, 5 (2009), 545–559.

[16] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Lett. 27*, 8 (2006), 861–874.

[17] Goh, R., Liu, L., Liu, X., and Chen, T. The CMU face in action (FIA) database. In *Proceedings of the Second International Conference on Analysis and Modelling of Faces and Gestures*. 2005, pp. 255–263.

[18] Goswami, G., Bhardwaj, R., Singh, R., and Vatsa, M. MDLFace: Memorability augmented deep learning for video face recognition. In *IEEE International Joint Conference on Biometrics (IJCB)* (2014), pp. 1–7.

[19] Grother, P., Quinn, G., and Phillips, P. Multiple biometric evaluation (MBE) 2010, report on the evaluation of 2D still-image face recognition algorithms. *NIST Interagency Report 7709* (2010).

[20] Jain, A., and Huang, J. Integrating independent components and linear discriminant analysis for gender classification. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition* (2004), pp. 159–163.

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

[22] Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., and Jain, A. K. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security 7*, 6 (2012), 1789–1801.

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

[24] Mislav, G., Kresimir, D., and Grgic, S. Scface surveillance cameras face database. *Multimedia Tools and Applications 51*, 3 (2011), 863–879.

[25] Ng, C. B., Tay, Y. H., and Goi, B.-M. Recognizing human gender in computer vision: A survey. In *Proceedings of the 12th Pacific Rim International Conference on Trends in Artificial Intelligence* (2012), PRICAI'12, pp. 335–346.

[26] Ojala, T., and Pietikinen, M.and Harwood, D. A comparative study of texture measures with classification. *Pattern Recognition 29* (1996), 51–59.

[27] Ojala, T., Valkealahti, K., Oja, E., and Pietikinen, M. Texture discrimination with multidimensional distributions of signed gray level differences. *Pattern Recognition 34* (2001), 727–739.

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

[29] Pietikainen, M., Hadid, A., Zhao, G., and Ahonen, T. *Computer Vision Using Local Binary Patterns*. Computational Imaging and Vision. 2011.

[30] Singh, M., Nagpal, S., Singh, R., and Vatsa, M. On recognizing face images with weight and age variations. *IEEE Access 2* (2014), 822–830.

[31] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1701–1708.

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

[33] Wong, Y., Chen, S., Mau, S., Sanderson, C., and Lovell, B. C. Patch-based probabilistic image quality assessment for face selection and improved video-based

face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2011), pp. 81–88.

[34] Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. Face recognition: A literature survey. *ACM Computing Surveys 35*, 4 (2003), 399–458.