

Gender Classification using RGB-D Videos

Navin Agrawal

IIIT-D-MTech-CS-GEN-13-043

Dec 15, 2015

Indraprastha Institute of Information Technology Delhi
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

© Agrawal, 2015

Keywords : Gender Classification, RGB-D Kinect Video, Face Biometrics

Certificate

This is to certify that the thesis titled “**Gender Classification using RGB-D Videos**” submitted by **Navin Agrawal** 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 him 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

Gender classification is used in applications as a soft feature or attribute in biometrics to help identify people. Using gender classification as an indexing technique can boost the performance of facial-biometric. If the face images are obtained using high quality camera then only RGB information is sufficient for gender classification. However, in surveillance scenario RGB face images are of low quality and have covariates such as pose, illumination, expression and distance. Therefore, in such scenarios depth information can be utilised to improve the performance of gender classification. Low-cost depth sensors such as Microsoft Kinect provide the depth images along with corresponding RGB color images. These low cost (Kinect) devices can be used for video surveillance; however, not much research has been focused on RGB-D (RGB and Depth data) video data obtained from these devices. In this research, we present a novel gender classification algorithm that extracts features using multiple algorithms from RGB-D videos. While most of the work in gender classification has focused on handcrafted feature extraction techniques such as Uniform Local Binary Pattern and Gradient Local Binary Pattern, we have also studied effectiveness of learned feature extraction techniques such as Stacked denoising autoencoder on gender classification. We also present a score level fusion of handcrafted features and learned features, which significantly improves the performance of gender classification. The proposed algorithm is evaluated on KaspAROV dataset, which contains RGB-D video data obtained from Microsoft Kinect device. This dataset encompasses challenges of varying conditions related to illumination, pose, expression, low image quality and distance. The experiments are also performed on Eurecom Kinect dataset. On both the databases the proposed algorithm achieves state-of-the-art results.

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 Anurag Chowdhary, without whose support this work would not have to be done. This section can not be complete without a vote of thanks to academic department for their help and never ending support.

Contents

1	Introduction	2
1.1	Overview and Research Motivation	2
1.2	Literature Review	4
2	Preliminaries	8
2.1	Preprocessing: Layered Bilateral Filtering	8
2.2	Feature Extraction	10
2.2.1	ULBP: Uniform Local Binary Pattern	11
2.2.2	GLBP: Gradient Local Binary Pattern	13
2.2.3	SDAE: Stacked Denoising AutoEncoder	14
2.3	Classification using Support Vector Machine	16
3	Gender Classification	18
3.1	Proposed Algorithm	18
3.1.1	Preprocessing	18
3.1.2	Feature Extraction and Classification	20
3.2	Dataset Specification	22
3.2.1	KaspAROV Kinect Video Dataset	22
3.2.2	Eurecom Kinect Face Dataset	23
4	Result and Conclusion	25
4.1	Experimental Protocol	25
4.2	Results	26
4.3	Conclusion and Future Work	30

List of Figures

1.1	Sample RGB images under constrained environment.	3
1.2	Sample RGB images obtained from surveillance camera.	3
2.1	KaspAROV dataset a) RGB Image. b) Depth image before preprocessing. c) Depth image after applying Layered bilateral filter.	10
2.2	Example of LBP operator [28].	12
2.3	Uniform LBP histogram obtained from the face image divided into patches. . . .	12
2.4	Illustration of LBP image obtained after applying LBP operator.	12
2.5	Example of Gradient LBP [18].	13
2.6	Illustration of GLBP images obtained corresponding to four orientations after applying GLBP operator.	14
2.7	Illustration of Autoencoder [6].	15
2.8	Maximum Margin Hyperplanes H_1 and H_2 , samples on margin hyperplane are support vectors [11].	17
3.1	Algorithm for RGB-D face based gender classification.	19
3.2	KaspAROV dataset a) RGB Image. b) Depth image before preprocessing. c) Depth image after applying Layered bilateral filter.	19
3.3	Architecture of two layer stacked denoising autoencoder.	21
3.4	Sample RGB and depth images of KaspAROV dataset along with covariates. The first two rows contains images captured using Kinect v1 device and the last two rows contains images captured using Kinect v2 device.	23
3.5	Sample images from Eurecom dataset a) RGB Image light on. b) Depth image light on. c) RGB Image neutral. d) Depth image neutral. e) RGB Image smile. f) Depth image smile.	24
4.1	ROC on KaspAROV and Eurecom dataset, positive class is female and negative class is male.	28

List of Tables

1.1	Summary of Literature Review	7
3.1	Summary of databases	24
4.1	Training and testing split of the dataset used for experiments.	26
4.2	χ^2 value using McNemar Test.	27
4.3	Area Under Curve (AUC) and Equal error rate (EER - %) reported for Kinect v1 and Kinect v2 of KaspAROV dataset.	27
4.4	Accuracy (%) on KaspAROV dataset Kinect v1 device using 3 fold cross validation.	28
4.5	Accuracy (%) on KaspAROV dataset Kinect v2 device using 3 fold cross validation.	29
4.6	Accuracy (%) on Eurecom dataset Kinect v1 device with two experimental protocols.	29
4.7	Area Under Curve (AUC) and Equal Error Rate (EER - %) of Eurecom dataset Kinect v1 device with two experimental protocols.	29
4.8	Images correctly classified (\surd) and misclassified (\times) by ULBP and SDAE	30

Chapter 1

Introduction

1.1 Overview and Research Motivation

Automatic gender classification has several emerging practical use cases for example, it can be used in surveillance camera settings of departmental stores such as Walmart, which records the footfall of people in their stores to estimate the gender ratio visiting them and let them stock different products accordingly. It can be used in security surveillance cameras, as gender detection from facial features can narrow down the search space of face recognition systems. Gender recognition can also be used in human computer interaction systems such as in creating avatars for virtual world and gaming.

In case of face images obtained from conventional cameras under constrained conditions as shown in Figure 1.1, RGB information is sufficient to achieve good performance for gender recognition task. On the other hand, in unconstrained environment, face images obtained from surveillance cameras, as shown in Figure 1.2, may not be of good quality and depth information may be exploited along with RGB information to boost the performance. Classical 3D cameras are generally expensive; however, with the introduction of low-cost depth sensors such as Microsoft Kinect, which captures depth information along with RGB color images, face analysis tasks using RGB-D information has gained significant attention. There has been relatively very less amount of research in the field of gender recognition using RGB-D Kinect videos. In this research, we

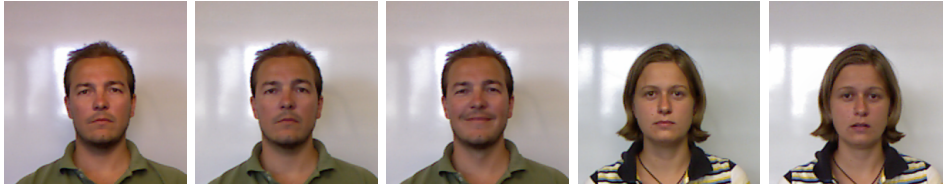


Figure 1.1: Sample RGB images under constrained environment.

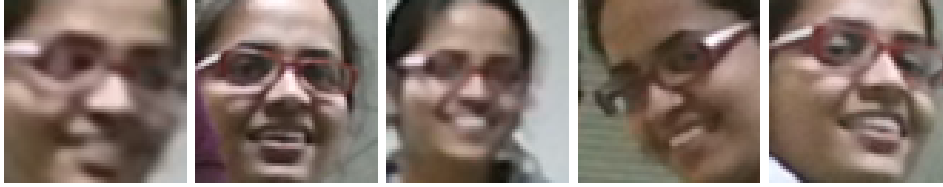


Figure 1.2: Sample RGB images obtained from surveillance camera.

have proposed a novel RGB-D based gender classification approach.

During literature review, we observed that most of the research in gender classification is based on handcrafted feature extraction techniques such as Uniform Local Binary Pattern [28], Histogram of Oriented Gradients [12] and Gradient Local Binary Pattern [18]. Not much attention has been given on data-driven representation learning based feature extraction techniques such as stacked denoising autoencoder [36]. In this study, we have evaluated such data-driven learned feature extraction algorithms on gender classification problem. We have also shown that the performance of gender classification can be improved using the score level fusion of handcrafted features and learned features.

During the course of this research work, we have evaluated the algorithms on KaspAROV and Eurecom Kinect face datasets [26]. The key contributions of our research are:

- Utilizing depth information for improving the performance of gender classification.
- To study the effectiveness of learning based feature extraction techniques such as stacked denoising autoencoder on gender classification problem.
- The score level fusion of handcrafted and learning based feature extraction techniques.
- Evaluate the above algorithms on KaspAROV Kinect video dataset. This dataset is rel-

actively challenging as faces are extracted from RGB-D videos which are taken in unconstrained environment (with respect to pose, illumination, expression and distance).

1.2 Literature Review

Gender classification using 2D face images has been well studied in the literature; however, very few studies have focused on utilising 3D information, especially depth information obtained from low cost Microsoft Kinect devices. In this section, we present an overview of research work in the field of gender classification.

- Alexandre *et al.* [5] have shown how the decision fusion of features obtained from different image sizes improves the performance of the gender classification system. Three varying image sizes (20×20 , 36×36 and 128×128) are used to evaluate the gender classification performance on FERET [30] and UND collection B dataset [14]. They have used Local binary pattern [28] and Histogram of oriented gradients [12] for feature extraction. These features are extracted from varying scale images and classification is performed using Support Vector Machine (SVM) classifier. Decision fusion is applied on the results obtained from these classifiers. The best accuracy of 99.07% on FERET dataset and 91.19% on UND collection B dataset using decision fusion across feature types and image sizes are reported.
- Dhamecha *et al.* [13] have studied gender classification across ethnicity on 2D face images. They have created a heterogenous dataset by combining face images from different publicly available datasets such as CMU PIE [34], Georgia Tech [2], GTAV [1] and FERET [30]. These datasets have subjects belonging to different ethnicities and nationalities. They have selected the face images with covariates such as expression and illumination. They have evaluated the performance and generalization capability of Principal Component Analysis (PCA) [19], Linear Discriminant Analysis (LDA) [32] and Subclass Discriminant Analysis (SDA) [40] on gender classification problem across ethnicity. Three techniques PCA, PCA + LDA and PCA+SDA are evaluated and the best accuracy of 86.47% is reported using PCA.

- Shan *et al.* [33] have studied gender classification on real world 2D face images using Labelled Faces in the Wild (LFW) dataset [17]. This real life dataset contains face images with covariates such as pose, expression, illumination and occlusion. Viola-Jones face detector [37] is used for detecting faces and all the face images are aligned using a commercial software (Wolf *et al.* [38]). Local Binary Pattern (LBP) is used as feature descriptor and Adaboost is applied on top of LBP to select the most discriminative LBPH bins. By using SVM as a binary classifier, accuracy of 94.81% is reported on LFW dataset.
- Influence of automatic and manual alignment method on gender classification accuracy has been reviewed by Makinen *et al.* [23]. IMM Face dataset [35] and FERET dataset [30] are used for the experiments. They have evaluated three automatic alignment methods, one profile alignment method and manual alignment method using four different classification algorithms namely SVM with LBP, Neural network on face pixels, SVM on face pixels and Adaboost with haar like features. They observed that manual alignment method provides better performance than automatic face alignment methods which suggest further enhancement in automatic face alignment methods and SVM provides the best classification accuracy compared to other classification algorithms.
- Above research papers have focused on 2D RGB face images. Some researcher have tried to explore the 3D domain for gender classification. Lu *et al.* [22] use the intensity and range information of human face obtained using Minolta Vivid 910 (3D laser scanner) for gender classification. The range and intensity face images are first normalized and segmentation scheme is applied on it to get the feature vector. SVM is used as classifier for its two class classification problem. Further, they have also used sum rule to combine the posterior probabilities obtained using SVM on range and intensity data. The combined dataset of University of Notre Dame (UND) [10] and Michigan State University (MSU) [22] is used in the experiment and an accuracy of 91% is obtained.
- There are very limited research works on the RGB-D data obtained from low cost Kinect sensors for gender classification. Hyunh *et al.* [18] have made use of depth data obtained from Kinect device in gender classification. They have proposed a novel Gradient LBP

technique on depth images for feature extraction. The technique is evaluated on Eurecom Kinect Face dataset [26] and Texas 3DFR dataset [16]. SVM is used as a classifier. Classification rate of 87.18% is obtained using Gradient LBP compared to 87.82% using Uniform LBP on Eurecom Kinect Face dataset. This experiment is performed on unseen subjects of testing set. Further, they have improved the accuracy to 90.38% on Eurecom dataset using the weighted combination of Uniform LBP and Gradient LBP.

- Boutellaa *et al.* [7] have reviewed the use of Kinect depth data in face analysis problem such as face, gender and ethnicity recognition. Local Binary Pattern (LBP), Local Phase Quantization (LPQ) [4], Histogram of oriented gradients (HOG), Binarized statistical image features (BSIF) [20] are used as feature extraction technique on both RGB and depth images of the dataset. The classification is performed using a SVM classifier. The results are evaluated on three Kinect face datasets namely, Facewarehouse [8], IIIT-D [15] and CurtinFaces [21]. They have shown the importance of depth information in gender classification task. Table 1.1 provides the summary of literature review.
- Ng *et al.* [27] have provided a comprehensive literature review of gender recognition using 2D face images and whole body. Commonly used feature extraction techniques are also discussed in the paper. A comparison of different gender classification methods has been provided by Mäkinen *et al.* [24]. They have performed experiments on FERET [30] and WWW image databases [24]. Comparison of methods such as SVM, Neural Network, LBP+SVM and Adaboost is provided with and without normalization of face images. They found that results obtained from all four methods are similar and no statistically significant difference is found. However, the accuracy improves when face images are normalized before classification. Also the combination of classifiers increases the accuracy over individual classifier.

Table 1.1: Summary of Literature Review

Paper	2D or 3D	Dataset	Technique	Classifier	Accuracy
Alexandre <i>et al.</i> [5]	2D	FERET and UND collection B	HOG, ULBP and decision fusion across features and image scales	SVM	99.07% (FERET) and 91.19% (UND)
Dhamecha <i>et al.</i> [13]	2D	Heterogenous dataset	PCA, PCA+LDA, PCA+SDA	Bayesian	86.47%
Shan <i>et al.</i> [33]	2D	LFW	Adaboost on Uniform LBP	SVM	94.81%
Makinen <i>et al.</i> [23]	2D	IMM Face and FERET	Automatic and manual alignment method. LBP, face pixels as features	SVM, NN	87.1% (Manual alignment)
Hyunh <i>et al.</i> [18]	3D	Eurecom	Weighted combination of ULBP and GLBP	SVM	90.38%
Lu <i>et al.</i> [22]	3D	UND and MSU	Range and Intensity information combination at decision level	SVM	91%
Boutellaa <i>et al.</i> [7]	3D	Face warehouse, IIIT D, Curtinface	LBP, LPQ, HOG, BSIF	SVM	87.7% (Curtinface)

Chapter 2

Preliminaries

Color and depth images obtained from Kinect devices of low resolution are often noisy. Therefore, images often need a preprocessing step before feature extraction. In this chapter, we have explained Layered bilateral filtering technique for preprocessing of depth images. In this section, we have also given a brief overview of different handcrafted and learned feature extraction techniques used in our study. Finally, Support Vector Machine which has been proven as an effective technique for gender classification has been described in this section.

2.1 Preprocessing: Layered Bilateral Filtering

The depth images acquired from the Kinect sensors are of relatively lower resolution and fidelity as compared to the RGB images. Hence, to deal with the problem of low quality depth images, we used layered bilateral filtering technique as introduced in [39]. This technique makes use of the registered high resolution color image to supersample and enhance the quality of the corresponding depth image. By using color image, it gets the true edges of the depth image. This approach for depth enhancement improves the input depth image by passing it through an iterative refinement module. The approach is divided into three steps: Iterative refinement module, bilateral filtering and sub pixel refinement.

- **Iterative Refinement Module**

A cost volume (\mathcal{C}) is first build upon the input depth image, D . Then bilateral filtering is applied on each slice of the iterative cost module, by making use of the high resolution color images, to produce a new cost volume \mathcal{C}' . The refined depth map is generated by passing the new cost volume through a sub pixel refinement stage. The cost function is calculated as:

$$\mathcal{C}_i(y, x, d) = \min \left(\eta * L, (d - D_i(y, x))^2 \right) \quad (2.1)$$

where η is constant, L is search range, d is depth candidate and $D_i(y, x)$ is currently selected depth.

• Bilateral Filtering

Bilateral filtering phase takes the input cost volume \mathcal{C} and registered color image and returns a new cost volume \mathcal{C}' . For each slice in \mathcal{C} , a patch of fixed size is taken and moved through all the pixels in a sliding window fashion. For each pixel in the current slice, a patch from the fixed neighbourhood of the slice is element-wise multiplied to the gaussian filter response of the R, G and B channels of the corresponding registered color image patch, (Equation 2.2). The resultant patch of the previous operations is then averaged over all its pixels and the center pixel of the patch is replaced by the same. Once the operations has been done on all the pixels of all the slices the new cost volume \mathcal{C}' is obtained which is then passed over to the sub pixel refinement stage.

$$F(y + u, x + v) = f_c(W_c(y, x, u, v)) f_s(W_s(u, v)) \quad (2.2)$$

$$f_c(x) = \exp \left(-\frac{|x|}{\gamma_c} \right) \quad (2.3)$$

$$f_s(x) = \exp \left(-\frac{|x|}{\gamma_s} \right) \quad (2.4)$$

$$\begin{aligned} W_c(y, x, u, v) = \frac{1}{3} (&|R(y + u, x + v) - R(y, x)| \\ &+ |G(y + u, x + v) - G(y, x)| \\ &+ |B(y + u, x + v) - B(y, x)|) \end{aligned} \quad (2.5)$$

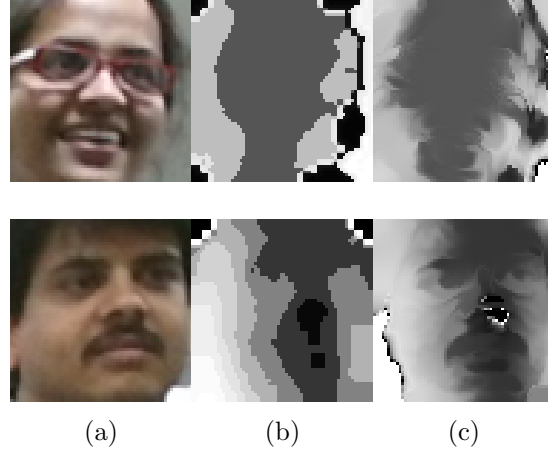


Figure 2.1: KaspAROV dataset a) RGB Image. b) Depth image before preprocessing. c) Depth image after applying Layered bilateral filter.

$$W_s(u, v) = \sqrt{u^2 + v^2} \quad (2.6)$$

- **Sub Pixel Refinement**

Since the input depth image in previous step is converted into a cost volume of fixed number of quantization levels, this can lead to discontinuities in the resultant depth image. In order to smoothen out the discontinuities a sub-pixel estimation algorithm is proposed in [39], based on quadratic polynomial interpolation.

Upon passing the low quality depth image and the registered color image through this iterative refinement module, we obtain the enhanced depth image, which we have used in gender classification. Figure 2.1 shows examples of this preprocessing technique.

2.2 Feature Extraction

An effective feature extraction technique is required for gender classification. Uniform LBP [28] which extracts the local texture features on RGB images and Gradient LBP [18] on depth images is widely used for gender classification. In this research, we have used these two feature extraction technique along with learned feature extraction technique, namely stacked denoising

autoencoder [36].

2.2.1 ULBP: Uniform Local Binary Pattern

Local Binary Pattern, as described by Ojala *et al.* [28], is calculated by taking the difference between centre pixel and neighbouring pixel in a 3×3 grid. Thresholding function is applied on the calculated values to obtain an 8 bit binary pattern. This 8 bit binary pattern is converted into decimal to get the value of the center pixel. An example is shown in Figure 2.2. A histogram of all such 256 possible values is used as feature descriptor when the neighborhood consist of 8 pixels. Thresholding function is given by:

$$s(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.7)$$

$LBP_{P,R}$ is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \quad (2.8)$$

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

As observed by Ojala *et al.* [28] some binary patterns appear more commonly than others. A uniform pattern is made of at most two 0-1 or 1-0 transitions. Each uniform pattern is given a separate bin in the computed histogram whereas, all other non-uniform patterns are given a single bin which helps in reducing the size of feature vector. In 3×3 grid with 8 neighbouring pixels, uniform pattern reduces the size of histogram to 59 from 256 in original LBP. This Uniform LBP is used in this research, by first dividing the face image into patches and then using Uniform LBP operator to each patch separately. Further we concatenate the histograms from all patches to obtain the final feature descriptor. As shown in Figure 2.3, the face image is divided into 16 patches and then Uniform LBP operator is applied on each patch to get final histogram of size $16 \times 59 = 944$. An illustration of LBP image obtained from original RGB

image is shown in Figure 2.4.

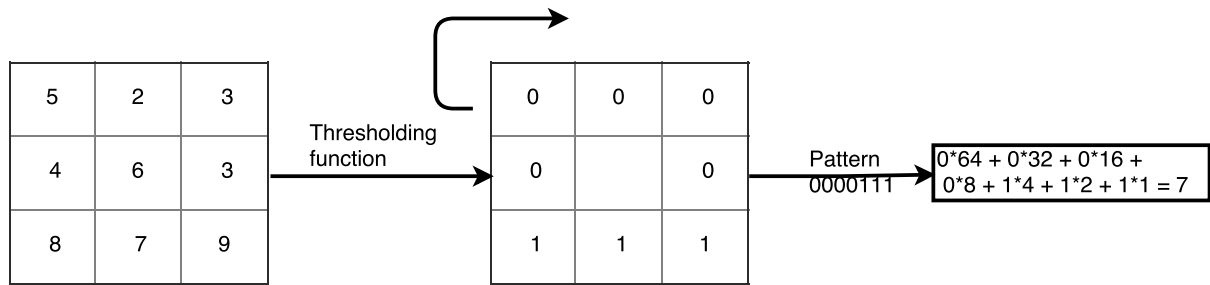


Figure 2.2: Example of LBP operator [28].

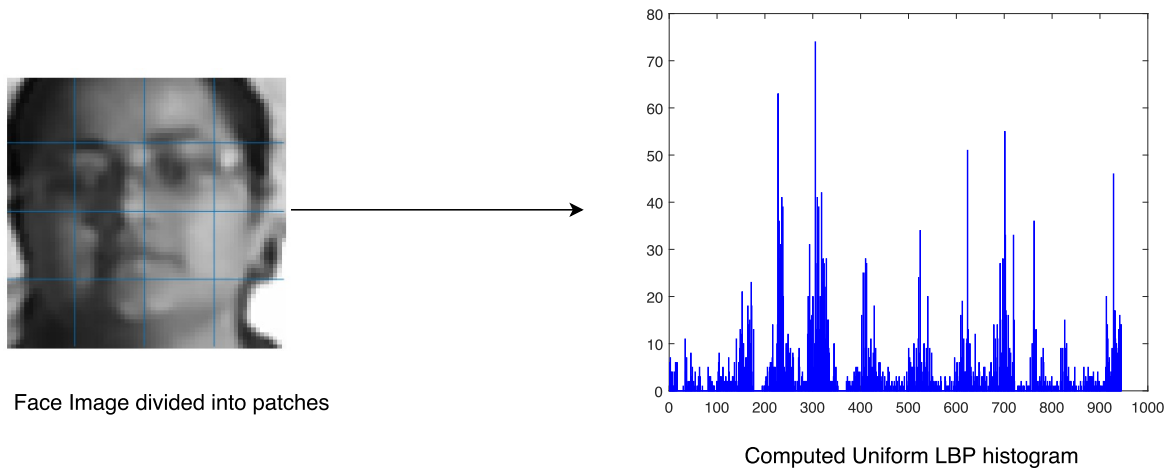


Figure 2.3: Uniform LBP histogram obtained from the face image divided into patches.



Figure 2.4: Illustration of LBP image obtained after applying LBP operator.

2.2.2 GLBP: Gradient Local Binary Pattern

Gradient LBP proposed by Huynh *et al.* [18] has been proven as an effective LBP based descriptor for facial depth images. The 8 neighbouring pixels of the original LBP of 3×3 grid correspond to eight orientations. The actual depth difference value is calculated between centre pixel and neighbouring pixels which creates eight depth difference images. Minimum value of this depth difference is stored as -8 and maximum value as 7. All the depth difference values below -8 are assigned value -8 and in same way all depth difference values above 7 are assigned value as 7. This way depth difference has 16 possible values. A histogram with 16 bins is computed along each orientation. An example of GLBP computation is shown in Figure 2.5. Also the Uniform LBP is applied on depth image and 59 bin histogram is extracted from it. As the 8 orientations of depth differences are pairwise symmetric, only half of the 8 orientations are used. The 4 histogram of depth difference and 1 histogram of uniform LBP are concatenated to get the final feature vector. In case of 3×3 grid of LBP having 8 neighbouring pixels, the size of final Gradient LBP feature vector will be $59 + 16 \times 4 = 123$ where 59 bins are of Uniform LBP histogram and 16 bins are of depth difference histogram calculated over four orientations. The main advantage of Gradient LBP is, it stores the actual depth difference value which preserves the sign of the depth difference and also the size of the feature vector is small. In this research, we have divided the depth image into 4×4 patches and then we have applied the Gradient LBP operator on each patch. Final histogram is the concatenation of all individual histograms of patches. An example of GLBP images obtained corresponding to four orientations from the depth image is shown in Figure 2.6.

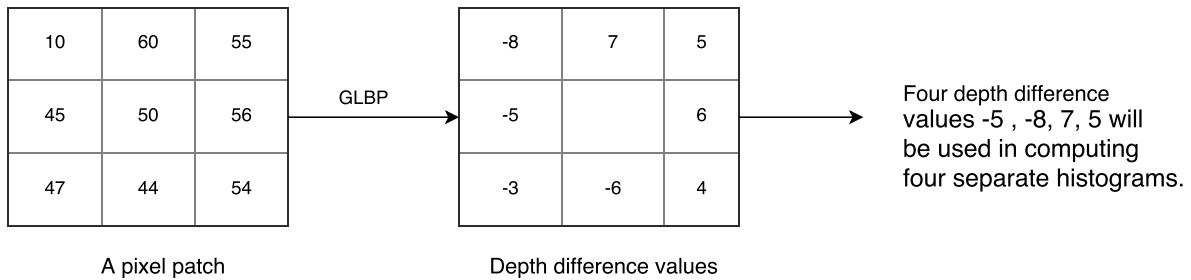


Figure 2.5: Example of Gradient LBP [18].

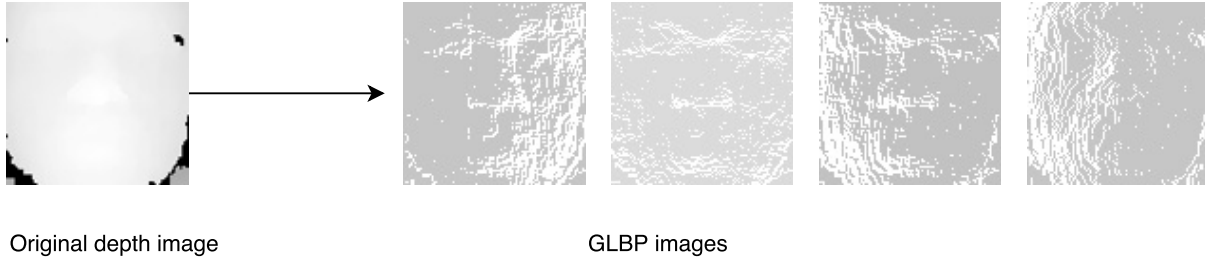


Figure 2.6: Illustration of GLBP images obtained corresponding to four orientations after applying GLBP operator.

2.2.3 SDAE: Stacked Denoising AutoEncoder

Autoencoder is an unsupervised learning algorithm which tries to reconstruct the given input vector. It has architecture similar to the neural network as shown in Figure 2.7. Autoencoder consist of encoding and decoding step. It first takes the input vector x and produces the hidden representation y (encoding step), as follows:

$$y = s(Wx + b) \quad (2.9)$$

where $s(\cdot)$ is a sigmoid activation function, W (Weight) and b (Bias) are parameters. Then Approximate reconstruction \hat{x} of the input vector x using the hidden representation y (decoding step) is calculated as follows,

$$\hat{x} = s(W'y + b') \quad (2.10)$$

where W' (Weight) and b' (Bias) are parameters. Autoencoder learns to minimize the reconstruction error between x and \hat{x} .

$$\operatorname{argmin} \|x - \hat{x}\|^2 \quad (2.11)$$

L_2 regularization is used to prevent overfitting. Cost function with added regularization term is given as:

$$\mathcal{C} = \frac{1}{2n} \sum_{k=1}^n \|x - \hat{x}\|^2 + \frac{\lambda}{2n} \sum_w w^2 \quad (2.12)$$

where n is the size of input data and λ is the regularization parameter. Typically for the purpose of classification, decoding layer is discarded in the end and the hidden representation is used

as new feature vector which is then given as input to any classifier. Denoising autoencoder is similar to normal autoencoder except that, instead of input vector x a noisy input x' is used. However, the reconstruction error is calculated between x and \hat{x} even though the input vector is a noisy input x' . It makes the autoencoder resistant to the noise in the input. Encoding and decoding equations in denoising autoencoder are represented as follows:

$$y = s(Wx' + b) \quad (2.13)$$

$$\hat{x} = s(W'y + b') \quad (2.14)$$

Stacked denoising autoencoder [36] is a concatenation of multiple layers of denoising autoencoder where output (hidden representation) of the previous layer is given as input to the next layer. The hidden representation of the last layer is used as new feature vector for classification.

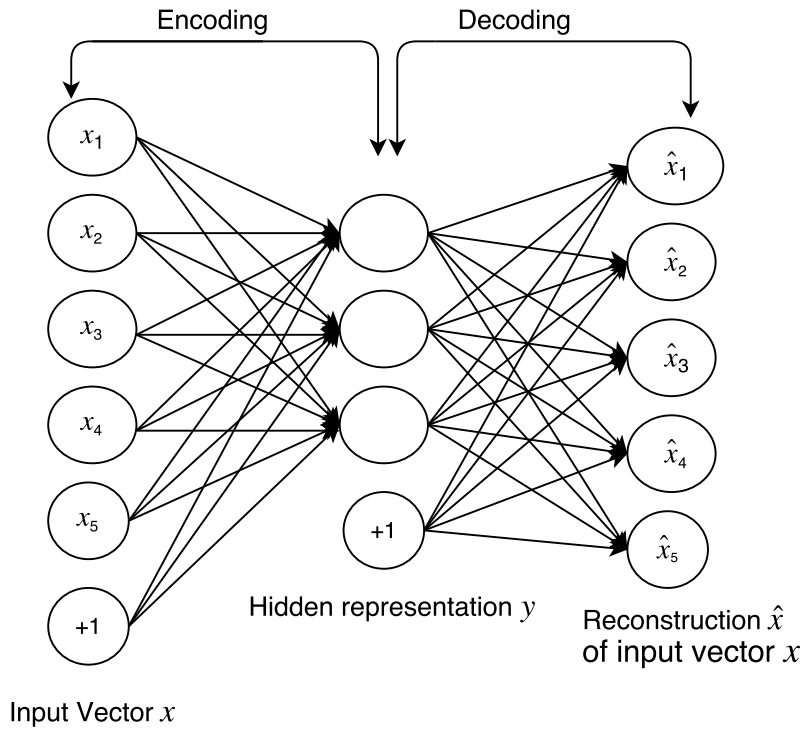


Figure 2.7: Illustration of Autoencoder [6].

2.3 Classification using Support Vector Machine

Support Vector Machine [11] is a supervised learning algorithm which tries to divide the input feature vector by finding the hyperplane such that mis-classification rate is minimum. SVM is primarily built for two class classification problem; however, it can be extended to multiclass classification. As shown in Figure 2.8, if the data is linearly separable, SVM tries to find the hyperplane separating the data points and if not, SVM maps the input data to higher dimensions by using the kernel trick.

Given a set of N data points, (x_i, y_i) where, x_i is a input feature vector and y_i is the corresponding label, SVM finds the hyperplane given by equation:

$$H_3 : w^T x + b \quad (2.15)$$

such that distance between two class distributions is maximum. In two class classification problem, equations of two boundary hyperplanes (H_1 and H_2) are:

$$H_1 : w^T x + b = 1 \quad (2.16)$$

$$H_2 : w^T x + b = -1 \quad (2.17)$$

The objective is to maximize the margin which is defined as maximum distance between the two boundary hyperplanes (H_1 and H_2). This margin is calculated as $\frac{2}{w^T w}$. This optimization function equivalently can also be stated as follows:

$$\text{argmin} \quad \frac{1}{2} w^T w \quad (2.18)$$

subject to the constraint that

$$y_i(w^T x_i + b) \geq 1 \quad \forall i \in \{1, 2, \dots, N\} \quad (2.19)$$

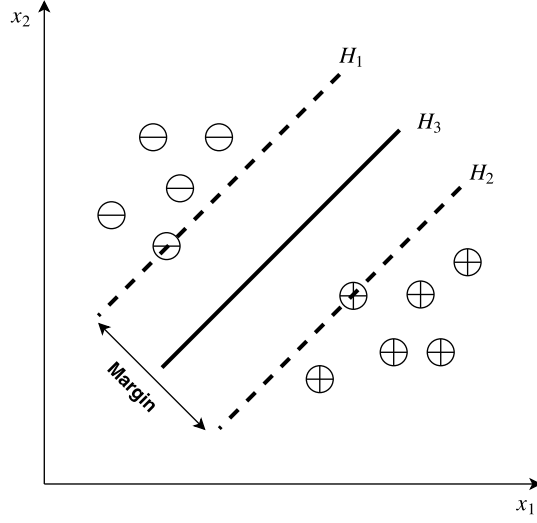


Figure 2.8: Maximum Margin Hyperplanes H_1 and H_2 , samples on margin hyperplane are support vectors [11].

which can be converted to solving the optimization problem having Lagrangian dual formulation as :

$$\operatorname{argmax} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (2.20)$$

where α_i are lagrangian multipliers, such that $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$.

After calculating α_i , one can infer from the Karush-Kuhn-Tucker (KKT) condition, that only support vectors (points on the boundary hyperplanes H_1 and H_2) have $\alpha_i \neq 0$. The final solution has the form:

$$w = \sum_{i=1}^N \alpha_i y_i x_i \quad (2.21)$$

where, x_i are support vectors. Slack variable ε_i is added to allow misclassification of difficult or noisy data points. The optimization function becomes:

$$\operatorname{argmin} \frac{1}{2} w^T w + C \sum_i \varepsilon_i \quad (2.22)$$

where, C is the cost parameter which controls the over-fitting, subject to the constraints

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i \quad \text{for } i = 1, \dots, N \quad (2.23)$$

Chapter 3

Gender Classification

3.1 Proposed Algorithm

Figure 3.1 shows the proposed architecture for gender classification. Depth images are preprocessed using layered bilateral filter and features are extracted using Gradient LBP. For a given RGB face image features are extracted using Uniform LBP and stacked denoising autoencoder. These features are provided as input to SVM and distance score metric is obtained from each SVM classifier. This distance score metric is then used for performing score level fusion.

3.1.1 Preprocessing

For Eurecom dataset, we have used similar preprocessing as mentioned in [18] for cropping the face image. Face images are cropped using nose as the center having width and height equal to twice the distance between left eye and right eye. No image enhancement is applied on the RGB images of either datasets. Depth images from KaspAROV dataset are enhanced using Layered bilateral filtering technique [3] [39]. As seen from the Figure 3.2, depth image quality after applying layered bilateral filtering has improved significantly. Morphological closing operation is performed on depth images of Eurecom Kinect Face Dataset to fill the holes as the depth images of Eurecom dataset are relatively better.

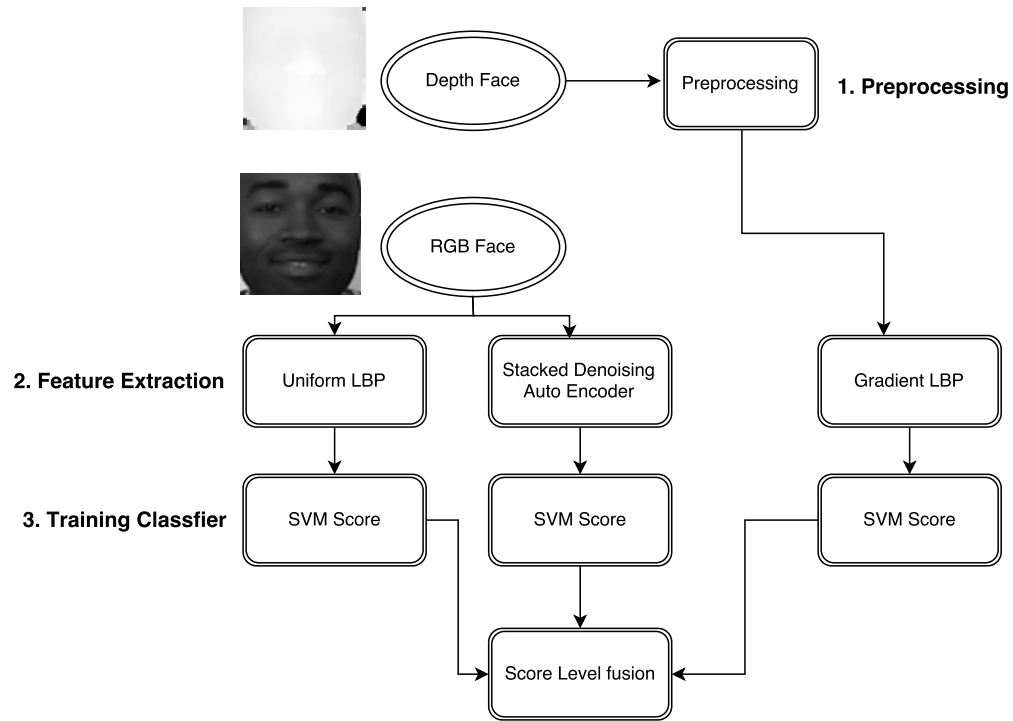


Figure 3.1: Algorithm for RGB-D face based gender classification.

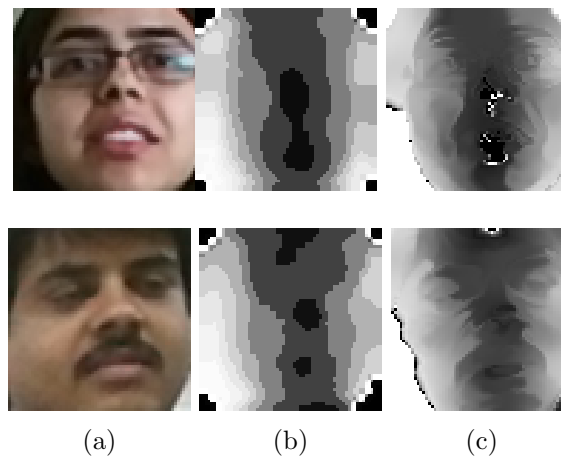


Figure 3.2: KaspAROV dataset a) RGB Image. b) Depth image before preprocessing. c) Depth image after applying Layered bilateral filter.

3.1.2 Feature Extraction and Classification

For feature extraction, we have used Uniform LBP and stacked denoising autoencoder on RGB images and Gradient LBP on depth images. SVM is used for two class classification i.e. (male, female). LIBSVM [9] is used for training SVM models for all experiments. This section describes the score level fusion of Uniform LBP, Gradient LBP and stacked denoising autoencoder.

- **ULBP:**

In the experiments, after converting the RGB face image to gray scale image, we have divided the gray scale image into 4×4 blocks and Uniform LBP (with $P=8$ and $R=1$) is applied on each block separately to form the histograms, and then we have concatenated these histograms to form the final feature vector. The size of final feature vector is $16 \times 59 = 944$. The feature vector is then given as input to SVM for classification.

- **SDAE:**

Stacked denoising autoencoder is trained on gray scale face images of size 64×64 pixels. These input images are first converted into vector form of $[1 \times 4096]$ and are given as input to the two layer stacked denoising autoencoder. In each hidden layer, number of hidden nodes are one fourth of the previous layer, therefore the first hidden layer contains 1024 nodes and the second hidden layer contains 256 nodes. Stacked denoising autoencoder is trained using the layer-by-layer greedy approach and the weights are learned using the backpropagation algorithm. The final hidden layer of size 256 gives us new feature representation of the given input vector. This final feature vector is then given to SVM for classification. The architecture is shown in Figure 3.3. Stacked denoising autoencoder implementation provided by deep learning toolbox [29] is used in all experiments. The same architecture is followed for both KaspAROV and Eurecom datasets; however, as the number of training images are less in Eurecom dataset, we have first trained SDAE on KaspAROV Kinect v1 dataset and used the same parameters for initialization of SDAE on Eurecom dataset. Then we learned the parameters using the training images of Eurecom dataset.

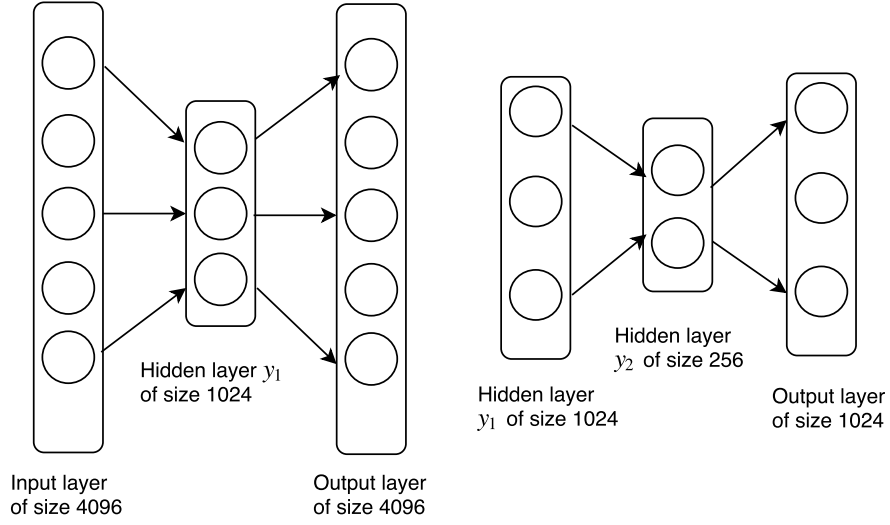


Figure 3.3: Architecture of two layer stacked denoising autoencoder.

- **GLBP:**

The depth face data is converted to gray scale depth image of size 64×64 pixels. Layered bilateral filtering technique is used for preprocessing to improve the quality of the depth face images. Thereafter, the depth face image is divided into 4×4 blocks and Gradient LBP operator is applied on each block to get the histograms. The histograms are then concatenated get the final feature vector of the depth images. The feature vector is then given to SVM for two class classification. In case of Eurecom dataset, only closing operation is performed as preprocessing step. Depth images are divided into 4×4 blocks and the same procedure is repeated.

- **Score level fusion of ULBP, GLBP and SDAE:**

We proposed an approach wherein, we have performed the SVM score level fusion of handcrafted features (Uniform LBP and Gradient LBP) and learned features obtained from stacked denoising autoencoder. The distance scores obtained after using SVMs is used for score level fusion. As suggested in literature [31], score level fusion can improve the performance of gender classification compared to feature extraction technique applied

individually. Overall, we have performed four score level fusion experiments, namely 1) Uniform LBP and Gradient LBP SVM score fusion, 2) Uniform LBP and SDAE SVM score fusion, 3) Gradient LBP and SDAE SVM score fusion and 4) Uniform LBP, Gradient LBP and SDAE SVM score fusion. We have found that all three techniques (ULBP, GLBP, SDAE) are statistically different using McNemar test [25].¹ In score level fusion, match scores from multiple algorithms are consolidated via sum rule, (Equation 3.2) to get a single score.

$$score = \sum_{i=1}^N \mathcal{W}_i S_i \quad (3.2)$$

where, S_i are individual scores of the classifier and \mathcal{W}_i are corresponding assigned weights. This single score is now used for classification.

3.2 Dataset Specification

To evaluate the proposed algorithm, we have used KaspAROV Kinect video dataset and Eurecom Kinect Face dataset. Both dataset contains RGB and Depth images. KaspAROV dataset is relatively large in terms of number of images and also challenging as face images are obtained from RGB-D surveillance videos. Table 3.1 provides the summary of the databases used.

3.2.1 KaspAROV Kinect Video Dataset

The dataset contains 108 subjects with two videos per subject on both Kinect version1 device and Kinect version2 device. Therefore, there are four videos per subject. These videos are taken in unconstrained environment with respect to pose, expression, illumination and distance. Figure 3.4 shows example frames from this database. Both RGB and depth information are captured in these videos. Frame resolution of RGB frames in Kinect v1 is 640×480 and for

¹McNemar test calculates the statistical correlation of two classifiers with regard to their classification performance. Two classifiers are said to be statistically different if the $\chi^2 \geq 3.8415$ at 0.05 significance level. The formula for McNemar test is given as follows:

$$\chi^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2} \quad (3.1)$$

where n_1 and n_2 are the number of misclassification made by one of the classifier while it is correctly classified by the other classifier.

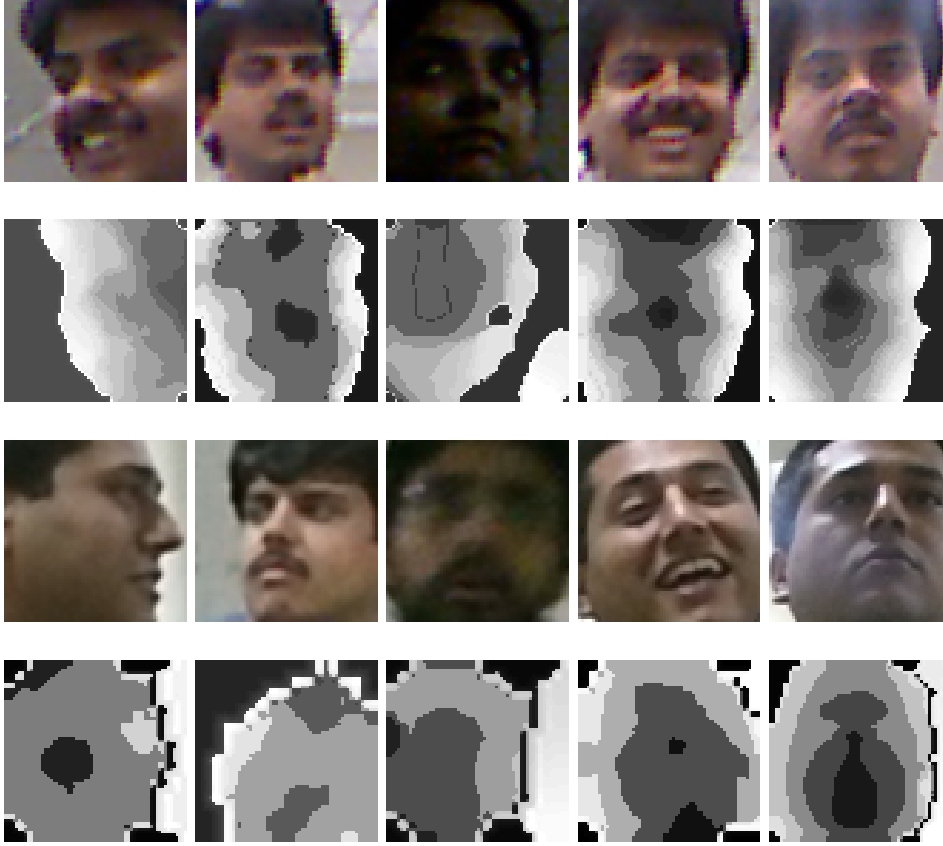


Figure 3.4: Sample RGB and depth images of KaspAROV dataset along with covariates. The first two rows contains images captured using Kinect v1 device and the last two rows contains images captured using Kinect v2 device.

Kinect v2 is 1920×1080 . Resolution of depth frame for Kinect v1 is 320×240 and for Kinect v2 is 512×424 . The dataset provides detected and cropped face images of the subjects from the video frames. All the face images are detected and resized to 64×64 resolution. For each RGB face image in the dataset, it also provides the corresponding raw depth image of size 64×64 .

3.2.2 Eurecom Kinect Face Dataset

Eurecom Kinect face dataset [26] contains RGB and depth images of 52 subjects out of which 38 are males and 14 are females, taken in two sessions using Kinect v1 device. Each session has nine face images with varying states. Nine states are neutral, light on, smile, left profile, right profile, open mouth, occlusion on eyes, occlusion on mouth, occlusion with paper. It also

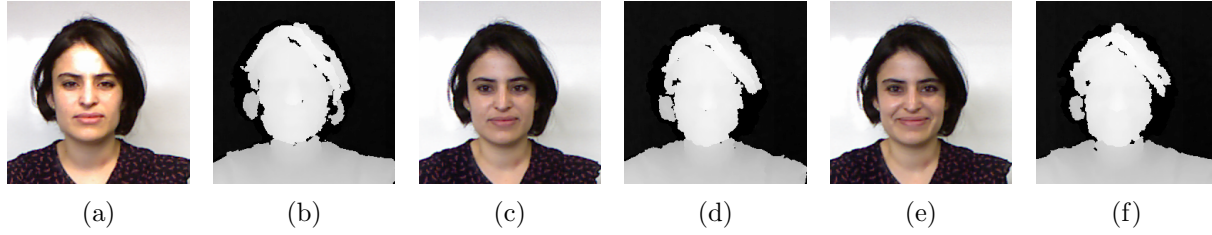


Figure 3.5: Sample images from Eurecom dataset a) RGB Image light on. b) Depth image light on. c) RGB Image neutral. d) Depth image neutral. e) RGB Image smile. f) Depth image smile.

Table 3.1: Summary of databases

Database	No. of male subjects	No. of female subjects	Total No. of images	Image resolution	Devices used
KaspAROV	79	29	22,251 (Kinect v1) 34,128 (Kinect v2)	64×64	Kinect v1 Kinect v2
Eurecom	38	14	936 (Kinect v1)	256×256	Kinect v1

provides 6 facial landmark points such as left and right eye, nose tip, chin, left side of the mouth and right side of the mouth in text file. Figure 3.5 shows the sample images from Eurecom dataset.

Chapter 4

Result and Conclusion

4.1 Experimental Protocol

KaspAROV dataset consists of 108 subjects out of which 29 are female and rest are male. In order to maintain unbiased class data division, we have taken only 29 male subjects for our task. Hence, the subset of KaspAROV dataset that we will be using in this work throughout consists 58 subjects. For our experiment, we have applied 3 fold cross-validation with random sub sampling where each fold consist of 15 male and 15 female subjects in training set, and 14 male and 14 female (unseen) subjects in testing set. 64×64 pixel size images have been used in the experiment.

Experimental protocol for Eurecom dataset is same as in [18]. Only three states (Neutral, Smile and Light on) of Eurecom dataset are used. Two experimental setup of the dataset are used: in the first setup, all the images from session 1 are used for training and images from session 2 for testing; in the second setup, half the number of males and females belonging to the both sessions are used for training while remaining half are used for testing. The details of training and testing split are shown in Table 4.1.

Table 4.1: Training and testing split of the dataset used for experiments.

	KaspAROV						Eurecom	
	Kinect v1			Kinect v2			Setup 1	Setup2
	Fold 1	Fold 2	Fold 3	Fold 1	Fold 2	Fold 3		
No. of Training images	9186	8038	7849	16747	13181	15487	156	156
No. of Testing images	5590	6738	6927	9266	12832	10526	156	156

4.2 Results

As stated in Section 3.1, the algorithms are evaluated on KaspAROV and Eurecom datasets using the experimental protocol mentioned in Section 4.1. The accuracies on KaspAROV dataset are reported in Tables 4.4 and 4.5, whereas accuracies on Eurecom dataset are reported in Table 4.6. Along with accuracy, Area Under Curve (AUC) and Equal Error Rate (EER) are also reported in Table 4.3 and Table 4.7. ROC (Receiver operating characteristic) for both the dataset are shown in Figure 4.1. The best accuracies are marked in bold. Following analysis can be drawn from the obtained results:

- ULBP outperforms SDAE and GLBP in all experimental setups except for KaspAROV Kinect v1 device where SDAE has produced marginally better accuracy than ULBP. The performance of GLBP on depth data of KaspAROV dataset is lower as compared to ULBP and SDAE on RGB data due to low quality of facial depth images obtained from RGB-D videos captured in unconstrained environment. For Eurecom dataset, the performance of GLBP is at par with ULBP and SDAE due to high quality depth images captured under constrained environment.
- As observed from Table 4.2, the χ^2 value between all three classifiers is greater than 3.8415 which proves that all three techniques are statistically different. Therefore, score level fusion can be performed using the weighted sum of scores of SVM.
- Score level fusion improves the accuracy in all experimental setups. In case of KaspAROV Kinect v1 device, score level fusion of ULBP, GLBP and SDAE provides the best average accuracy of 93.02% with an Equal error rate of 7.08% over all three folds. Similarly for

Table 4.2: χ^2 value using McNemar Test.

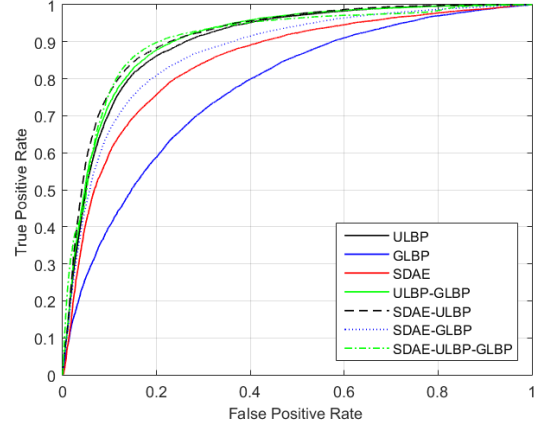
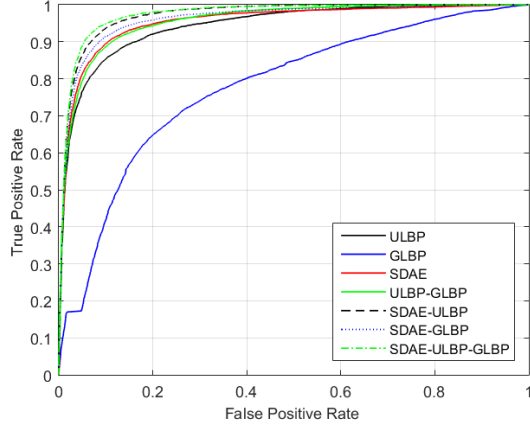
Classifiers	χ^2 value
ULBP-SDAE	479.89
SDAE-GLBP	41.91
ULBP-GLBP	478.69

Table 4.3: Area Under Curve (AUC) and Equal error rate (EER - %) reported for Kinect v1 and Kinect v2 of KaspAROV dataset.

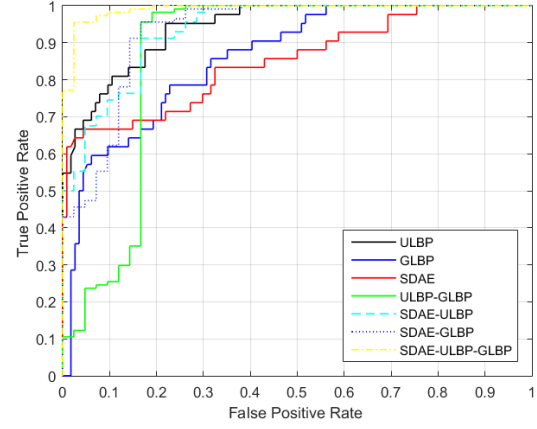
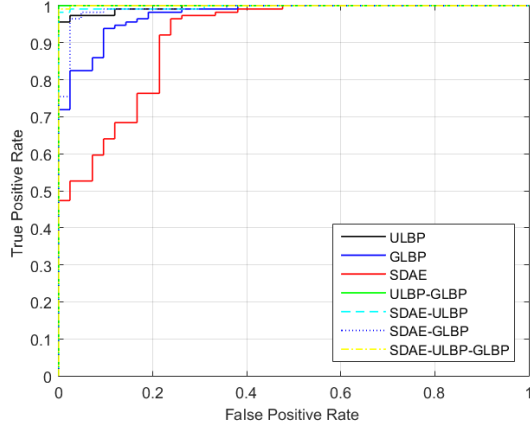
	Kinect v1		Kinect v2	
	AUC	EER	AUC	EER
ULBP (Gray)	94.33 \pm 1.11	12.54 \pm 2.38	89.22 \pm 1.09	17.83 \pm 0.57
GLBP (Depth)	78.30 \pm 1.43	27.79 \pm 0.38	77.71 \pm 2.88	28.91 \pm 2.85
SDAE (Gray)	95.72 \pm 1.71	9.92 \pm 2.06	83.57 \pm 3.08	23.00 \pm 2.61
ULBP+GLBP	95.52 \pm 0.35	10.88 \pm 1.57	90.11 \pm 1.62	16.74 \pm 0.34
ULBP+SDAE	97.21 \pm 0.47	8.00 \pm 0.09	90.29 \pm 0.24	16.38 \pm 1.29
SDAE+GLBP	96.32 \pm 1.75	8.91 \pm 2.24	86.60 \pm 1.39	20.28 \pm 1.55
SDAE+ULBP+GLBP	97.51\pm0.85	7.08\pm0.90	91.63\pm0.90	15.36\pm0.81

KaspAROV Kinect v2 device the best average accuracy of 84.97% with an Equal error rate of 15.36% is obtained by fusion of all three descriptors. Accuracy for Kinect v2 device is lower compared to Kinect v1 because Kinect v2 has more number of detected faces compared to Kinect v1 which results in more variations of face images. Also for Eurecom dataset the combination of all three descriptors in experimental setup 2 outperforms combination of ULBP and GLBP. The best accuracy of 95.51% with an Equal error rate of 4.76% is obtained on Eurecom dataset with experimental setup 2.

- Score level fusion also improves the individual class accuracies of males and females. Eurecom dataset consist of 38 males and 14 females which clearly makes any classifier biased towards male class; however, as seen from Table 4.6 in setup 2, female accuracy is increased to 85.71% from 73.80% (fusion of ULBP and GLBP) when we perform score level fusion of ULBP, GLBP and SDAE. Table 4.8 shows some sample examples of correctly classified or misclassified face images by ULBP and SDAE.



(a) ROC on KaspAROV Kinect v1 over three folds (b) ROC on KaspAROV Kinect v2 over three folds



(c) ROC on Eurecom using experimental setup 1 (d) ROC on Eurecom using experimental setup 2

Figure 4.1: ROC on KaspAROV and Eurecom dataset, positive class is female and negative class is male.

Table 4.4: Accuracy (%) on KaspAROV dataset Kinect v1 device using 3 fold cross validation.

	Kinect v1		
	Male	Female	Overall
ULBP (Gray)	94.67 \pm 2.99	75.76 \pm 11.48	88.54 \pm 1.59
GLBP (Depth)	72.09 \pm 4.93	72.31 \pm 3.66	72.34 \pm 1.91
SDAE (Gray)	88.00 \pm 4.50	91.40 \pm 2.99	89.23 \pm 2.42
ULBP+GLBP	95.02 \pm 3.11	77.89 \pm 10.97	89.46 \pm 1.45
ULBP+SDAE	94.01 \pm 2.84	88.21 \pm 5.88	92.15 \pm 0.43
SDAE+GLBP	89.44 \pm 5.01	91.98 \pm 3.37	90.41 \pm 2.92
SDAE+ULBP+GLBP	93.92 \pm 3.07	91.07 \pm 4.17	93.02\pm1.01

Table 4.5: Accuracy (%) on KaspAROV dataset Kinect v2 device using 3 fold cross validation.

	Kinect v2		
	Male	Female	Overall
ULBP (Gray)	79.82±3.01	85.01±4.63	82.64±1.54
GLBP (Depth)	66.67±3.14	75.07±2.23	70.47±1.67
SDAE (Gray)	76.18±4.47	77.44±8.19	77.66±2.91
ULBP+GLBP	80.47±3.15	86.36±3.19	83.60±0.75
ULBP+SDAE	83.26±3.08	84.18±6.08	84.40±1.77
SDAE+GLBP	77.88±4.69	80.95±6.47	79.98±2.11
SDAE+ULBP+GLBP	82.50±4.72	86.05±6.44	84.97±1.43



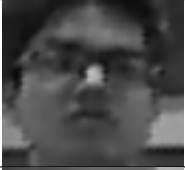




Table 4.6: Accuracy (%) on Eurecom dataset Kinect v1 device with two experimental protocols.

	Setup1			Setup2		
	Male	Female	Overall	Male	Female	Overall
ULBP (Gray)	99.12	88.09	96.15	98.24	61.9	88.46
GLBP (Depth)	97.36	80.95	92.94	95.61	57.14	85.25
SDAE (Gray)	98.24	66.66	89.74	97.36	61.90	87.82
ULBP+GLBP	100	97.61	99.35	100	73.8	92.94
ULBP+SDAE	98.24	100	98.71	98.24	71.42	91.02
SDAE+GLBP	98.24	95.23	97.43	99.12	71.42	91.66
SDAE+ULBP+GLBP	99.12	100	99.35	99.12	85.71	95.51

Table 4.7: Area Under Curve (AUC) and Equal Error Rate (EER - %) of Eurecom dataset Kinect v1 device with two experimental protocols.

	Setup1		Setup2	
	AUC	EER	AUC	EER
ULBP (Gray)	99.52	2.38	83.33	5.75
GLBP (Depth)	97.47	9.52	76.19	14.05
SDAE (Gray)	91.42	21.43	73.81	14.90
ULBP+GLBP	100	0.00	86.97	16.67
ULBP+SDAE	99.71	2.38	93.59	16.67
SDAE+GLBP	99.02	2.38	92.90	14.29
SDAE+ULBP+GLBP	99.98	0.00	99.06	4.76

Table 4.8: Images correctly classified (\checkmark) and misclassified (\times) by ULBP and SDAE

		ULBP	SDAE
		\checkmark	\checkmark
		\checkmark	\times
		\times	\checkmark
		\times	\times

4.3 Conclusion and Future Work

In this research, we have utilized RGB and depth information obtained from Kinect sensor for gender classification. We have also shown the effectiveness of learned feature extraction techniques, namely stacked denoising autoencoder, in gender classification problem. In this study, we have also shown that score level fusion of Uniform LBP, Gradient LBP and stacked denoising autoencoder improves the performance of gender classification. Experiments are performed on KaspAROV and Eurecom Kinect datasets and state-of-the-art gender classification accuracies are achieved. KaspAROV dataset is a challenging dataset as the face images are extracted from RGB-D videos captured in surveillance settings. Therefore, there is scope for further improvement in performance of gender classification in such settings. The accuracy on low quality depth images can be improved further by developing new preprocessing and feature extraction techniques. Also, the effectiveness of other learning based feature extraction techniques such as convolution neural network and deep belief network can be evaluated.

Bibliography

- [1] F. Tarres and A. Rama. GTAV face database, 2011. <http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm>.
- [2] Georgia tech face database, 2011. <ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/>.
- [3] Matlab Toolbox for Depth Enhancement. <https://bitbucket.org/shshzaa/depth-enhancement>.
- [4] AHONEN, T., RAHTU, E., OJANSIVU, V., AND HEIKKILÄ, J. Recognition of blurred faces using local phase quantization. In *19th International Conference on Pattern Recognition*, (2008), IEEE, pp. 1–4.
- [5] ALEXANDRE, L. A. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters* 31, 11 (2010), 1422–1427.
- [6] BENGIO, Y., LAMBLIN, P., POPOVICI, D., LAROCHELLE, H., ET AL. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, (MIT Press 2007), 153–160,.
- [7] BOUTELLAA, E., HADID, A., BENGHERABI, M., AND AIT-AOUDIA, S. On the use of kinect depth data for identity, gender and ethnicity classification from facial images. *Pattern Recognition Letters* 68 (2015), 270–277.
- [8] CAO, C., WENG, Y., ZHOU, S., TONG, Y., AND ZHOU, K. Facewarehouse: a 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425.
- [9] CHANG, C.-C., AND LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [10] CHANG, K. I., BOWYER, K. W., AND FLYNN, P. J. Multimodal 2d and 3d biometrics for face recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures* (2003), IEEE, pp. 187–194.

- [11] CORTES, C., AND VAPNIK, V. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.
- [12] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2005), vol. 1, IEEE, pp. 886–893.
- [13] DHAMECHA, T., SANKARAN, A., SINGH, R., VATSA, M., ET AL. Is gender classification across ethnicity feasible using discriminant functions? In *International Joint Conference on Biometrics (IJCB)* (2011), IEEE, pp. 1–7.
- [14] FLYNN, P. J., BOWYER, K. W., AND PHILLIPS, P. J. Assessment of time dependency in face recognition: An initial study. In *Audio-and Video-Based Biometric Person Authentication* (2003), Springer, pp. 44–51.
- [15] GOSWAMI, G., VATSA, M., AND SINGH, R. Rgb-d face recognition with texture and attribute features. *IEEE Transactions on Information Forensics and Security* 9, 10 (2014), 1629–1640.
- [16] GUPTA, S., CASTLEMAN, K. R., MARKEY, M. K., AND BOVIK, A. C. Texas 3d face recognition database. In *IEEE Southwest Symposium on Image Analysis & Interpretation* (2010), IEEE, pp. 97–100.
- [17] HUANG, G. B., RAMESH, M., BERG, T., AND LEARNED-MILLER, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [18] HUYNH, T., MIN, R., AND DUGELAY, J.-L. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Computer Vision-ACCV 2012 Workshops* (2013), Springer, pp. 133–145.
- [19] JOLLIFFE, I. *Principal component analysis*. Wiley Online Library, 2002.
- [20] KANNALA, J., AND RAHTU, E. Bsif: Binarized statistical image features. In *21st International Conference on Pattern Recognition* (2012), IEEE, pp. 1363–1366.
- [21] LI, B. Y., LIU, W., AN, S., AND KRISHNA, A. Tensor based robust color face recognition. In *21st International Conference on Pattern Recognition* (2012), IEEE, pp. 1719–1722.
- [22] LU, X., CHEN, H., AND JAIN, A. K. Multimodal facial gender and ethnicity identification. In *Advances in Biometrics*. Springer, 2005, pp. 554–561.

- [23] MÄKINEN, E., AND RAISAMO, R. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 3 (2008), 541–547.
- [24] MÄKINEN, E., AND RAISAMO, R. An experimental comparison of gender classification methods. *Pattern Recognition Letters* 29, 10 (2008), 1544–1556.
- [25] MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (1947), 153–157.
- [26] MIN, R., KOSE, N., AND DUGELAY, J.-L. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 11 (2014), 1534–1548.
- [27] NG, C. B., TAY, Y. H., AND GOI, B.-M. Recognizing human gender in computer vision: a survey. In *PRICAI 2012: Trends in Artificial Intelligence*. Springer, 2012, pp. 335–346.
- [28] OJALA, T., PIETIKÄINEN, M., AND MÄENPÄÄ, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 971–987.
- [29] PALM, R. B. Prediction as a candidate for learning deep hierarchical models of data. *Technical University of Denmark* (2012).
- [30] PHILLIPS, P. J., MOON, H., RIZVI, S., RAUSS, P. J., ET AL. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 10 (2000), 1090–1104.
- [31] ROSS, A. A., NANDAKUMAR, K., AND JAIN, A. K. *Handbook of multibiometrics*, vol. 6. Springer Science & Business Media, 2006.
- [32] SCHOLKOPFT, B., AND MULLERT, K.-R. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX 1* (1999), 1.
- [33] SHAN, C. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters* 33, 4 (2012), 431–437.
- [34] SIM, T., BAKER, S., AND BSAT, M. The cmu pose, illumination, and expression (pie) database. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition* (2002), IEEE, pp. 46–51.
- [35] STEGMANN, M. B., ERSBØLL, B. K., AND LARSEN, R. Fame-a flexible appearance modeling environment. *IEEE Transactions on Medical Imaging* 22, 10 (2003), 1319–1331.

- [36] VINCENT, P., LAROCHELLE, H., LAJOIE, I., BENGIO, Y., AND MANZAGOL, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11 (2010), 3371–3408.
- [37] VIOLA, P., AND JONES, M. J. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.
- [38] WOLF, L., HASSNER, T., AND TAIGMAN, Y. Similarity scores based on background samples. In *Computer Vision–ACCV 2009*. Springer, 2010, pp. 88–97.
- [39] YANG, Q., YANG, R., DAVIS, J., AND NISTÉR, D. Spatial-depth super resolution for range images. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007), IEEE, pp. 1–8.
- [40] ZHU, M., AND MARTINEZ, A. M. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 8 (2006), 1274–1286.