RGB-D Face Recognition in Surveillance Videos

Anurag Chowdhury

IIIT-D-MTech-CS-GEN-14-002

June 23, 2016

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

Keywords : RGB-D, Kinect, Face Detection, Face Recognition, Deep Learning

Certificate

This is to certify that the thesis titled **"RGB-D Face Recognition in Surveillance Videos"** submitted by **Anurag Chowdhury** 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

Biometric analysis of surveillance videos carries inherent challenges in form of variations in pose, distance, illumination and expression. To address these variations, different methodologies are proposed, including utilizing temporal and 3D information. With the introduction of consumer level depth capturing devices such as Microsoft Kinect, research has been performed in utilizing low cost RGB-D depth data for characterizing and matching faces.

Face detection being the foremost task in face biometric pipeline has a cascading effect on the performance of any face recognition system that follows. Face detection algorithms generally work best for frontal face images with good illumination and low standoff distance. Developing a face detection system robust to the variates of a surveillance scenario is a highly challenging task. Recognition of the detected faces in surveillance scenarios is a challenging task owing to high variance in pose, illumination, expression and resolution. Also, the quality of depth data in RGB-D videos deteriorates with increase in standoff distance, thus adding to the challenges of RGB-D face recognition.

This research introduces the KaspAROV RGB-D video face database which provides face videos and images from Kinect device for over 100 subjects. The database encompasses challenges such as pose, distance, and illumination. Further, a novel face detection system for RGB-D videos taken in unconstrained scenario is proposed. The proposed system makes use of human body detection in color images and fuses it with the corresponding depth map to provide a robust solution for face detection at a distance in RGB-D videos. For recognizing the detected faces we introduce a RGB-D face recognition algorithm which can also work with only RGB probe images in absence of depth data in probe images . The proposed algorithm generates a shared representation from RGB images which contains discriminative information from both the RGB and depth images. This representation is much more discriminative than the RGB images as it gives substantially higher identification accuracy than a conventional fusion based RGB-D recognition pipeline.

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. It has been a real pleasure and a blessing to have been placed under their tutelage here at IIIT Delhi. Their impeccable attention to detail and expert guidance can only be matched by their unwavering dedication to the cause of conducting quality research. Their support and guidance all these times have enabled me to gain more knowledge and build a resilient research temperament. I would like to specifically thank all the mates at Image Analysis and Biometrics lab with whom I have shared a major portion of my time here at IIIT Delhi. Their friendly advice and professional wisdom has always helped me gain new perspective of my research work. This section can not be complete without a vote of thanks to academic department for their help and never ending support.

Dissemination of Research Results

1. **Chowdhury, A.**, Ghosh, S., Vatsa, M., and Singh, R. RGB-D Face Recognition via Reconstruction based Shared Representation. IEEE International Conference on Biometrics: Theory, Applications, and Systems 2016 *Accepted*.

2. Chhokra, P., **Chowdhury, A.**, Goswami, G., Vatsa, M., and Singh, R. KaspAROV: Unconstrained Kinect Video Face Database. IEEE Transactions on Information Forensics and Security 2016 *Under Review*.

3. **Chowdhury, A.**, Chhokra, P., Vatsa, M., and Singh, R. Face detection at a distance in RGB-D videos. IEEE International Conference on Identity, Security and Behaviour Analysis 2017 *Under Internal Review*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Research Motivation

Face recognition aims to establish identity using the face image of a person. Face is an easily accessible trait and requires little user cooperation to capture [28]. However, in a non-cooperative scenario such as surveillance, the covariates that can deter face recognition performance cannot be controlled. Face images captured in such environments can pose a significant challenge to existing recognition algorithms that perform significantly worse in unconstrained conditions [5]. Besides improving the various facets of a face recognition algorithm, such as feature extraction and matching methodologies, image acquisition itself has an important role to play in the overall efficiency of a recognition framework. For instance, near infrared (NIR) imagery can help in cases with poor illumination and 3D images can provide much more information about a face which help in both detection and recognition. However, such data can be expensive to acquire due to sensor cost and deploying such a system is not feasible in most cases, hence limiting its applicability.

A potential application of low cost RGB-D sensors and recognition algorithms is in video surveillance. Dedicated 3D sensors have high associated costs which prevent their usage in such scenarios as compared to a low cost Kinect version 2 device. While offering substantially low costs as compared to pure 3D sensors, it offers RGB-D and infrared data as opposed to pure 2D sensors.

Figure 1.1: Sample RGB and Depth frames from KaspAROV [27] RGB-D video dataset

.

Besides capturing RGB-D and infrared data, it offers a wide field of vision and audio recording capabilities, all of which are functionalities required in a surveillance device.

Data captured using RGB-D imaging devices like Microsoft Kinect deploy a RGB camera along with a depth sensor for capturing RGB and depth images of a scene in synchronization. A sample RGB and depth video frame from captured using Kinect v2 could be seen in Figure 1.1. Sample images from Kinect device of a person's face under varying facial poses and expressions could be seen in Figure 1.2. Raw depth data is usually captured in form of a 16 bit image where each image pixel contains distance of the corresponding world point from the camera in millimeters. For representation sake an 8 bit image of the same scene could be generated as seen in Figure 1.1. Higher the intensity value of a pixel in the depth image, lower is its depth value i.e. it is placed closer to the camera than the pixels with lower intensities.

Face recognition systems for RGB-D surveillance videos is an open research problem in the biometrics and computer vision community. Development of such systems would enable effective monitoring of challenging surveillance scenarios like banks, airports, secure facilities etc. Image processing and machine learning algorithms have a major role to play in development of such systems. Application of Machine learning and Image processing in face recognition system could be sub-divided under following heads:

- **Preprocessing:** The acquired frames from the RGB-D cameras are usually ridden with sparse salt and pepper noise and they also suffer from low contrast problems. Image

Figure 1.2: Sample RGB-D images of a person showing facial different expressions taken from IIITD RGBD face database [26].

processing techniques such as median filtering and morphological operations are used to remove salt and pepper noise. Also histogram equalization is used in certain cases to deal with issues of contrast.

- **Face Detection:** Faces in surveillance video frames are often captured at a high standoff distance and non-frontal poses. Also, the frames suffer from sensor and environmental noise. Detecting faces in such situations is a arduous image processing task and its effectiveness has a direct impact on the overall performance of the recognition system. The Viola-Jones face detector [43] detects face images well in constrained conditions where the face images are mostly in frontal/semi-frontal pose and are well illuminated. The Everingham face detector [12] works better than standard Viola-Jones face detector [43] in unconstrained scenarios but it still fails to capture a huge proportion of the face images in the frames. Face detection systems making use of depth data could be developed for improved face detection performance in case of RGB-D data.

- **Face Image Post-processing:** Since the segmented face images from the video frames are very small in pixel resolution. Most of the finer details of face images crucial towards face recognition is lost. In order to restore the quality of face images , super-resolution based techniques could be used.

- **Face Recognition:** The processed face images are then labeled manually into their identity based classes and are divided into training and testing sets. Several machine learning

based models are learnt from the labeled training data and is then tested on testing set to predict the identity of the data samples in the testing set.

1.2 Literature Review

In recent years, there has been increased focus on usage of RGB-D cameras for development of 3D scene understanding and object detection algorithms [33]. Kinect sensor based surveillance systems have also been deployed for border control [15]. Usage of Kinect sensors for indoor surveillance systems is an interesting research problem due to it's capability of capturing RGB, Depth and NIR footage from a single camera unit. Recently, decreased cost of depth sensors has made it feasible to be used in surveillance activities and has consequently led to increased interest in RGB-D face detection and recognition. In presence of covariates such as pose and illumination, it has been shown that 3D images perform better than their 2D counterparts in face recognition [30]. The depth map provides additional discriminative information which enhances the recognition performance.

Face is one of the highly investigated biometric modality. A large number of methods exist in literature [48] for identification and verification of face images under controlled scenarios. Introduction of covariates such as distance from the camera, pose, illumination, and resolution makes the problem challenging and requires novel and sophisticated algorithms. With the advent of depth sensors, Han et al. [21] introduced the use of utilizing 3D images (RGB and Depth) have been introduced for face recognition.

RGB-D images have been used in a variety of applications including Indoor scene segmentation [39], Human action recognition [44], Face anti-spoofing [11], Head pose estimation [13], Object recognition [35], Object discovery [29], Face detection [22], Gender recognition [25].

1.2.1 Previous Work on RGB-D Face Detection

The first step in a face recognition pipeline is face detection. The variations in pose, illumination, and distance in surveillance videos increase the complexity of the task thereby making face

detection in surveillance scenario a challenging research problem. To improve the detection performance, research have explored usage of depth cues. However, majority of the research work in RGB-D based face detection has been focused on improving the speed of face detection system. Walker et al. [6] used depth information to select areas in the image where there are higher probability of faces to be found, thus improving the speed of the face detector. Wu et al. [46] used stereo camera to estimate depth information of the scene and then used it to accelerate face detection while pruning false positives.

1.2.2 Previous Work on RGB-D Face Recognition

Inspired by the low cost availability of multi-modal data, researchers have proposed several algorithms that utilize RGB-D data to perform face recognition [4, 20, 24, 31, 34]. Several researchers have explored the applicability of image fusion algorithms to improve the performance where multimodal information is available. For instance, Singh et al. proposed wavelet fusion based algorithm to combine images from multiple spectrums [41]. However, these algorithms either have fixed weighting scheme to generate the fused image or utilize quality assessment to select local image regions and their weights. Recent introduction of multimodal deep learning paradigms [40, 42] has provided the researchers a new spectrum of applications where multiple modalities are involved and not all the modalities are required during testing to perform an accurate match.

Existing RGB-D algorithms utilize the depth data for improving various facets of face recognition such as face detection, landmark detection, image alignment, achieving pose invariance, and extracting additional discriminative information.

Most of the well known RGB-D face recognition algorithms have utilized the discriminative information from both RGB and depth images using sophisticated information fusion algorithms. These existing algorithms demonstrate the effectiveness of RGB-D face data in improving recognition performance. However, all but one of these algorithms have been developed with data obtained from a Kinect version 1 device.

- Li *et al.* [31] have explored the use of depth map in preprocessing as well as feature

extraction.

- Li *et al.* [31] presented a face recognition algorithm from low resolution 3D images. Texture transformation and a sparse coding based reconstruction method is used to perform face matching.

- Beretti *et al.* [4] have utilized the depth map to create a 3D face model after applying super-resolution techniques.

- Mantecon *et al.* [34] have devised a new Depth Local Quantized Pattern (DLQP) descriptor which extracts features only from the depth data and performs recognition using these features with a SVM classifier.

- Hsu *et al.* [24] have not utilized the depth information for feature extraction but instead rely on it for facial landmark detection and pose estimation followed by feature extraction and matching based on pose normalized color images.

- Goswami *et al.* [20] proposed using a descriptor based on entropy of RGB-D images and saliency feature from the RGB image. Geometric facial features are also utilized and a sophisticated fusion method is proposed to use the RGB-D images for face recognition.

- Li *et al.* [32] proposed a 3D keypoint based face matching algorithm using multi-task sparse representation.

- Elaiwat *et al.* [10] used a multimodal keypoint detector for identifying keypoints on a 3D surface, and both texture and 3D local features are utilized.

- Ming [37] proposed a regional bounding spherical descriptor for facial recognition and emotional analysis which uses regional and global regression mapping for classification.

1.3 Research Contributions

RGB-D sensors could be used to solve the challenges of pose and expression in challenging unconstrained surveillance scenarios. Availability of real time depth data of the scenes could be

used in developing algorithms which could help improve current state of the art face detection and recognition performances in challenging scenarios. Owing to the extensive research potential and large amount of real life applications we have taken up face detection and recognition in using RGB-D data in surveillance videos as our prime area of research.

The key contributions of this research are:

- Create and benchmark a RGB-DI video face dataset, titled KaspAROV, collected using Kinect sensors(both version 1 and 2): The dataset consists of video frames of over 100 subjects taken in surveillance like scenarios. Kinect v2 data consists of RGB, Depth and NIR face images, while Kinect v1 data consists of only RGB and Depth face images.

- Develop a novel RGB-D face detection algorithm: This algorithm leverages the fact that detecting human bodies from a distance is easier than detecting faces. The proposed algorithm first detects humans bodies in the RGB images and makes use of the corresponding depth data to segment tightly cropped human bodies from the scene. The upper portion of the segmented human bodies are then scanned for presence of faces for detecting faces in the scene.

- Propose a novel "shared representation based reconstruction network" for RGB-D face recognition: Here, we pre-trained a cross-modality reconstruction network which learns a mapping between two input modalities, RGB and depth data. The network is trained using RGB and Depth data from the dataset under consideration, during testing only RGB data is used to reconstruct depth image which encodes discriminative properties of both RGB and depth data. The reconstructed depth images are then used for classification.

In the next chapter, we discuss the KaspAROV RGB-D video face dataset. In third chapter we propose the KarPOV face detection system in RGB-D videos at a distance and in the same chapter we also put forth the proposed shared representation based RGB-D face recognition system. Finally in the last chapter we summarize the contributions of the research works carried out as part of this thesis and we also talk about future extensions of the work done here.

Chapter 2

KaspAROV: Unconstrained Kinect Video Face Database

Unconstrained face recognition poses several challenges to existing algorithms and to address these variations, different methodologies are proposed, including utilizing video and 3D information. With the introduction of consumer level depth capturing devices such as Microsoft Kinect, research has been performed in utilizing low cost RGB-D depth data for characterizing and matching faces. Recently, next generation of Kinect device, the Kinect version 2, has been released which provides higher resolution color and depth images at a comparable sensor cost. This research work introduces the KaspAROV RGB-D video face database which provides face videos and images from both versions of the Kinect device for over 100 subjects which will be made available to the research community. The database encompasses challenges such as pose, distance, and illumination. We include baseline results using a few existing algorithms and provide standard experimental protocols for ease of comparative evaluation on the database in future research.

2.1 Introduction

Existing RGB-D databases encompass only a small variety of covariates (mainly pose, illumination, and expression only) and are captured in largely constrained conditions. These databases are limited in either challenges, subjects, or samples. For example, while the CurtinFaces [31] database contains pose, illumination, and expression covariates, it has data pertaining to only 52 subjects. For video based recognition, the database with the highest number of videos, BIWI [14], contains data pertaining to only 20 subjects which is not suitable for evaluating recognition performance. Also, there is only one existing database with Kinect version 2 data, HRRFaceD [34], which contains data for only 18 subjects. None of the existing databases address cross-distance face recognition or the unconstrained scenario. Therefore evaluating an algorithm on these databases does not address all the challenging covariates with unconstrained face recognition. Table 2.1 provides an overview of the existing RGB-D databases

The Kinect version 2 device provides improvements in acquisition technology for both color and depth. It utilizes the Time of Flight (TOF) technology to obtain a more accurate and higher resolution depth map (512×424) compared to its predecessor (320×240) while still offering low sensor cost. Along with RGB and depth, it can simultaneously capture near infrared video, enabling exploration of multi-modal techniques as well. While Kinect version 1 device can also capture infrared data, it cannot do so simultaneously while capturing RGB-D data. The Kinect version 2 device does not have this limitation. Figure 2.1 presents a comparison of the RGB-D images obtained using the first and second versions of the Kinect device. The improved acquisition technology can bolster the performance of recognition algorithms even further. However, the first requirement to conduct research and evaluation on such algorithms is a large and challenging Kinect 2 database. A challenging benchmark RGB-DI database with large number of subjects, samples, and challenges is essential in order to develop better algorithms and further the state-of-the-art in unconstrained face recognition. In this paper, we present the KaspAROV [27] RGB-DI database, which contains videos pertaining to 108 individuals captured using both versions of the Kinect sensor. Since the Kinect version 2 device captures both RGB-D and near infrared data, we term the data captured as RGB-DI (denoting

Table 2.1: Summarizing the characteristics of existing RGB-D face databases.

Device	**Database**	**No. of subjects**	**No. of samples**	**Total samples**	**Covariates**
Kinect 1	BIWI [14]	20	24 (video)	480 (video)	Pose and expression
	CurtinFaces [31]	52	97	4,784	Pose, expression, illumination
	Eurecom [36]	52	18	936	Expression, illumination, occlusion
	FaceWarehouse [7]	150	20	3000	Pose and expression
	Florence [4]	50	1 (video)	50 (video)	Pose
	IIIT-D [20]	106	11-254	4,605	Pose and expression
	VAP [23]	31	51	1581	Pose and expression
Kinect 2	HRRFaceD [34]	18	200	3,600	Pose
Kinect 1 and 2	KaspAROV [27] (Proposed)	108	4 (video)	432 (video) 117,831 (frames)	Pose, illumination, expression, sensor interoperability, resolution, distance

Figure 2.1: Sample RGB-D frames. The first two columns contain frames captured using Kinect v1 device (from left to right: visible and depth) and the last three columns contain frames captured using the Kinect v2 device (from left to right: visible, depth, and NIR).

color, depth, and near infrared). The proposed database provides two videos per sensor for each subject, captured in two different sessions and under unconstrained pose, expression, and illumination conditions. The database also encompasses the problem of cross-distance face recognition. We provide face identification and verification protocols to encourage and facilitate comparative evaluation as well as a set of baseline results obtained using existing algorithms. The remainder of this paper is organized as follows: Section 2 describes the proposed database, Section 3 presents the experimental protocol and results, and Section 4 presents concluding remarks and future research directions.

Figure 2.2: Sample frames corresponding to two subjects from the KaspAROV [27] RGB-D database captured using the Kinect v2 device. The first two columns are visible spectrum images, the next two columns correspond to depth images, and the final two columns are NIR images.

2.2 The KaspAROV Database

The KaspAROV [27] RGB-DI database contains videos pertaining to 108 subjects captured using Kinect version 1 and version 2 devices. Consent for capturing these videos is obtained from all the participants. There are two videos for each subject for each sensor, resulting in a total of four videos per subject. These videos are captured in two different sessions. For any given video, both the Kinect sensors are placed at the same viewpoint within a location which is not uniformly illuminated. Two subjects are then asked to walk back and forth within the field-of-view (FOV) of the sensors while not imposing any limitation on expression, pose, or gesture. Therefore, the database contains unconstrained pose, illumination, and expression variations along with variations in capture distance. Further, each video contains a full-body capture of at least two subjects of interest (for whom labeled data is available in the database) while sometimes containing additional individuals in the background (without labeled data) which mimics a surveillance scenario. Further details of the database are described below.

2.2.1 Data Acquisition

A total of 432 videos are captured, 216 of which are captured using a Kinect v1 device and the other 216 are captured using a Kinect v2 device from 108 subjects. The color, depth, and near infrared (in case of Kinect v2) streams are split into frames using the Kinect SDK v1.8 for the Kinect v1 streams and Kinect SDK v2.0 for the Kinect v2 streams. A total of 67,984 frames are extracted from all the Kinect v1 videos, whereas, a total of 70,518 frames are extracted from all the Kinect v2 videos. The native resolution of color frames for Kinect v1 videos is 640×480 and for the Kinect v2 videos is 1920×1080. The native resolution of depth frames for Kinect v1 is 320×240 and 512×424 for Kinect v2. The native resolution of the near infrared data captured by Kinect v2 device is also 512×424. Figure 2.2 shows sample color images, depth images, and NIR images from the database.

No. of individuals	108
No. of videos - Kinect V1	216
No. of videos - Kinect V2	216
No. of frames - Kinect V1	67,984
No. of frames - Kinect V2	70,518

Table 2.2: Characteristics of the KaspAROV [27] face database.

2.2.2 Face Detection

The first step in a face recognition pipeline is face detection. Given the challenging nature of the database, we asserted that every step of the face recognition pipeline will be challenging. Therefore, we manually detected all the frames present in all the videos captured using Kinect v1 and Kinect v2 devices. From Kinect v1, manual annotation provides 55,712 faces and from Kinect v2, 62,119 faces faces are detected. A bounding box is prepared around all the faces. These manual annotations can serve as the ground truth to evaluate face detection performance in RGB-DI scenario.

We then perform automatic face detection to understand the detection performance on these videos. There are multiple face detectors available in literature. In our evaluation, we used five different face detectors including the Kinect's face detector. First, we used the face API of

Kinect SDK v2.0 to detect faces in the frames captured from Kinect v2 device. It is important to note that the API only works on Kinect v2 videos. The Everingham face detector [12] is then utilized to detect faces from the frames in which no faces are detected by the API. From 70,518 frames of Kinect 2 containing 62,119 faces, the API + Everingham face detector detected 38,544 faces. Since the API does not work for Kinect v1 videos, Everingham face detector forms the first stage of face detection. We obtained a total of 22,217 faces out of 55,712 frames. We observed that a lot of faces are not detected even after using the two-level detection in case of Kinect v2. These are primarily faces with non-frontal pose and captured at relatively large distance from the sensor. To detect such faces, we further employed Histogram of Oriented Gradients (HOG) based human body detection [8] on the frames where the previous two detectors could not detect faces.

The HOG detector detects the human boundary present in the frames which is then refined using the depth information. The upper 30% of the segmented human body is enlarged, followed by applying Everingham detector on the specific region. Using this additional level of detection, the number of detected faces for Kinect v2 increased to 51,401. Unfortunately, the depth data obtained using Kinect v1 is not suitable for applying this procedure. Therefore, this approach could not be applied on Kinect v1 videos. A few examples of faces that are detected manually and not by the automated detection algorithms discussed above are presented in Figure 2.3. Analyzing these images reveal that poor quality, distance from the camera, and pose are the three biggest challenges for designing an efficient face detection algorithm that yields good results in surveillance scenarios like the one captured in the KaspAROV [27] database. It can also be observed that the Kinect v2 device provides higher fidelity face images in both color and depth modes, however, both the devices suffer from holes and spikes in the depth map which should be addressed during preprocessing.

2.2.3 Face Metadata

Along with face detection, Kinect SDK v2.0 face API also provides facial landmarks and attributes corresponding to RGB images. These additional information can be incorporated with

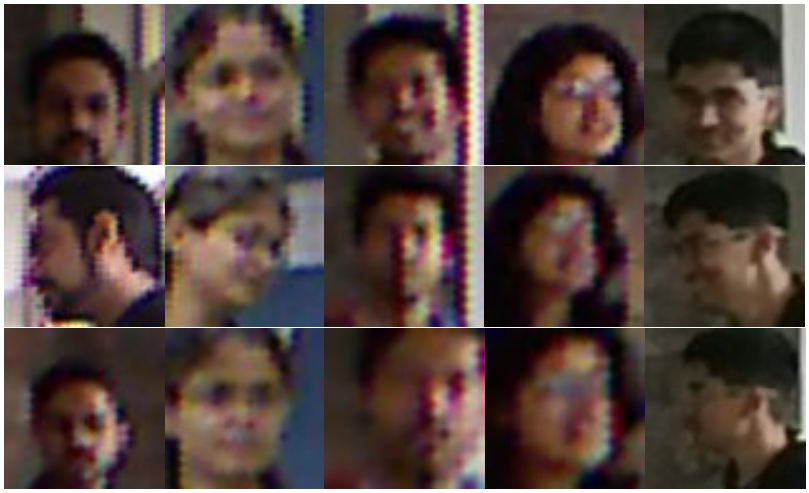


Figure 2.3: Sample faces from the KaspAROV [27] database that are not detected successfully by automatic detection.

the recognition pipeline to improve the performance. We extracted five facial landmark points, left eye, right eye, nose tip, and mouth corners, for all the faces that are detected through the Kinect SDK v2.0 face API and Everingham's face detector [12] corresponding to the videos captured from the Kinect v2 device. Head poses (yaw, pitch and roll angles) and expressions are also recorded for a total of 21,043 extracted faces. The pose distribution of these faces is presented in Figure 2.4. Among the faces for which pose data is available from the Kinect SDK, the majority are semi-profile, a small number are completely profile, and almost 10% of these faces are frontal. However, it should be noted that since facial landmarks in profile faces are much harder to label than semi-profile faces for the API to label, the actual number of profile faces in the complete database may be much higher. Finer divisions of pose can be obtained by quantization of the yaw, pitch, and roll values and used for training and evaluation of pose estimation algorithms. The pose distribution also showcases the challenging nature of the proposed database as the API is only able to provide pose labels for approximately 34% of the faces. We also observe that the labels are obtained successfully when the faces are relatively close to the sensor. As the distance of the subjects increases the API fails to obtain pose/expression labels. This further highlights the challenge of distance.

The database also consists of attribute labels that are available for a subset of images. Table 2.3 provides a summary of these attributes in the database. These attributes and their labels

Attributes	Results			
	No	Yes	Maybe	Total
Happy	11,852	2,419	2,107	16,378
Engaged	11,851	5,616	2,149	19,616
WearingGlasses	10,328	5,064	986	16,378
LeftEyeClosed	10,311	5,274	793	16,378
RightEyeClosed	10,537	3,882	1,959	16,378
MouthOpen	10,821	2,853	2,704	16,378
MouthMoved	5,597	10,180	601	16,378
LookingAway	12,098	4,688	2,830	19,616

Table 2.3: Attribute labels available for a subset of the database.

are also obtained using the Kinect v2 API. Most of the attributes are available for a subset of 16,378 images and two attributes are available for 19,616 images. Identifying attributes such as LookingAway and RightEyeClosed/LeftEyeClosed are useful for video recognition applications since such frames can be discarded and considered as failure to process frames. On the other hand, the research and development of auto-capture and auto-group image processing applications can benefit from the accurate identification of attributes such as happy and engaged.

2.2.4 Potential Usage of KaspAROV Database

Owing to the variety of information collected in the database, it can be utilized in several potential ways. We list a few of these below:

- **Face Detection:** Since the database contains unaltered frames captured using both

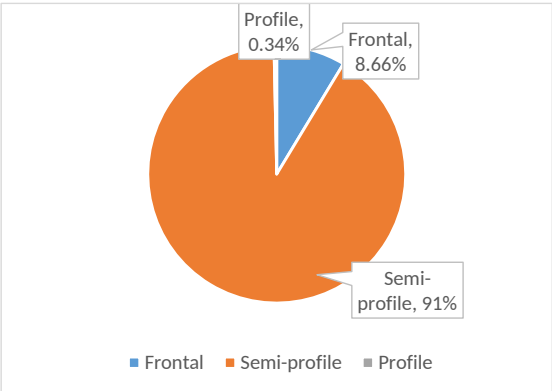


Figure 2.4: Illustrating the distribution of faces grouped by pose.

Kinect v1 and v2 sensors, it can be utilized for face detection experiments. Researchers can compare the detection performance across sensors and evaluate their detection algorithm with faces at varying resolution of both depth and color image, as well as at varying distances from the sensor. As mentioned earlier, using the Everingham face detection algorithm [12], HOG based body detection (for Kinect v2), and the face API from Kinect SDKs, detects only 73,618 faces out of 117,831 (obtained using manual annotation) faces from both Kinect 1 and Kinect 2 videos. This represents a detection rate of only 62.47%, thus indicating a large scope of research and improvement in face detection in unconstrained conditions. It is interesting to note that since the database contains three information sources: RGB image, depth information, and NIR image, the detection algorithms can also be designed for all three modalities of information.

- **Pose Recognition:** The database provides 21,043 faces with pose parameter (yaw, pitch, and roll) information as described in Section 2.2. This subset of the database can be utilized to design and evaluate pose recognition algorithms on challenging cross-distance RGB-DI imagery. Moreover, the complete database contains 41,076 face images which the Kinect v2 API failed to label automatically. These images, if manually labeled for pose, can be utilized for creating an even more challenging RGB-DI pose benchmark.

- **Video Frame Selection:** Due to the large size of videos, researchers generally have used different techniques to reduce the number of frames, for instance, sampling and frame selection. Recently, frame selection in video based face recognition for 2D videos has demonstrated significant improvement in performance [16]. Frame selection can also utilize cues from the depth map and may offer further improvements and robustness to video based recognition algorithms. Since the number of frames per video ranges between 150-800, these ideas and algorithms can be evaluated on the KaspAROV database [27].

- **Face Recognition:** The database contains individuals in multiple frames of a video and hence it can be used for both identification (one-to-many) and verification (one-to-one) experiments. Given the unconstrained nature of the database, both are important and challenging problems.

- **Face Reidentification:** Since the database contains an individual in multiple frames, it can be used to re-identify an individual in video frames. In surveillance scenarios, reidentification is an important problem that requires attention from the research community.

- **Gait Recognition:** Since the videos are captured by asking the subjects to walk, the database captures the subject's gait and it can be utilized for both gait analysis and recognition. As a combined biometric, gait can be combined with face to further improve the recognition performance. This combination is also very useful for surveillance applications since gait is an important cue in such scenarios.

- **Cross-sensor and Cross-resolution:** Since the database contains images from two RGB-D sensors, the possible combinations encompass same-sensor and cross-sensor scenarios. If data from only one version of Kinect (v1/v2) is utilized, it represents the same-sensor scenario. On the other hand, if Kinect v2 samples are utilized as gallery and only Kinect v1 samples are included as probes this is a cross-sensor cross-resolution experiment since the native resolutions of the two devices are different. If samples from both the sensors are included as probes, it represents a hybrid scenario with both same-sensor and cross-sensor probes. The cross-sensor problem encompasses the issue of sensor interoperability and can explore whether algorithms designed using Kinect v1 databases perform similarly on Kinect v2 data and vice versa. Cross-resolution is captured in two ways: (a) Kinect v1 and v2 have different native resolutions and (b) the resolution of faces captured in different frames of the same video vary based on subject distance from the camera. Cross-resolution is an important problem to address, especially in the case of a surveillance scenario where the resolution of obtained face might vary greatly due to positioning of cameras at different locations. The proposed database contains data pertaining to each of these cases and can be utilized to evaluate algorithm performance under these challenging constraints.

It is interesting to note that for each of the problems discussed above, the database can be used in RGB, RGB-D, or RGB-DI mode for Kinect v2 data and RGB or RGB-D mode for Kinect v1.

2.3 Experimental Protocol

Out of the potential applications listed in the previous section, in this research, we perform the benchmarking for image and video verification and identification. We first prepared four different protocols related to face verification and identification in images and videos followed by evaluating/benchmarking the performance of several face recognition algorithms on the database. The subsections below present the protocols and the corresponding baseline recognition results. The number of gallery and probe images present under each protocol for training, testing, and validation are summarized in Table 2.4. In order to compute these results, both Kinect v1 and Kinect v2 videos are converted to sets of images by sampling a subset of frames. Even though Kinect v2 provides a higher quality depth map, it is still prone to noise in the form of holes and spikes. A median filter of size 5×5 is utilized to denoise the depth map. Face detection is performed using Everingham's face detector [12] on the color image and a mapping between the face rectangle in the color image and the depth map is obtained. This mapping is then utilized to crop the face image in both color and depth images.

The benchmark results on the protocols are computed using the following descriptors.

- Local Binary Pattern (LBP) [2]
- Three patch local binary pattern (TPLBP) [45]
- Histogram of oriented gradients (HOG) [8]

These descriptors have been applied successfully in existing RGB-D algorithms and are well-established in their capability of encoding RGB-D face data efficiently. In addition to these feature descriptors, we also utilize the below mentioned algorithms for comparing face recognition performance on the KaspAROV [27] dataset.

- FaceVACS [1]
- RISE [19]
- mRISE

FaceVACS is used as Commercial-off-the-Shelf (COTS) face recognition software for RGB based matching and the RISE [19] algorithm is used for RGB-D matching. mRISE is proposed as a simple extension to the RISE algorithm [20] for RGB-DI data matching. The RISE algorithm combines four HOG feature vectors, two each extracted from the entropy maps of the RGB and depth components of an RGB-D image. mRISE includes two additional feature vectors extracted from the NIR component of the RGB-DI image. All other parameters and components of the algorithm are kept unchanged. The feature vectors of two RGB-D images are matched using the cosine, Euclidean, and χ^2 distance metrics.

We have also used two algorithms for enhancing the quality of depth images captured by Kinect v2 sensor.

- Markov Random Fields (MRF) [9]
- Layered Bilateral Filter (LBF) [47]

Markov Random Fields and Layered Bilateral filter both make use of registered high quality RGB images to improve the quality of corresponding depth images. Since the registration between RGB and Depth images in Kinect v1 is not optimal, we have applied depth enhancement techniques only on Kinect v2 data. Given a coarse depth map and RGB image, Markov Random Fields (MRF) assigns depth labels to all pixels which is the most likely estimate of the ground truth depth value for a given point. Layered Bilateral filtering uses edges obtained from the RGB image for improving depth image. In this, bilateral filtering is applied iteratively on the input depth image followed by a sub-pixel refinement stage to obtain the refined depth map.

The protocols include image verification and video verification. Since all the matchers used for benchmarking operate on images, the video based protocols essentially run on images and the scores of each frame are then combined using fusion rules. In order to combine the scores obtained by using multiple frames of the same video, we have used the below mentioned statistical rules.

- Average
- Min-rule

Table 2.4: Overview of the experimental protocols defined on the KaspAROV [27] database. 1. Single gallery identification, 2. Video based identification. K1-K1: only Kinect v1 data, K2-K2: only Kinect v2 data, K2-K1: Kinect v2 data as gallery and only Kinect v1 data as probe.

Protocol	Set	Subjects	No. of samples (Images/frames/videos)			
			Gallery		Probe	
			K2-K2	K1-K1	K2-K2	K1-K1
1.	Training	54	54	54	31,943	27,920
	Testing	54	54	54	30,122	27,273
2.	Training	54	16,121	14,100	15,822	13,820
	Testing	54	15,929	14,049	14,247	13,278

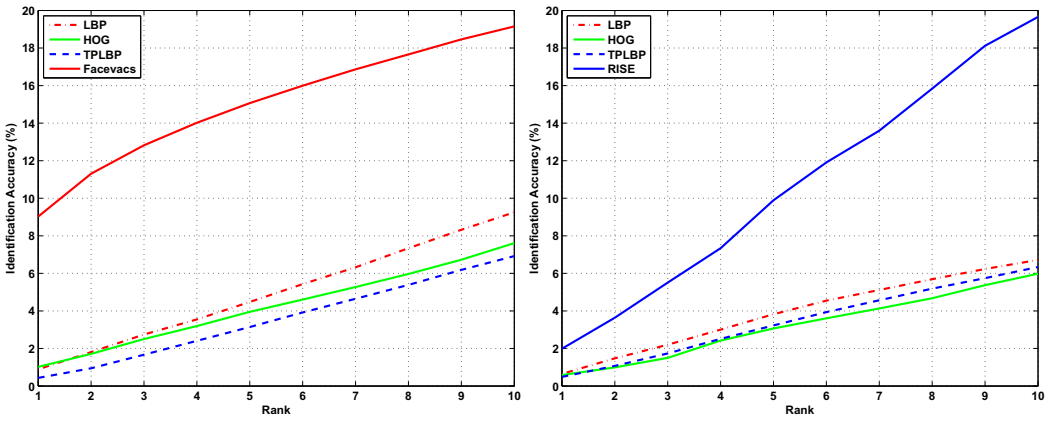
Aggregated scores obtained from all possible pairs of frames are used to decide the final subject label for each probe video. We next describe the four protocols proposed and the results obtained on the KaspAROV [27] database are given in . A summary of the protocols is also given in Table 2.4.

2.4 Benchmark Results on KaspAROV dataset

2.4.1 Single Gallery Identification

The aim of this protocol is to evaluate the performance in a $1 : N$ matching scenario, i.e., identification, while using only one gallery image per subject. Out of the total 108 subjects, frames pertaining to 50% of the subjects are utilized for training and the remaining 50% subjects are utilized for testing. The subjects in each partition are selected randomly and the process is repeated five times for cross validation. An exemplar frame is selected as the gallery image for each subject and the remaining images of the subject are utilized as probe images. The training partition is used to learn the model and the performance on validation set is used for parameter estimation. The algorithm is then evaluated on the test partition to report the final results. For each fold, each set is mutually exclusive to the others so as to prevent bias in the reported results by considering only those subjects that are unseen during training and validation. Figure 2.5 and 2.6 and Tables 2.5, 2.6 and 2.7 present the benchmark results obtained on this protocol.

With RGB data, both Kinect v1 and v2 exhibit similar performance on single gallery identification experiments on all measures except in case of FaceVACS. In FaceVACS, Kinect v2 performs

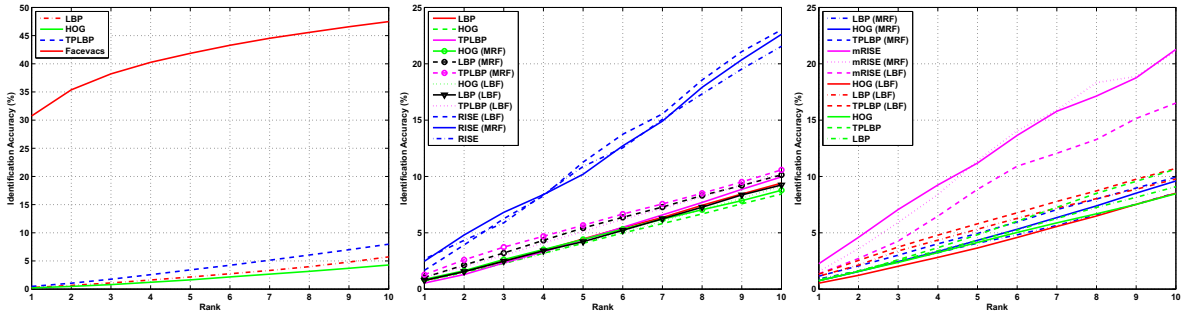


(a) RGB (b) RGB-D

Figure 2.5: CMC curves for single gallery image identification experiments using standard face recognition algorithms on Kinect v1 data.

much better than Kinect v1. Kinect v1 suffers due to it lower quality RGB camera leading to poor quality images especially in low illumination scenes. Also a large number of images from Kinect v1 fail to enroll in FaceVACS system due to poor quality. Also, effect of an improved depth sensor in Kinect v2 is made evident by better identification results on RGB-D data, which improve further by using depth enhancement techniques on Kinect v2 depth images. Identification accuracies on Kinect v2 data increase even more when RGB-DI data is used, although absolute identification performance at rank 1 remains very poor with the best case performance of approximately 1%. We also observe that using RGB-D data improves identification over just utilizing visible spectrum data.

Furthermore, the performance of Kinect v1 and Kinect v2 in face recognition is further evaluated using an existing RGB-D algorithm [20] for the case of single gallery image identification experiment. Table 2.8 presents the statistics of the number of probes that were successfully identified using data from both, none, or individual sensor, whereas Figure 2.9 presents a few examples of probes belonging to each category. The categories included in Table 2.8 are treated as mutually exclusive and exhaustive for calculating the number of probes, i.e., the probes belonging to 'Either Kinect v1 or v2' category are not counted for the individual sensor categories.



(a) RGB (b) RGB-D (c) RGB-DI

Figure 2.6: CMC curves for single gallery image identification experiments sing standard face recognition algorithms and depth enhancement techniques on Kinect v2 data.

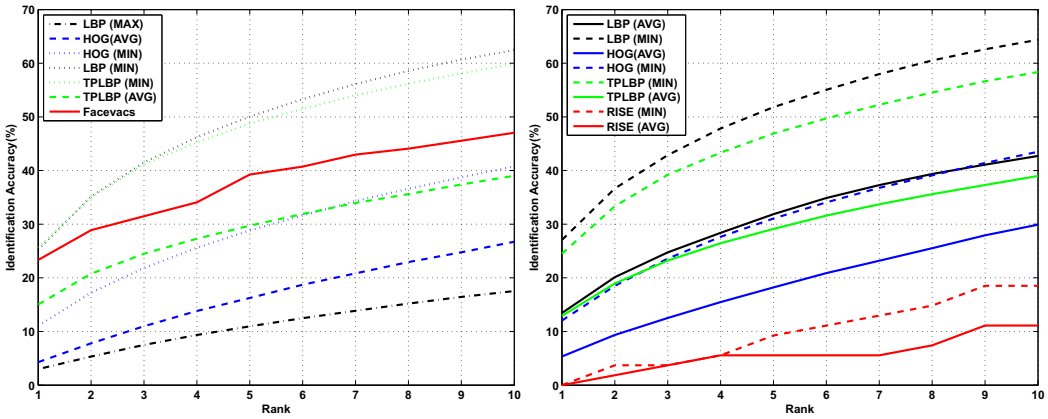
2.4.2 Video Based Identification

The aim of this protocol is to evaluate the performance in identification scenario, using one entire video as the gallery per subject. The data into 50% training and 50% testing is utilized. Each set is mutually exclusive to the others. However, instead of specifying a single frame as gallery for each subject, one of the videos is provided as gallery for each subject. The remaining videos are then used as probe, with each video comprising a single probe. Similar to the previous protocol, this partition is performed with five times random cross validation. Benchmark results on this protocol are presented in Figure 2.7 and Figure 2.8 and Tables 2.5, 2.6 and 2.7.

The min rule seems to work best for video identification, outperforming mean and max rules consistently when all the frames are utilized. The performance is observed to be the best for Kinect v2 data and worst for the cross-sensor scenario. For the same-sensor experiments, the best case performance is obtained using only RGB data with FaceVACS However with other face recognition algorithms, the performance trend remains the same, RGB $<$ RGB-D $<$ RGB-DI Overall, the LBP descriptor seems to work best on all of the data as compared to other features. In terms of depth enhancement techniques, MRF generally provides better results.

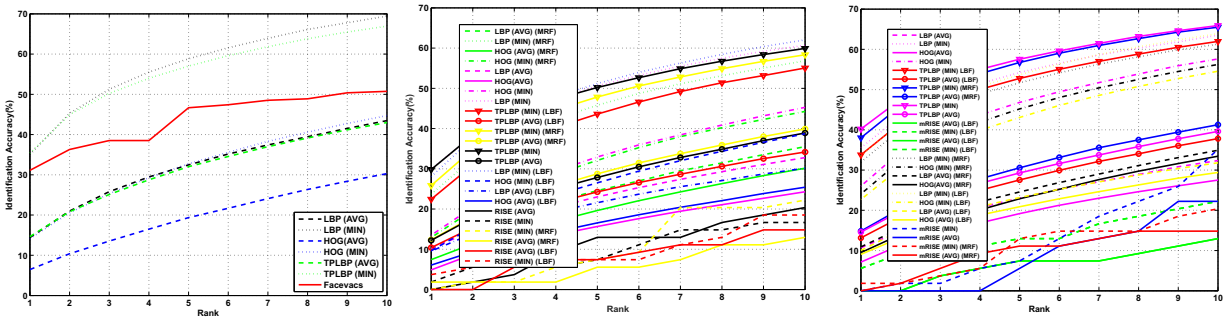
2.4.3 Observations Across Multiple Experiments

It is observed that Kinect v2 is successful for a higher number of probes, as compared to Kinect v1. Still there were cases where Kinect v1 provided better results compared to Kinect v2. This



(a) RGB (b) RGB-D

Figure 2.7: Baseline CMC curves for video based identification experiments on the proposed KaspAROV [27] database using different algorithms and Kinect v1 data.



(a) RGB (b) RGB-D (c) RGB-DI

Figure 2.8: Baseline CMC curves for video based identification experiments on the proposed KaspAROV [27] database using different algorithms and Kinect v2 data.

Table 2.5: Baseline identification results (Rank 10) on the proposed KaspAROV [27] database using only RGB data.

Protocol	Method	Fusion Rule	Feature			
			LBP	TPLBP	HOG	FaceVACS
1. Single gallery	K1-K1	-	9.93 ± 3.04	7.78 ± 1.48	8.83 ± 2.45	19.14
	K2-K2	-	7.61 ± 2.35	9.26 ± 2.83	7.80 ± 3.01	47.48
2. Multiple gallery	K1-K1	Mean	41.19 ± 4.66	39.02 ± 3.96	26.73 ± 4.66	47.03
		Min	62.45 ± 4.66	59.88 ± 3.96	40.77 ± 4.66	
	K2-K2	Mean	43.46 ± 3.94	42.98 ± 4.51	30.29 ± 3.94	50.74
		Min	69.44 ± 3.94	66.95 ± 4.51	44.71 ± 3.94	

Table 2.6: Baseline identification results (Rank 10) on the proposed KaspAROV [27] database using only RGB-D data.

Protocol	Method	Depth Enhancement	Fusion Rule	Feature			
				LBP	TPLBP	HOG	RISE
1. Single gallery	K1-K1	None	-	7.96 ± 1.83	7.12 ± 1.70	7.32 ± 1.69	19.65
	K2-K2	None	-	9.40 ± 1.85	9.95 ± 1.97	8.43 ± 2.58	21.55
		LBF	-	9.23 ± 1.85	10.23 ± 2.58	9.13 ± 1.97	23.01
		MRF	-	10.14 ± 1.85	10.58 ± 2.58	8.76 ± 1.97	22.6
2. Multiple gallery	K1-K1	None	Mean	32.78 ± 2.43	38.90 ± 3.40	28.28 ± 2.43	47.03
		None	Min	60.70 ± 2.43	59.92 ± 3.40	45.00 ± 2.43	11.11
	K2-K2	None	Mean	34.66 ± 2.44	37.94 ± 2.62	23.63 ± 2.44	20.37
			Min	49.59 ± 2.44	47.16 ± 2.62	26.82 ± 2.44	16.66
		LBF	Mean	30.16 ± 2.65	34.17 ± 2.89	25.42 ± 2.62	14.81
			Min	62.03 ± 1.72	55.05 ± 2.54	38.71 ± 2.06	18.51
		MRF	Mean	35.50 ± 2.97	39.89 ± 3.27	30.15 ± 1.74	12.96
			Min	56.83 ± 1.92	58.33 ± 2.30	44.26 ± 1.65	22.22

Table 2.7: Baseline identification results (Rank 10) on the proposed KaspAROV [27] database using only RGB-DI data.

Protocol	Method	Depth Enhancement	Fusion Rule	Feature			
				LBP	TPLBP	HOG	mRISE
1. Single gallery	K2-K2	None	-	9.08 ± 1.32	10.65 ± 1.92	8.46 ± 1.81	21.27
		LBF	-	9.77 ± 1.32	10.71 ± 1.81	8.51 ± 1.92	16.51
		MRF	-	8.53 ± 1.32	9.93 ± 1.81	9.59 ± 1.92	21.07
2. Multiple gallery	K2-K2	None	Mean	32.36 ± 2.46	39.60 ± 3.47	34.59 ± 2.46	22.22
			Min	63.51 ± 2.46	65.91 ± 3.47	58.85 ± 2.46	35.18
		LBF	Mean	31.80 ± 3.25	37.81 ± 3.38	29.40 ± 1.46	12.96
			Min	63.90 ± 1.57	61.97 ± 2.03	54.55 ± 1.49	22.22
		MRF	Mean	34.91 ± 3.04	41.26 ± 3.46	33.43 ± 1.65	14.81
			Min	61.26 ± 1.43	65.49 ± 1.83	56.24 ± 1.33	20.37

(a) Either Kinect v1 or v2

(b) Kinect v2 only

(c) Kinect v1 only

(d) Neither Kinect v1 nor v2

Figure 2.9: Examples of probe images belonging to each of the four cases of sensor choice. The captions indicate the sensor choice that would lead to correct identification for the probes as observed in our experiments, i.e., (a) denotes the probes for which data from either sensor suffices, (b) denotes probes where only Kinect v2 data leads to the correct output and so on.

Table 2.8: The no. of probes successfully identified in single gallery image identification experiment.

Data Source	No. of Probes successfully identified
Either Kinect v1 or v2	75
Neither Kinect v1 nor v2	3042
Only Kinect v2	1215
Only Kinect v1	677

demonstrates that even with better sensor quality of Kinect v2, it still doesn't completely supersede Kinect v1's recognition performance. Also, a vast majority of the probes are misclassified using data from either sensor which suggests a large scope of improvement in the applied algorithms and methodologies for RGB-D face recognition. From Figure 2.9, we can see that the image samples presented in the worst scenario (d) where neither sensor works, are the most challenging cases containing pose, expression, illumination, occlusion, missing information, and poor quality. For videos shot in the daytime where the person of interest is standing against the sunlight, the recognition performance is very poor. Furthermore, in the night time videos, there are cases when it is too dark for the camera to record anything significant in the RGB spectrum. Therefore, it is our assertion that illumination in general, could be addressed to a large degree in the future by incorporating the NIR data which is simultaneously available only with Kinect v2 sensor. Addressing multiple frames from the video can be leveraged to provide pose invariance.

2.5 Summary

Researchers have explored the utility of RGB-D data in improving face recognition. However, most such effort has utilized RGB-D information obtained using the Kinect consumer level device. Recently, the Kinect version 2 device has been released which provides higher quality depth data at comparable sensor cost. In addition, it can also simultaneously capture NIR data which can further augment detection, recognition, and other tasks. In this paper, we present a large RGB-DI database with videos of 108 subjects captured using both Kinect v1 and Kinect v2 devices. The database includes standard experimental protocols and encompasses

the challenges of pose, illumination, expression, quality, cross-distance, and video based RGB-DI face detection and recognition. The experimental results indicate that the existing approaches cannot fully address these challenges and further research is required in order to effectively utilize RGB-DI devices in face recognition.

Chapter 3

Face Detection and Recognition Algorithms

In this research, we have developed face detection and recognition algorithms for RGB-D videos obtained in surveillance scenario. Proposed face detection algorithm makes use of depth images in conjunction with RGB images for segmenting humans and detecting faces in the video frames. On the other hand, the proposed face recognition algorithm learns a deep learning based shared representation of RGB and Depth images for recognizing faces in the videos.

3.1 Face Detection at a Distance in RGB-D Videos

Face detection algorithms generally provide good results for frontal face images with good illumination and close proximity to the imaging device. However, in the surveillance scenario, cameras are often placed far away from the subjects and the collected frames from surveillance video suffers from variations in pose and illumination. Developing a face detection system, robust to all these variations, is a highly challenging task. Among many approaches, depth information can be utilized for improving face detection results. This research focuses on developing a novel face detection algorithm for RGB-D videos taken in unconstrained scenarios. The proposed face detection algorithm utilizes human body detection in color images and combines it with the

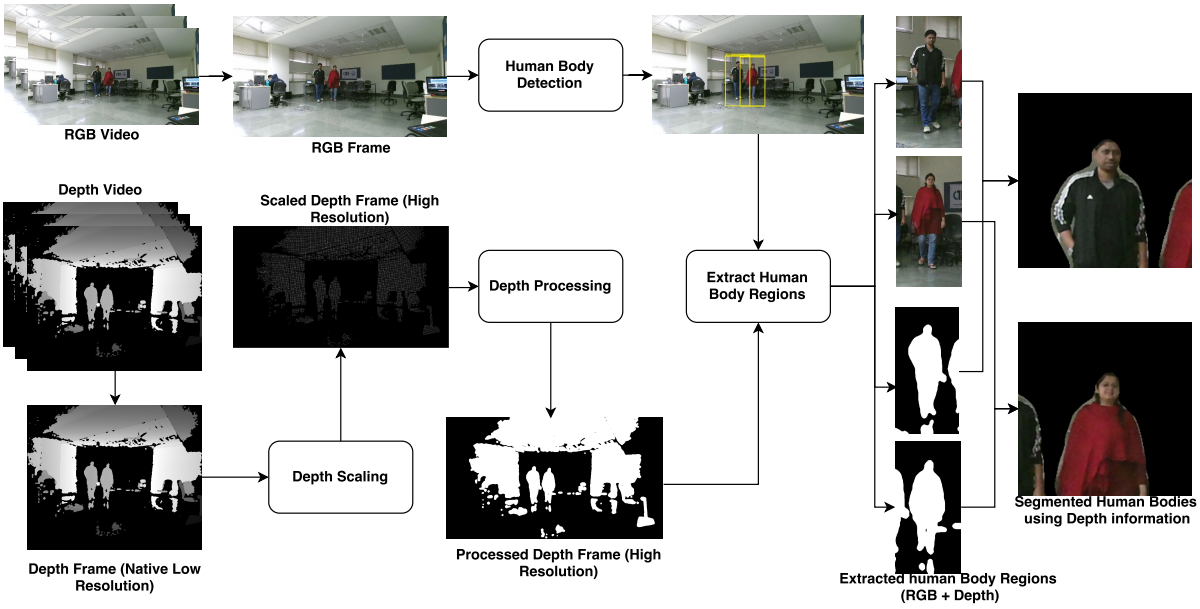


Figure 3.1: Illustrating the steps of the proposed KarPOV algorithm for human segmentation and face detection in a sample frame.

corresponding depth map to provide a robust solution for face detection at a distance in RGB-D videos.

MS Kinect also has a software development kit (SDK) [38] for face detection. Along with that, there are other face detectors in literature, such as Haar face detector [43] and Everingham detector [12], that operate on RGB images. However, on surveillance quality videos obtained from Kinect, they yield a significant number of false positives and false negatives. Figure 3.2 shows face detection results of Kinect SDK and Everingham detector. Therefore, in this research, we focus on detecting faces in video frames where the subjects are at a considerable distance from the imaging device, which is typical to surveillance scenarios. The proposed approach, KarPOV, is based on the assumption that if the video is captured from a distance, the body of the individual should also be visible along with the face. Since *body* will be larger, it will be easier to extract that and then determine the region of interest around the top portion of the detection body. The proposed approach first uses HOG based human body detector [8] for detecting humans in the video frames and then utilizes the corresponding depth information for segmenting detected human bodies from background.

Figure 3.2: Sample frames showing the comparative results of Kinect SDK + Everingham and KarPOV.

3.1.1 KarPOV - Proposed RGB-D Face Detector

For Kinect v2 sensor, the quality of depth data deteriorates with increase in distance between the subject and the sensor. Therefore, human body detection is performed primarily on the RGB frame and corresponding depth data is then used to aid segmentation. The step involved in the KarPOV face detector are demonstrated in Figure 3.1 and explained below.

1. *Human Body Detection and Tracking*: RGB frames are captured at a native resolution of 1920×1080 from Kinect v2 device. As mentioned before, major challenges for face detection are observed when there is a large stand-off distance between the subject and camera. Interestingly, in such frames, it is easier to detect human bodies due to the fact that complete bodies (from head to toe) are only captured for the subjects that are at a minimum distance from the camera. Thus, Histogram of Oriented Gradient based human body detection [8] is applied for efficiently capturing subjects of interest in the frames where the Kinect Face API fails to capture faces.

Algorithm 1 Pseudo-code for implementation of KarPOV face detector

procedure KARPOV
  for each *frame* in *Video* **do**
    $d(x,y) \leftarrow$ *Depth* frame
    $c(x,y) \leftarrow$ *Color* frame
    $mData \leftarrow$ *Color* To *Depth* coordinate mapping
    Depth Processing:
    $d_{med}(x,y) \leftarrow$ 5x5 *median filter* on d(x,y)
    Scale $d_{med}(x,y)$ to 1920 times 1080 using $mData$
    $d_{scaled}(x,y) \leftarrow Normalize(d_{norm}(x,y))$
    $d_{canny}(x,y) \leftarrow Canny(d_{norm}(x,y))$
    $d_{or}(x,y) \leftarrow Or(d_{canny}(x,y), d_{norm}(x,y))$
    $d_{close}(x,y) \leftarrow Closing(d_{or}(x,y), `disk', `3')$
    $d_{med}(x,y)' \leftarrow$ 5x5 *median filter* on $d_{close}(x,y)$
    $d_{open}(x,y) \leftarrow Opening(d'_{med}(x,y), `disk', `5')$
    $d'_{close}(x,y) \leftarrow Closing(d_{open}(x,y), `disk', `3')$
    $d_m(x,y) \leftarrow$ 15x15 *median filter* on $d'_{close}(x,y)$
    $d_{dilate}(x,y) \leftarrow Dilate(d_m(x,y), `disk', `4')$
    $d_{final}(x,y) \leftarrow Threshold(d_{dilate})$
    Human Body Detection and Tracking:
    $DetectedBodies \leftarrow DetectHumanBodies(c(x,y))$
    $bROI \leftarrow$ Detected Bodies location in c(x,y))
    if $count(DetectedBodies) = 0$ **then**
      $c(x,y)' \leftarrow PreviousFrame(c(x,y))$
      $DetectedBodies \leftarrow BodyTracking(c(x,y)')$
    Depth based Body Segmentation:
    for each *body* in *DetectedBodies* **do**
      $d_{crop}(x,y) \leftarrow Crop(d_{final}(x,y), bROI)$
      $c_{seg}(x,y) \leftarrow And(body(x,y), d_{crop}(x,y))$
      $c_{torso}(x,y) \leftarrow Crop(c_{seg}(x,y), [h/2 \quad w]))$
      $c_{faceReg}(x,y) \leftarrow Resize(c_{torso}(x,y), 3))$
      $faces \leftarrow Everingham(c_{faceReg}(x,y)))$

2. *Depth Frame Pre-Processing*: Since face recognition algorithms would like to utilize both RGB and depth information for recognition, it is important to process the depth map as well for face detection. The depth sensor on the Kinect2 samples input scene relatively sparsely as compared to the RGB camera. Hence, the native resolution of the depth sensor on Kinect v2 is lower, 512×424, than the RGB counterpart. Different native resolutions of the RGB and depth sensors require registering the RGB and depth frames. The color to depth coordinate mapping data is used to register the two information sources and generate a sparse depth image in matching resolution to RGB camera of Kinect v2.

This is then subjected to a series of canny edge detection [18], morphological and median filtering operations and is then thresholded to generate the processed binary depth frame, as detailed in Algorithm 1 and shown in Figure 3.1. The final processed depth frame is well registered to the RGB frame at its native resolution i.e. 1920×1080.

3. *Human Body Segmentation*: A pixel wise product of the detected human bodies from the RGB image and corresponding regions of the processed depth image yields the tightly segmented human bodies from the given RGB frame.

4. *Face Detection*: Upper 50 % of the segmented human bodies is then enlarged to 300% and the Everingham face detector is used to detect the faces. The detection boundary in RGB frames is mapped to depth map as well.

3.2 RGB-D Face Recognition via Learning-based Reconstruction

It is challenging to apply RGB-D based face recognition in surveillance scenarios due to the large distance of such cameras from the subject. The depth information captured in such situations is of poor quality and may not contribute to recognition. We introduce a RGB-D face recognition system which only needs RGB probe images. This is accomplished using a novel learning-based reconstruction model presented in our work. The proposed method generates a feature rich representation from RGB images which contains discriminative information from both the RGB and depth images. Thus, this representation is much more discriminative than the RGB images and gives substantially higher identification accuracy than a conventional fusion based RGB-D recognition pipeline.

Figure 3.3a show as sample of images captured at close distances, where the quality of captured face images is good. However, with large standoff distance between the camera and the subject, both RGB and depth cameras fail to capture good quality face images, as seen in Figure 3.3b.

We introduce a new neural network based algorithm to construct a feature rich representation from RGB images so that it contains discriminative information from both the RGB and depth

(a) (b)

Figure 3.3: RGB and depth images: (a) in controlled conditions (Eurecom RGBD database [36]) and (b) with large standoff distance and uncontrolled conditions (Kasparov database [27]).

images. During training, a mapping function is learnt between RGB and depth images that helps to construct feature rich representation that has the properties of both RGB and depth data, hence making it similar to a shared representation. The property of reconstructing depth images from the RGB probe images provides an added advantage that it is not necessary to capture depth information during testing. The major research contributions of our paper can be summarized as follows:

- We introduce a novel neural network architecture to learn a mapping function between two modalities M_1 and M_2. This mapping function can be used to construct a feature rich representation of both the modalities combined in one .
- The proposed architecture is applied on RGB and depth face images to construct a feature rich representation. The approach utilizes the discriminative properties of depth data without the need of capturing depth data for probe images. This approach can be deployed in scenarios where the standoff distance of the subject from the camera is too high to get good quality depth data.
- On the Kasparov database [27], the proposed algorithm provides state-of-the-art results and yields significant improvements in identification accuracy on low quality face images captured at a distance. On the IIITD RGB-D database [19] the proposed algorithm yield competitive accuracies with respect to [19].

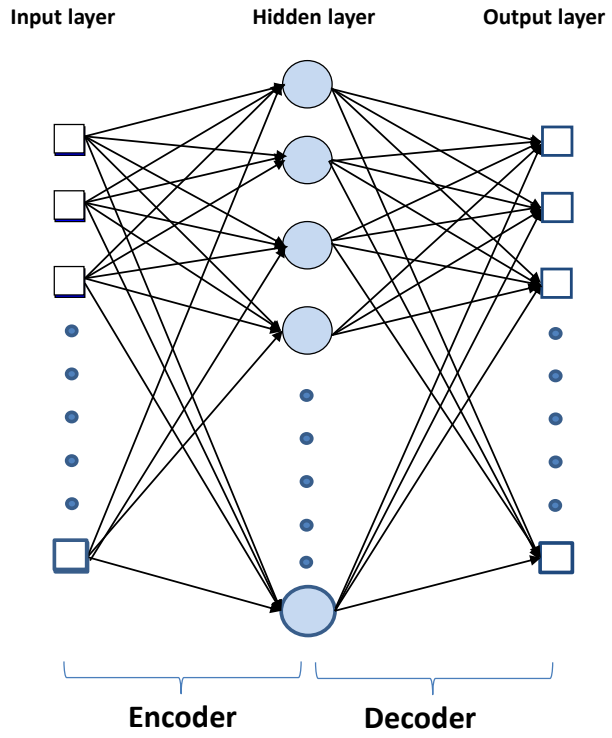


Figure 3.4: Illustration of Autoencoder [3].

3.2.1 VaNaND- Proposed Face Recognition Algorithm

This section presents the formulation of the proposed algorithm. Our proposed algorithm, named VaNaND, derives its basic motivation and idea from a deep learning based unsupervised feature extraction architecture called Stacked AutoEncoder.

3.2.1.1 SAE: Stacked AutoEncoder

Autoencoder is a deep neural network architecture geared towards unsupervised feature learning. An autoencoder, as illustrated in Figure 3.4, is comprised of two main steps: the encoding step and the decoding step. In the encoding step the data on the input layer is fed to the network in form of an input vector X, which is then encoded into a hidden representation H, as follows:

$$H = \phi\left(WX + b\right) \tag{3.1}$$

where, ϕ is the sigmoid function and W, b are the weights and bias respectively.

The hidden representation is then decoded back into the reconstructed data $\hat{X}$ in the output layer, as follows:

$$\hat{X} = \phi\left(W'H + b'\right) \tag{3.2}$$

where W', b' are the decoding weights and bias respectively.

The reconstruction error between X and $\hat{X}$ is then minimized across epochs in the autoencoder, by optimizing the below given loss function.

$$argmin_{\theta}(||X - \hat{X}||_2^2 + \lambda R) \tag{3.3}$$

where λ is the regularization parameter, R is the regularizer, and θ is the set of parameters $\{W_1, W_2, b_1, b_2\}$.

Once the loss function is minimized, the hidden layer representation is treated as a feature representation of the data at the input layer.

Our proposed network architecture, illustrated in Fig 3.5, instead of reconstructing the data at the input layer, it reconstructs the data at input layer into its another representation given at the output layer. After our network is trained, it could be used to reconstruct one representation of the input data into the desired representation at the output layer.

Our algorithm is presented as a generic model for learning the feature rich representation containing features from both the modalities namely M_1 and M_2 followed by a classifier for identification. The learning phase is composed of two main steps, learning the shared representation and learning the classifier for identification. Figures 3.5 and 3.6 illustrate the steps involved in the proposed pipeline where M_1 and M_2 are considered to be RGB and depth images respectively in this research.

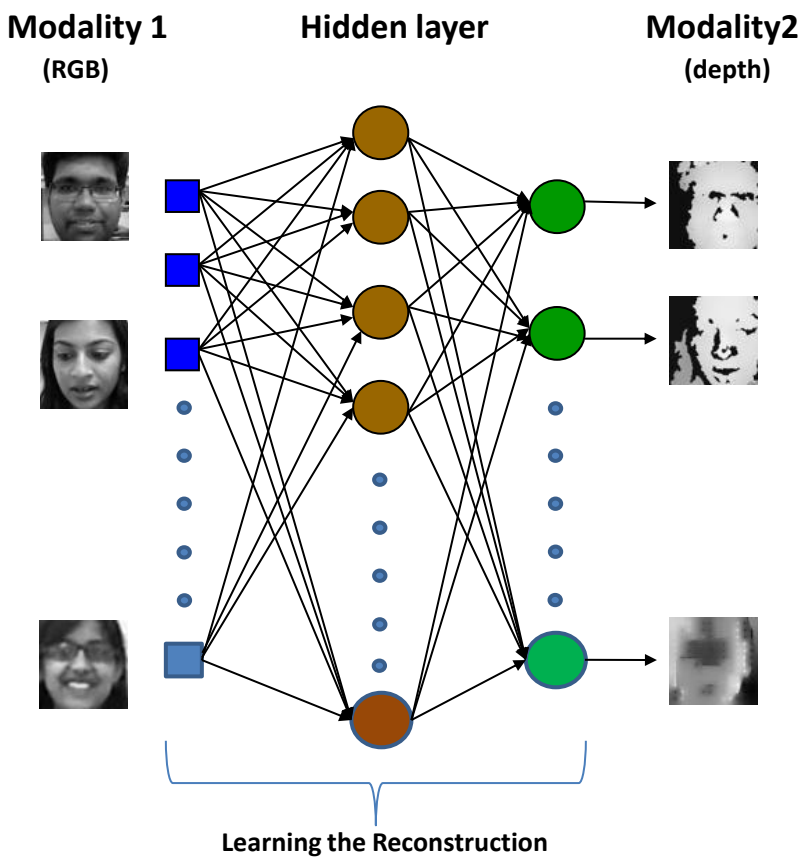


Figure 3.5: Illustrating the training module of the proposed algorithm.

3.2.1.2 Learning Mapping and Reconstruction using Neural Network

Let $X_{M_1} = \left\{x_{M_1}^{(1)}, x_{M_1}^{(2)}, x_{M_1}^{(3)}, ...x_{M_1}^{(n)}\right\}$ be the n data samples from the first modality (e.g. RGB or grayscale images). Similarly, let $X_{M_2} = \left\{x_{M_2}^{(1)}, x_{M_2}^{(2)}, x_{M_2}^{(3)}, ...x_{M_1}^{(n)}\right\}$ be the n data samples pertaining to the second modality (e.g. depth data). In this research, we propose to learn a mapping function $R : X_{M_1} \longrightarrow X_{M_2}$ using an autoencoder style neural network architecture. In the proposed approach, the first layer termed as the mapping layer can be expressed as

$$H = \phi(W_1.X_{M_1} + b_1) \tag{3.4}$$

where, ϕ is the sigmoid function and W_1, b_1 are the weights and bias respectively. In the second layer called as reconstruction layer, we learn the mapping between X_{M_1} and X_{M_2} using Equations

3.5 to 3.7.

$$\begin{aligned} \hat{X}_{M_2} &= \phi(W_2.H + b_2) \\ &= \phi(W_2.\phi(W_1.X_{M_1} + b_1) + b_2) \end{aligned} \tag{3.5}$$

such that

$$argmin_\theta(||X_{M_2} - \hat{X}_{M_2}||_2^2 + \lambda R) \tag{3.6}$$

expanding Equation 3.6 using Equation 3.5,

$$argmin_\theta(||X_{M_2} - \phi(W_2.\phi(W_1.X_{M_1} + b_1) + b_2)||_2^2 + \lambda R) \tag{3.7}$$

where λ is the regularization parameter, R is the regularizer, and θ is the set of parameters $\{W_1, W_2, b_1, b_2\}$. In this formulation, we have applied $l_2 - norm$ regularization on the weight matrix, which prevents overfitting by performing weight decay. From equations 3.5 and 3.6, it can be inferred that $\hat{X}_{M_2}$ is the reconstruction of X_{M_2}. The network for reconstruction also provides us a feature map, H, in between X_{M_1} and X_{M_2}. Thus, there are two outcomes of the proposed network,

- $\hat{X}_{M_2}$ as the feature rich representation generated by using X_{M_1} as input.
- H as a mapping function between X_{M_1} and X_{M_2}.

This mapping and reconstruction algorithm can be applied to any relevant bimodal database. In this research, we utilize the proposed algorithm to improve the performance of RGB-D face recognition.

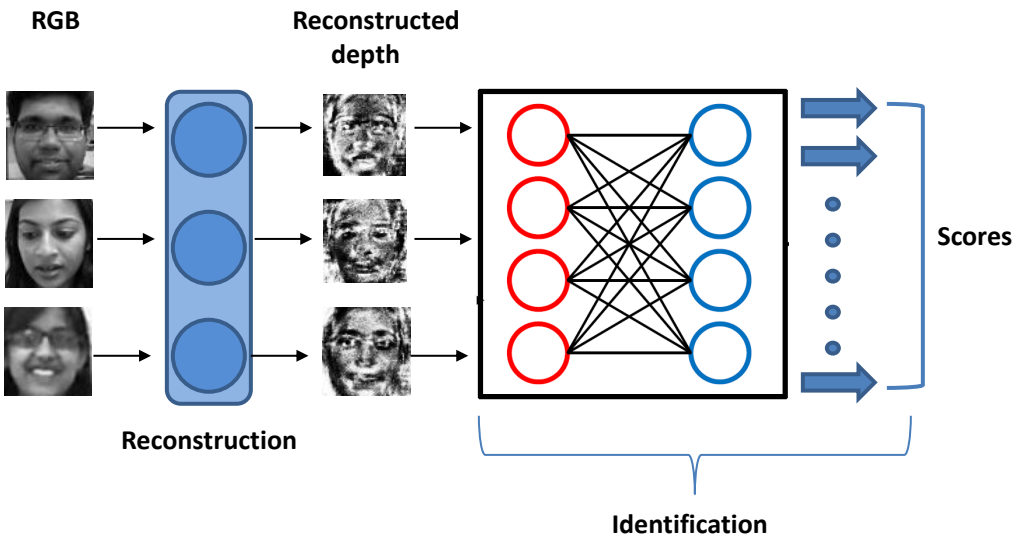


Figure 3.6: Illustrating the steps involved in testing with the proposed algorithm and identification using reconstructed data.

3.2.1.3 RGB-D Face Recognition

We next describe the RGB-D face recognition algorithm based on the proposed mapping and reconstruction algorithm described in the previous section. The proposed algorithm has two components: (1) **training**: to learn the mapping and reconstruction layers using a training set of RGB-D face images and (2) **testing**: determining the identity of the person using RGB or depth images.

With M_1 being the RGB modality (converted to grayscale) and M_2 being the depth data, we first learn the mapping between X_{RGB} and X_{depth} to obtain H and the reconstructed depth map $\hat{X}_{depth}$.

$$\hat{X}_{depth} = \phi(W_2.\phi(W_1.X_{RGB} + b_1) + b_2) \tag{3.8}$$

such that

$$argmin_\theta(||X_{depth} - \hat{X}_{depth}||_2^2 + \lambda R) \tag{3.9}$$

Figure 3.7 shows samples of feature rich representation obtained using the proposed algorithm.

(a) (b)

Figure 3.7: Visualizations of different representations used in the proposed method, (a) IIITD RGBD database [19], (b) KaspAROV database, where column 1: RGB image in grayscale, column 2: Captured depth image, column 3: Visualization of feature rich representation $\hat{V}_{shared}$.

It can be observed that compared to the original RGB image or depth map, the feature rich representation contains more discriminative information and should be able to provide better identification performance.

The next step is to use the mapping function and reconstructed output for identification. We learn a multiclass neural network classifier for face recognition. As shown in Figure 3.6, the input to the testing module is only the grayscale image. Given input RGB (grayscale) probe images the learned network is first used to generate $\hat{X}_{depth}$ using equation 3.8. This representation is then given as input to the trained neural network classifier for identification.

3.3 Summary

Detecting human faces in videos at a distance is a challenging research problem. RGB-D imaging devices can be effectively used to develop a novel solution to the problem. The algorithm proposed in this research work uses human body detection in RGB frames where face detection is challenging due to distance from camera, and combines it with depth data to effectively segment

human body from background in the frame and provide a novel face detection algorithm. Results of the proposed face detector shows considerable improvement in the number of faces detected over Kinect v2 SDK Face API and Everingham face detector. However, there still are cases where the proposed detector is unable to detect faces. We believe there is a scope for research in the domain of RGB-D face detection to further improve the results.

The proposed face recognition algorithm presents a novel and effective method of creating a feature rich representation of RGB and depth images. The presented method can be utilized in surveillance scenarios as well where obtaining depth data for the probe images is challenging. Experimental results show that this representation learned in the form of reconstructed depth images is highly discriminative and contain the properties of both depth and RGB data. Visualizations are also presented to support these intuitions. We would extend our method to investigate the effect of learning deeper reconstruction networks and test other classifiers for surveillance scenarios.

Chapter 4

Experimental Results

In this research work, we have proposed and detailed algorithms for both detecting and recognizing faces in RGB-D videos taken in surveillance like scenarios. For testing the efficiency of the proposed algorithms we have primarily used the KaspAROV [27] RGB-D video dataset. Both the proposed algorithms have been compared and contrasted with their respective state of the art counterparts in the following sections.

4.1 Face Detection Results of KarPOV Face Detector

KaspAROV RGB-D Video Dataset [27] contains video frames of 108 subjects, both male and female, captured in surveillance like scenarios. Each video contains a pair of subjects and each subject appears in two videos taken in two different sessions. The dataset also provides manually annotated face image details for each frame, hence serving as ground truth for evaluating the performance of face detection systems. We have also compared the performance of the proposed detector with existing face detection algorithms including Kinect v2 API and Everingham face detector [12].

The Face API of Kinect v2 SDK makes use of depth based human skeleton tracking for aiding face detection in RGB frames. However, since the maximum range of the depth sensor of Kinect v2 for effective operation is around 4.5 meter, the performance of Face API drops for faces that

Table 4.1: Face detection accuracy on the complete KaspAROV dataset.

Face Detection Algorithm	**Faces Detected**
Kinect SDK	54.31%
KarPOV	56.25%
Kinect SDK + Everingham	62.04%
Kinect SDK + KarPOV	**82.74%**

Table 4.2: Face detection accuracy on only *far frames* from the KaspAROV dataset.

Face Detection Algorithm	**Faces Detected**
Everingham	16.93%
KarPOV	**62.22%**

are outside the range of effective operation. Visually also we observed that the detection fails in the frames where the camera to person distance is large, Therefore, in our experiments, the frames on which Kinect v2 SDK Face API fails to detect faces are treated as *far frames* and the remaining ones are treated as *near frames*.

We have conducted experiments to first evaluate the performance on the entire dataset and also analyzed the performance specifically for the *far frames*. KaspAROV dataset contains $62,119$ manually annotated faces across 108 videos which serves as the ground truth for detection labels. For experiments on entire dataset, we have compared KarPOV with Kinect v2 SDK Face API as both the detectors use depth data for face detection. Of the 62,119 faces in the dataset, Kinect v2 SDK Face API is able to detect 33,737 faces while KarPOV yields slightly better accuracy and detected 34,942 face images. Specifically, comparing 28,383 *far frames* which are not detected by Kinect v2, we have compared the performance of KarPOV with Everingham face detector. Of these *far faces*, Everingham's approach is able to detect 4,807 faces while KarPOV shows considerably better results with 17,660 face detections. Figure 4.1 shows sample frames to illustrate the results. In the left frames, both existing algorithms and the proposed algorithm detect the faces but the existing algorithm detected a false positive. In the right frames, existing algorithms could not detect any face whereas the proposed KarPOV could detect both the faces.

We have also evaluated a combination face detectors which comprised of Kinect SDK for *near frames* and Everingham for *far frames* against a combination of Kinect SDK for *near frames* and KarPOV for *far frames*. While the former combination detected 38,544 face images, the latter

Figure 4.1: Sample frames comparing the results of Kinect SDK + Everingham and KarPOV.

detected 51,397 face images. Thus, showcasing the efficacy of the proposed algorithm in aiding face detection in surveillance scenarios. Tables 4.1 and 4.2 summarize the true positive accuracies of face detection obtained by individual detection algorithms along with the combination of algorithms for the complete database and specifically *far frames* from the database, respectively. It is to be noted that the detection results reported in Table 4.1 are with respect to the ground truth values, while the results in Table 4.2 are reported with respect to total number of *far faces*. As shown in Figure 4.2, the combination detector of Kinect SDK with KarPOV consistently performs better than the one with Everingham across all the subjects in KaspAROV dataset. While KarPOV shows good results on *far frames*, which makes it suitable for face detection in surveillance scenarios, according to ground truth, there are still $11,000$ frames that could not be detected. A sample of face detection results on KaspAROV dataset are shown in Figure 4.3. The analysis of such images shows that majority of the undetected face images suffer from heavy camera sensor noise, motion blur, and are of very low resolution, thus making them more challenging candidates for face detection.

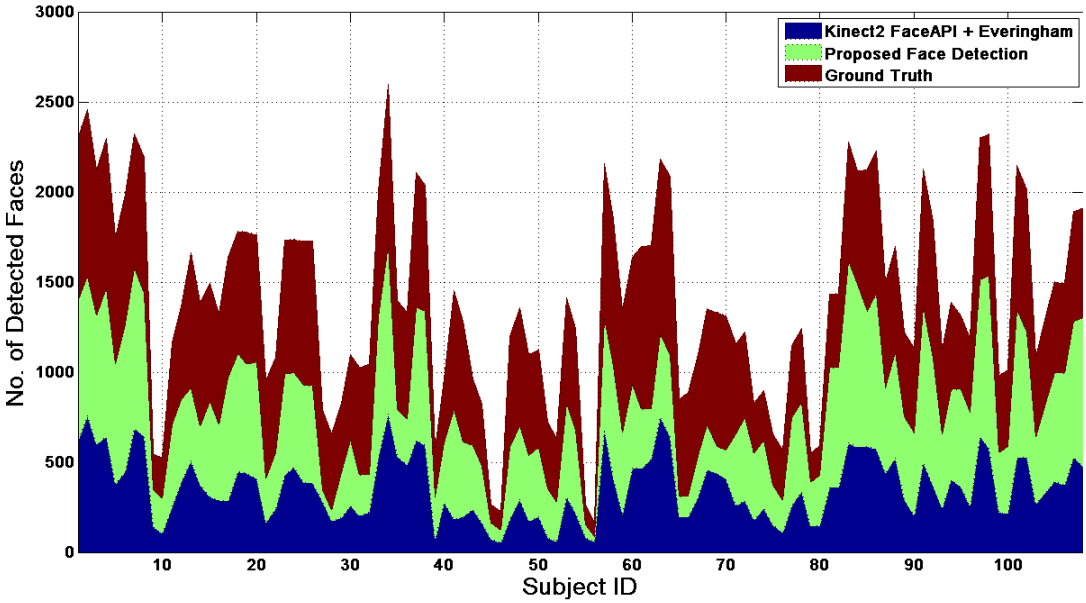


Figure 4.2: Subject wise comparison of the experimental results.

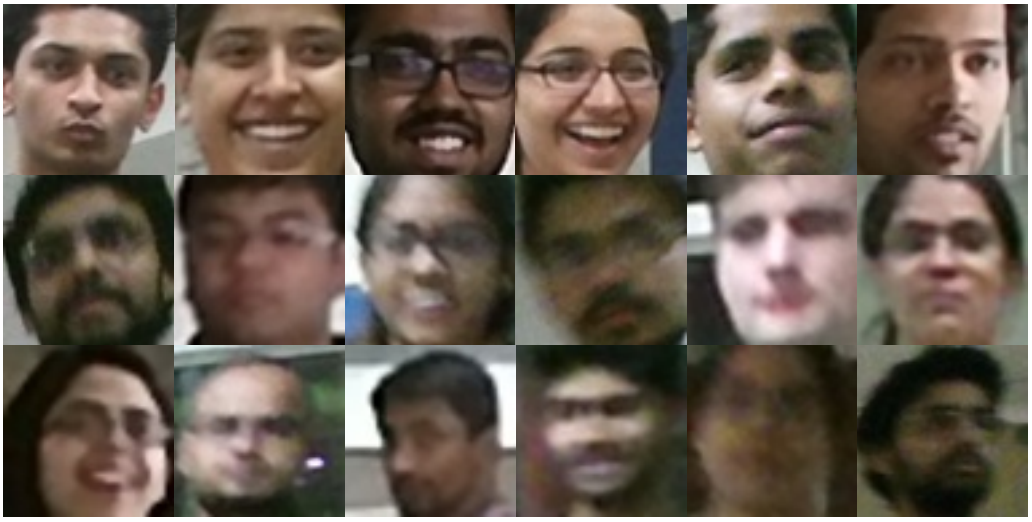


Figure 4.3: Sample face detection results from the KaspAROV RGB-D database. First row depicts faces detected by all the detectors, the next row depicts faces detected by KarPOV only, and the final row represents images not detected by any of the face detectors.

4.2 RGB-D Face Recognition Results of VaNaND Face Recognition Algorithm

For evaluating the performance of the proposed reconstruction based network, we have used two RGB-D face datasets, the IIITD RGB-D dataset [19] and the Kasparov dataset [27]. Since training the mapping function requires large training data, the shared representation model is first pretrained using the EURECOM [36] RGB-D face database.

1. **EURECOM RGB-D dataset :** The EURECOM dataset [36] contains high quality registered RGB and depth images images of 52 subjects captured using the Microsoft Kinect version 1 sensor. The dataset provides face images of each person with expressions, lighting, and occlusion. The dataset also provides 3D object files of the faces apart from RGB and depth images.

2. **KaspAROV dataset:**: KaspAROV is a RGB-D video dataset captured using both Kinect v2 and v1 sensors in surveillance like scenarios. Detected and cropped faces of 108 subjects, from the video frames under the variates of pose, illumination, distance and expression are provided in the dataset. For our experiments we have only used data from Kinect v2 sensor due to better registration of the RGB and depth images as compared to the Kinect v1 sensor data. The Kinect v2 sensor data in the KaspAROV dataset consists of $62,120$ face images. The resolution of the RGB videos is 1920×1080 and those of depth videos is 512×424.

3. **IIITD RGB-D dataset:** The IIITD RGB-D dataset contains images of 106 subjects captured using the Microsoft Kinect version 1 sensor. Each subject has multiple images,ranging between 254 to 11 images per subject per fold. The RGB and the depth images are captured as separate 24 bit images. The resolution of both RGB and Depth frames is 640×480.

Table 4.3: Details of databases used in the experiments

Dataset	Device	Classes	Image Size		Train set	Test set
			RGB	Depth		
Eurecom	Kinect 1	52	64×64	64×64	364	-
IIITD RGB-D	Kinect 1	106	64×64	64×64	9,210	13,815
KaspAROV	Kinect 2	108	64×64	64×64	31,060	31,060

4.2.1 Preprocessing

The images are converted into grayscale, followed by face detection. The detected facial regions from both grayscale and depth images are resized to a fixed resolution of 64×64 pixels. Since the IIITD RGB-D [19] and Eurecom [36] datasets contain good quality images, the cropped images (RGB and depth) provided in the database are utilized without any pre-processing. However, for the Kasparov dataset, faces are detected using Kinect Face API. The frames where faces are not detected, manual annotations given with the database are used to detect the faces. Due to high variance in distance of subjects from the camera sensor, the face images in KaspAROV dataset (both RGB and Depth) are very challenging. In order to improve the quality of depth images we have used Markov Random Field based depth enhancement technique. RGB images are used without any enhancement.

4.2.2 Protocol

Entire EURECOM database is used for pre-training the reconstruction networks. Even though the number of samples in Eurecom dataset is not large, it provides well registered RGB and depth images of good quality along with multiple variates in pose, illumination, expression.

The remaining two datasets are used for fine-tuning and testing. As shown in Table 4.3, they are divided into training and testing sets according to their pre-defined protocols. For identification experiments on the KaspAROV dataset, the pre-trained network is fine-tuned on 10% of the entire dataset and the neural network classifier is trained on 50% (which includes the data for finetuning). The remaining 50% is used for testing. For IIITD RGB-D dataset, a similar finetuning is performed, the classifier is trained on 40% of the dataset and tested on the remaining 60% of the images.

4.2.3 Experiments

To evaluate the efficacy of the proposed architecture, we have performed multiple identification experiments along with comparing the performance with state-of-the-art algorithms in literature. The experimental setup for all five experiments are described below and these are performed on both the testing databases.

1. *Recognition on raw features:* The raw depth and RGB images are used directly as features to train neural network classifiers. These are numbered as experiments 1 and 2.

2. *Recognition on hidden representation*: The learnt weights (W_1) of the proposed reconstruction network between two modalities X_{RGB} and X_{depth}, we can create a representation H as explained in Section 3.2.1.3. This is numbered as experiment 3.

3. *Recognition using VaNaND:* As explained in Section 2, the feature rich representation are obtained and used for identification. This is referred as experiment 4 and is our proposed VaNaND face recognition algorithm.

4. *Recognition using RISE [19] features*: To compare the performance with state-of-the-art algorithm, RISE [19] features are chosen. This is termed as experiment 5.

5. *Recognition using LBP [2] and Gabor [17] features*: To compare the performance with 2D features, LBP and Gabor features are chosen. These is termed as experiments 7 through 12.

4.2.4 Analysis of Results

Table 4.2.3 summarizes the rank-1 identification accuracies of all the experiments on both KaspAROV and IIITD RGB-D databases. The CMC curves for the same are outlined in Figure 4.4.

- The identification accuracies of raw RGB and depth data separately (Experiment 1 and 2) can be considered as the baselines against which we can compare all the other experiments. Depth information yields an accuracy of 11.80% and 26.81% whereas RGB input yields

Exp. No.	Modality1	Modality2	Feature	Accuracy (in %)	
				KaspAROV	IIITD RGBD
1	Depth	-	Raw	11.80	26.81
2	RGB	-	Raw	23.24	36.75
4	RGB	Depth	Hidden	**60.00**	98.08
5	RGB	Depth	VaNaND	**67.77**	**98.71**
6	RGB-D	-	RISE [19]	52.38	**98.74**
7	RGB	-	LBP [2]	1.63	9.53
8	Depth	-	LBP [2]	1.61	6.64
9	RGB-D	-	LBP [2]	1.96	8.22
10	RGB	-	Gabor [17]	2.53	36.16
11	Depth	-	Gabor [17]	1.97	32.90
12	RGB-D	-	Gabor [17]	2.03	35.03

Table 4.4: Identification results (Rank 1) on the IIITD RGB-D and Kasparov Databases

23.24% and 36.75% respectively on KaspAROV and IIITD RGB-D databases. Lower identification accuracy on depth data as compared to RGB on both the databases portrays that the discriminative information carried by depth alone is lower than that of RGB. Also the identification accuracy on KaspAROV database is lower than that on the IIITD RGB-D database. Hence, verifying the increased challenges of face recognition in surveillance scenarios.

- The hidden representation also gives competitive accuracies (experiment 3) with respect to the shared representation learnt by our proposed network for both KaspAROV and IIIT RGB-D databases. Hence, we can infer that the hidden representation H also learns discriminative information from both the modalities like the reconstructed depth, however the properties of the two are believed to be different. The visualization of the hidden layer weights W_1 of the reconstruction network is depicted in Figure 4.5.

- The feature rich representation $\hat{X}_{depth}$ obtained from the RGB to depth mapping network gives superior identification accuracy (experiment 4 in table 4.2.3) compared to raw RGB and depth as the representation(experiment 2 and 1 in table 4.2.3). We have also observed that $\hat{X}_{depth}$ gives much better results than learning features from the RGB data using a conventional deep autoencoder (feature learning on RGB data) and using the encoding weights to create a representation. The proposed method (experiment 4 in table 4.2.3)

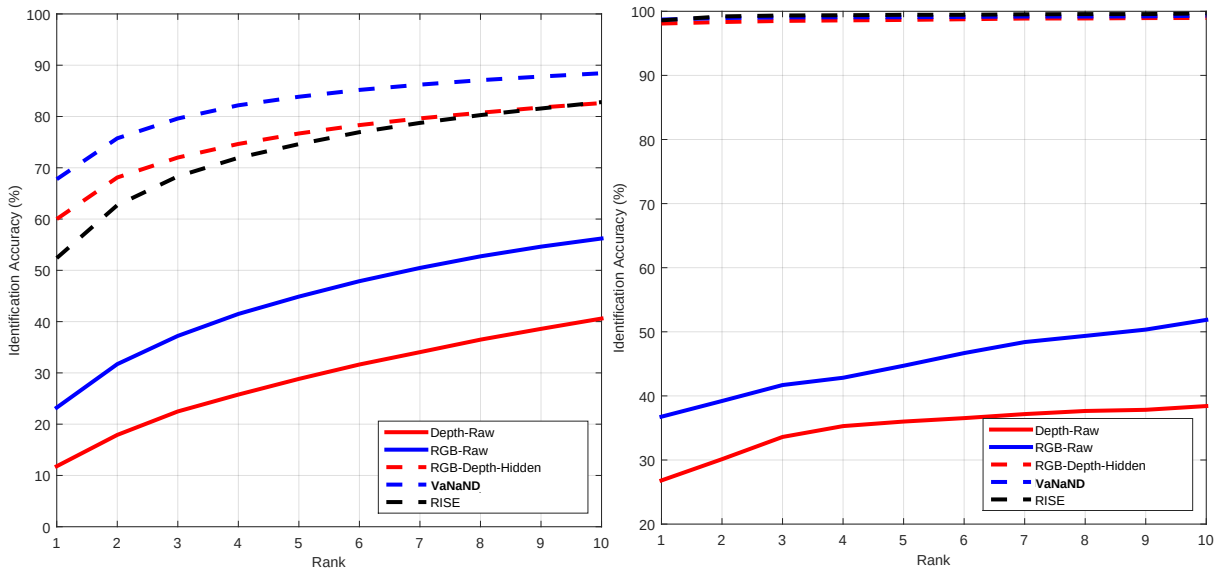


Figure 4.4: CMC curves of experiments on (a) KaspAROV Database, (b)IIITD RGBD Database

outperforms the state of art method [19] significantly on the KaspAROV dataset, where the images are of surveillance quality. This shows that the representation learned is robust to illumination, occlusion, pose and resolution variates.

- The feature rich representation $\hat{X}_{depth}$ for different samples in our experiments has two kinds of information, structural and discriminative. To visualize the fact that they look different for each subject we create a new visualization $\hat{V}_{shared}$ given by

$$\hat{V}_{shared} = \hat{X}_{depth} - mean(\hat{X}_{depth}) \tag{4.1}$$

where $mean(\hat{X}_{depth})$ is the mean feature rich representation of the entire dataset. This visualization is depicted in Fig 3.7, column 3 in both (a) and (b). It can be easily observed that they are different from each other and house important discriminative information. On closer examination of the $\hat{V}_{shared}$ images it can be observed that they contain the properties of both RGB and depth data.

- The experiments for comparison with 2D face features uses LBP [2] and Gabor [17] features. As seen in the identification results in experiments 7 to 12, both LBP and Gabor give much

Figure 4.5: Visualizations of weights (W_e) learnt by hidden layer of RGB to Depth mapping network given in Figure 3.5

lower accuracies on both the databases as compared to other features in experiments 1 to 5. Therefore, we can assert that the proposed algorithm is superior compared to these existing ones for RGB-D surveillance videos.

Chapter 5

Conclusions and Future Scope

In this research, we have addressed face detection and recognition in RGB-D video surveillance scenarios where conventional algorithms fail to scale across variantions in pose, illumination, expression and standoff distance. We have used depth data in conjunction with RGB data in our proposed face detection and recognition algorithms, which out perform current state of the art algorithms on the proposed KaspAROV dataset. Our algorithms also produce competitive results on IIITD RGB-D face dataset. In KarPOV face detection algorithm we hierarchically locate a subject's face in a given scene by first detecting and segmenting humans from the scene by using depth data of the same and then faces are detected from the segmented human bodies. In VaNaND face recognition algorithm we have put forth a novel deep learning based technique which creates a reconstruction based shared representation between two data modalities. Where the both the modalities are used while training the network but during testing phase either of the modalities suffice. Such an algorithm can find use in multi-modal learning problems where limited data is available from one of the modalities. Here, portion of data with both modalities available could be used to train the deep learning model and testing could be carried out on either of the modality from the testing dataset.

Such a deep learning architecture improves face recognition performance in domain of multi-modal recognition systems by creating a discriminative shared representation between the modalities. For instance, we can learn the network between RGB and Depth faces and use the trained

model on a RGB only dataset where the network will extract depth features from RGB images only and return it in the shared representation. This could enable improvement of face recognition performance, while working on only RGB data. For building the reconstruction based shared representation learning network we have used a stacked denoising autoencoder (SAE) as a base unit. Further research needs to be conducted in direction of:

- Implementing similar network architectures while using probabilistic graphical models such as Restricted Boltzmann Machines and Deep Belief Networks.
- Exploring the effect of shallow networks vs deeper networks in similar architectures.

Bibliography

[1] FaceVACS. http://www.cognitec.com/facevacs-videoscan.html.

[2] Ahonen, T., Hadid, A., and Pietikäinen, M. Face recognition with local binary patterns. In *European Conference on Computer Vision*. 2004, pp. 469–481.

[3] 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,.

[4] Berretti, S., Pala, P., and Del Bimbo, A. Face recognition by super-resolved 3D models from consumer depth cameras. *Transactions on Information Forensics and Security 9*, 9 (2014), 1436–1449.

[5] Bhatt, H. S., Singh, R., and Vatsa, M. Emerging covariates of face recognition.

[6] Burgin, W., Pantofaru, C., and Smart, W. D. Using depth information to improve face detection. In *Proceedings of the 6th international conference on Human-robot interaction* (2011), pp. 119–120.

[7] 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.

[8] 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.

[9] Diebel, J., and Thrun, S. An application of markov random fields to range sensing. In *Advances in Neural Information Processing Systems* (2006), pp. 291–298.

[10] Elaiwat, S., Bennamoun, M., Boussaid, F., and El-Sallam, A. A curvelet-based approach for textured 3d face recognition. *Pattern Recognition 48*, 4 (2015), 1235–1246.

[11] Erdogmus, N., and Marcel, S. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *Biometrics: Theory, Applications and Systems* (2013), pp. 1–6.

[12] 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.

[13] Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. Random forests for real time 3d face analysis. *International Journal of Computer Vision 101*, 3 (2013), 437–458.

[14] Fanelli, G., Dantone, M., Gall, J., Fossati, A., and Van Gool, L. Random forests for real time 3D face analysis. *International Journal of Computer Vision 101*, 3 (2013), 437–458.

[15] Forbes. Using kinect to patrol the dmz. `http://www.forbes.com/sites/erikkain/2014/02/03/south-korea-is-using-kinect-to-patrol-the-dmz/#7b0de395ca3d/`, 2014.

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

[17] Gabor, D. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering, 93*, 26 (1946), 429–441.

[18] Gonzalez, R., and Woods, R. *Digital image processing*. Pearson Education India, 2009.

[19] Goswami, G., Bharadwaj, S., Vatsa, M., and Singh, R. On RGB-D face recognition using kinect. In *Biometrics: Theory, Applications and Systems* (2013), pp. 1–6.

[20] Goswami, G., Vatsa, M., and Singh, R. RGB-D face recognition with texture and attribute features. *Transactions on Information Forensics and Security 9*, 10 (2014), 1629–1640.

[21] Han, J., Shao, L., Xu, D., and Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Computers 43*, 5 (2013), 1318–1334.

[22] Hg, R., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., and Tranchet, G. An RGB-D database using microsoft's kinect for windows for face detection. In *Signal Image Technology and Internet Based Systems* (2012), pp. 42–46.

[23] Hg, R. I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., and Tranchet, G. An RGB-D database using microsoft's kinect for windows for face detection. In *Signal Image Technology and Internet Based Systems* (2012), pp. 42–46.

[24] Hsu, G.-S., Liu, Y.-L., Peng, H.-C., and Wu, P.-X. RGB-D based face reconstruction and recognition. *Transactions on Information Forensics and Security 9*, 12 (2014), 2110–2118.

[25] 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 *Asian Conference on Computer Vision Workshops* (2012), pp. 133–145.

[26] Image Analysis and Biometrics Lab, IIIT Delhi. Iiitd rgbd kinect video dataset. `http://iab-rubric.org/resources/rgbd.html`, 2016.

[27] Image Analysis and Biometrics Lab, IIIT Delhi. Kasparov kinect video dataset. `http://iab-rubric.org/resources/Kasparov.html/`, 2016.

[28] Jain, A. K., and Li, S. Z. *Handbook of Face Recognition*. Springer-Verlag New York, Inc., 2005.

[29] Karpathy, A., Miller, S., and Fei-Fei, L. Object discovery in 3d scenes via shape analysis. In *IEEE IRCA* (2013), pp. 2088–2095.

[30] Kittler, J., Hilton, A., Hamouz, M., and Illingworth, J. 3d assisted face recognition: A survey of 3d imaging, modelling and recognition approachest. In *IEEE Computer Vision and Pattern Recognition Workshops* (2005), pp. 114–114.

[31] Li, B. Y. L., Mian, A. S., Liu, W., and Krishna, A. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Winter Conference on Applications of Computer Vision* (2013), pp. 186–192.

[32] Li, H., Huang, D., Morvan, J.-M., Wang, Y., and Chen, L. Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision 113*, 2 (2015), 128–142.

[33] Lin, D., Fidler, S., and Urtasun, R. Holistic scene understanding for 3d object detection with RGB-D cameras. In *IEEE International Conference on Computer Vision* (2013), pp. 1417–1424.

[34] Mantecon, T., del Bianco, C. R., Jaureguizar, F., and Garcia, N. Depth-based face recognition using local quantized patterns adapted for range data. In *International Conference on Image Processing* (2014), pp. 293–297.

[35] Mian, A. S., Bennamoun, M., and Owens, R. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *Pattern Analysis and Machine Intelligence 28*, 10 (2006), 1584–1601.

[36] Min, R., Kose, N., and Dugelay, J.-L. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics 44*, 11 (2014), 1534–1548.

[37] Ming, Y. Robust regional bounding spherical descriptor for 3d face recognition and emotion analysis. *Image and Vision Computing 35* (2015), 14–22.

[38] MSDN. Kinect v2 face API. `https://msdn.microsoft.com/en-us/library/microsoft.kinect.face.aspx/`, (accessed January 31, 2016).

[39] Nathan Silberman, Derek Hoiem, P. K., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision* (2012).

[40] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. In *International Conference on Machine Learning* (2011), pp. 689–696.

[41] Singh, R., Vatsa, M., and Noore, A. Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. *Pattern Recognition 41*, 3 (Mar. 2008), 880–893.

[42] Srivastava, N., and Salakhutdinov, R. R. Multimodal learning with deep boltzmann machines. In *Neural Information Processing Systems* (2012), pp. 2222–2230.

[43] Viola, P., and Jones, M. J. Robust real-time face detection. *International Journal of Computer Vision 57*, 2 (2004), 137–154.

[44] Wang, J., Liu, Z., Chorowski, J., Chen, Z., and Wu, Y. Robust 3d action recognition with random occupancy patterns. In *European Conference on Computer Vision*. 2012, pp. 872–885.

[45] Wolf, L., Hassner, T., and Taigman, Y. Descriptor based methods in the wild. In *Real-Life Images workshop at the European Conference on Computer Vision* (2008).

[46] Wu, H., Suzuki, K., Wada, T., and Chen, Q. Accelerating face detection by using depth information. In *Springer, Advances in Image and Video Technology*. 2009, pp. 657–667.

[47] 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.

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