RGB-D Face Recognition via Learning-based Reconstruction

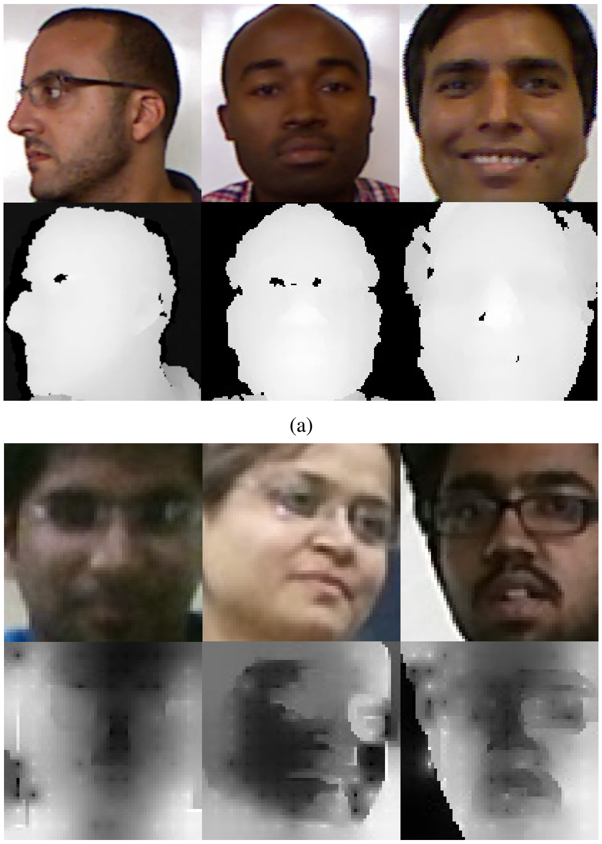
Anurag Chowdhury*, Soumyadeep Ghosh*, Richa Singh and Mayank Vatsa
IIIT-Delhi, India
{anurag1402,soumyadeepg,rsingh,mayank}@iiitd.ac.in

Abstract

Low cost RGB-D cameras are gaining significant popularity in surveillance scenarios. While RGB images contain good quality discriminative information, depth images captured in uncontrolled environment at a distance does not provide accurate depth map. In this research, we present a learning based reconstruction and mapping algorithm to generate a feature rich representation from the RGB images. These reconstructed images are then used for face identification. The experiments performed on both IIITD RGB-D database and the challenging Kasparov database show that the proposed algorithm yields significant improvements compared to when the original depth map is used for identification. Comparison with existing state-of-the-art algorithm also demonstrate the efficacy of the proposed architecture for RGB-D images.

1. Introduction

Face recognition is one of the highly investigated biometric modality. A large number of methods exist in literature [22] 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. [7] 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 such as indoor scene segmentation [17], human action recognition [21], face anti-spoofing [3], head pose estimation [4], object recognition [14], object discovery [10], face detection [8], and gender recognition [9].


(a)

(b)

Figure 1: RGB and depth images: (a) in controlled conditions (Eurecom RGBD database [15]) and (b) with large standoff distance and uncontrolled conditions (Kasparov database [1]).

In the last few years the cost of depth sensors has decreased, which has led to increased interest in RGB-D face 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 [11]. The depth map provides additional discriminative information which enhances the recognition performance. 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. Li et al. [12] presented a face recognition algorithm from low resolution 3D images. Texture transformation and

[*] Equal contribution from student authors

a sparse coding based reconstruction method is used to perform face matching. Goswami et al. [6] 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. [13] proposed a 3D keypoint based face matching algorithm using multi-task sparse representation. Elaiwat et al. [2] used a multimodal keypoint detector for identifying keypoints on a 3D surface, and both texture and 3D local features are utilized. Ming [16] proposed a regional bounding spherical descriptor for facial recognition and emotional analysis which uses regional and global regression mapping for classification.

Majority of the existing RGB-D face recognition research has focused on controlled environment. However, the sensors can also capture images from a distance in a real world deployment. As shown in Figure 1a, when captured at close distances, the quality of depth images is good. However, with large standoff distance between the camera and the subject, depth sensors fail to capture good quality depth images (Figure 1b). In such situations, it may not be advisable to use poor quality images. Therefore, in this research, we propose to utilize the complementary information available in both RGB and depth images to yield a feature rich representation for face recognition.

Several researchers have explored the applicability of image fusion algorithms to improve the performance with where multimodal information is available. For instance, Singh et al. proposed wavelet fusion based algorithm to combine images from multiple spectrums [19]. However, these algorithms either have fixed wighting 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 [18, 20] 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. Inspired from this, we introduce a new neural network based algorithm to yield feature rich representation from RGB images so that it contains more discriminative information than original depth images. During training, a mapping function is learnt between RGB and depth images that helps to yield a feature rich representation that has the properties of both RGB and depth data. The property of generating such representation 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 reconstruct one modality from the input information of the other.
- The proposed architecture is applied on RGB and depth face images to generate 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 [1], 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 [5] the proposed algorithm yield competitive accuracies with respect to [5].

2. Proposed Algorithm

This section presents the formulation of the proposed algorithm. The algorithm is presented as a generic model for learning a feature rich representation of two modalities namely M_1 and M_2 followed by a classifier for identification. The learning phase is composed of two main steps, learning the feature rich representation and learning the classifier for identification. Figures 2 and 3 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.

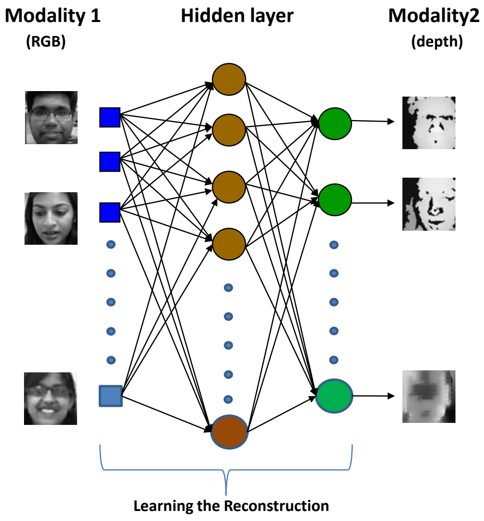


Figure 2: Illustrating the training module of the proposed algorithm.

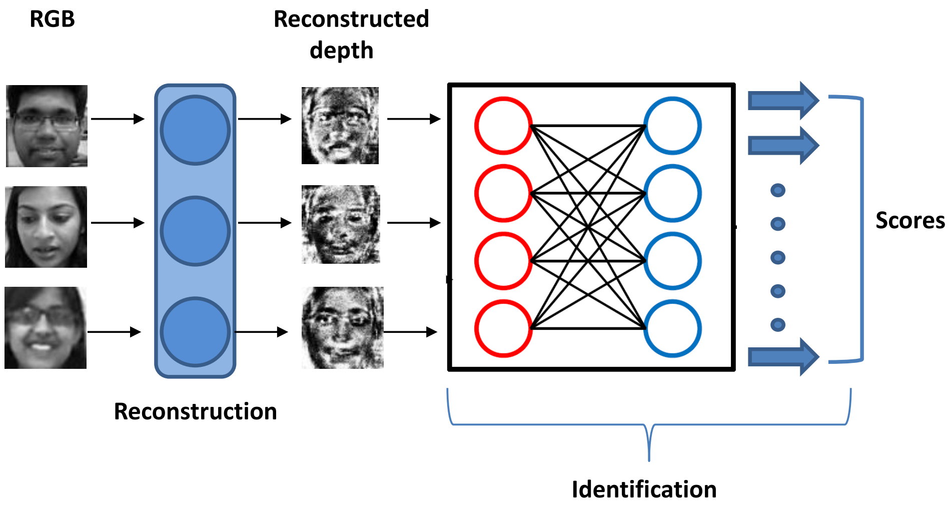


Figure 3: Illustrating the steps involved in testing with the proposed algorithm and identification using reconstructed data.

2.1. Learning Mapping and Reconstruction using Neural Network

Let $X_{M_1} = \left\{x^{(1)}_{M_1}, x^{(2)}_{M_1}, x^{(3)}_{M_1}, ...x^{(n)}_{M_1}\right\}$ be the n data samples from the first modality (e.g. RGB or grayscale images). Similarly, let $X_{M_2} = \left\{x^{(1)}_{M_2}, x^{(2)}_{M_2}, x^{(3)}_{M_2}, ...x^{(n)}_{M_1}\right\}$ 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

$$H = \phi(W_1.X_{M_1} + b_1) \tag{1}$$

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 2 to 4.

$$\begin{aligned} \hat{X}_{M_2} &= \phi(W_2.H + b_2) \\ &= \phi(W_2.\phi(W_1.X_{M_1} + b_1) + b_2) \end{aligned} \tag{2}$$

such that

$$argmin_\theta(||X_{M_2} - \hat{X}_{M_2}||_2^2 + \lambda R) \tag{3}$$

expanding Equation 3 using Equation 2,

$$argmin_\theta(||X_{M_2} - \phi(W_2.\phi(W_1.X_{M_1} + b_1) + b_2)||_2^2 + \lambda R) \tag{4}$$

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 2 and 3, 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 reconstructed depth data 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.

2.2. 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 $\hat{X}_{depth}$.

$$\hat{X}_{depth} = \phi(W_2.\phi(W_1.X_{RGB} + b_1) + b_2) \tag{5}$$

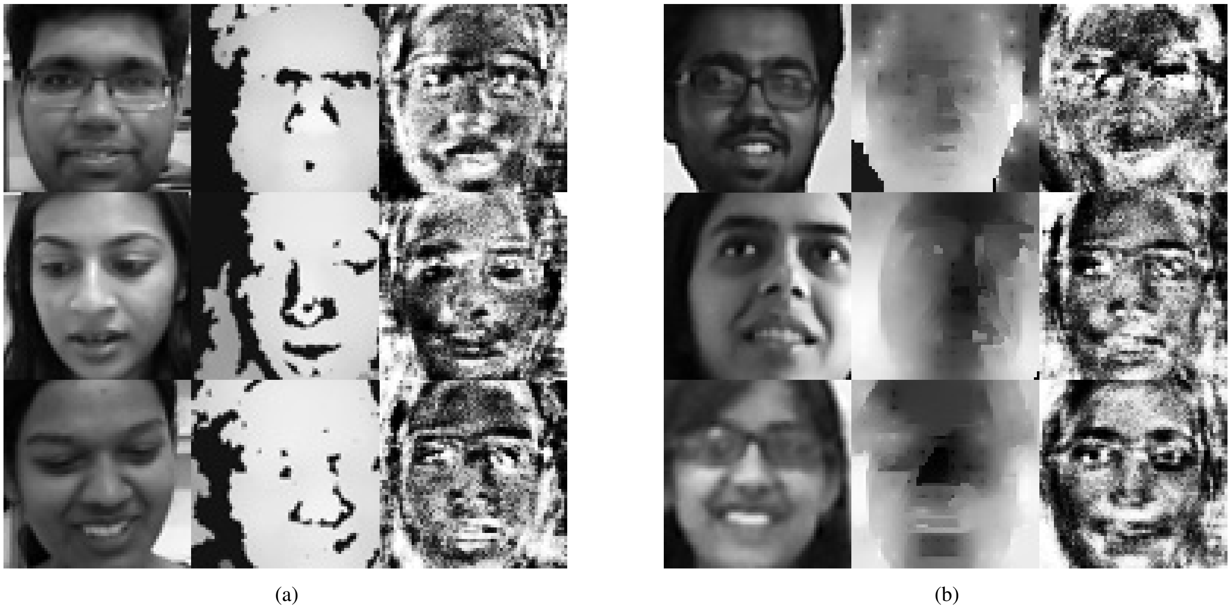


Figure 4: Visualizations of different representations used in the proposed method, (a) IIITD RGBD database [5], (b) KaspAROV database, where column 1: RGB image in grayscale, column 2: Captured depth image, column 3: Visualization of learned feature rich representation $\hat{V}_{shared}$, shows the discriminative properties of the reconstructed depth $\hat{X}_{depth}$.

such that

$$argmin_{\theta}(||X_{depth} - \hat{X}_{depth}||_2^2 + \lambda R) \qquad (6)$$

Figure 4 shows samples of the feature rich representation obtained using the proposed algorithm. It can be observed that compared to the original depth map, the obtained 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, 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 5. This representation is then given as input to the trained neural network classifier for identification.

3. Experimental Results and Analysis

For evaluating the performance of the proposed reconstruction based network, we have used two RGB-D face datasets, the IIITD RGB-D dataset [5] and the Kasparov dataset [1]. Since training the mapping function requires large training data, the proposed representation learning model (Fig. 2) is first pretrained using the EURECOM [15] RGB-D face database.

1. The EURECOM dataset [15] 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. **KaspaAROV**: 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** 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 1: Details of databases used in the experiments

Dataset	Device	Classes	Image Size		Train set	Test set
			RGB	Depth		
Eurecom	Kinect 1	52	256 × 256	256 × 256	364	-
IIITD RGBD	Kinect 1	106	Variable	Variable	9,210	13,815
KaspAROV	Kinect 2	108	64 × 64	64 × 64	31,060	31,060

3.1. 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 [5] and Eurecom [15] 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.

3.2. 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 1, 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.

3.3. 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 2.2. This is numbered as experiment 3.

3. *Recognition using feature rich representation:* As explained in Section 2, the reconstructed depth maps are feature rich representations obtained by projecting the depth image and are used for identification. This is referred as experiment 4.

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

3.4. Results and Analysis

Table 2 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 Fig. 6.

- 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 23.24% and 36.75% respectively on KaspAROV and IIITD RGB-D databases. Accuracies on some well known handcrafted features like LBP and Gabor (Experiment 5 and 6) are also provided.

- The hidden representation also gives competitive accuracies (experiment 3) with respect to the shared representation learnt by the proposed network for both KaspAROV and IIIT RGB-D databases. This shows that hidden representation H also learns discriminative information from both the modalities alike the reconstructed depth. We trained both the models (experiments 3 and 4) for equal number of epochs. The visualization of the hidden layer weights W_1 of the reconstruction network is depicted in Figure 5.

Table 2: Identification results on the IIITD RGBD and Kasparov databases.

Experiment No.	Modality 1	Modality 2	Feature	Rank-1 Identification Accuracy (in %)	
				KaspAROV Dataset	IIITD RGBD Dataset
1	Depth	-	Raw	11.80	26.81
2	RGB	-	Raw	23.24	36.75
3	RGB	Depth	Hidden	60.00	98.08
4	RGB	Depth	Reconstructed	**67.77**	**98.71**
5	RGBD	-	RISE [5]	52.38	98.74
6	RGB	-	LBP	1.63.	9.53
7	RGB	-	Gabor	2.59.	36.16

- The reconstructed depth $\hat{X}_{depth}$ obtained from the RGB to depth reconstruction network gives higher identification accuracy (experiment 4 in Table 2) compared to raw depth and RGB as the representation (experiments 1 and 2 in Table 2 respectively). 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 algorithm (experiment 4 in Table 2) significantly outperforms state-of-the-art method RISE[5] on the KaspAROV dataset, where the images are of surveillance quality. The IIITD RGB-D database contains high quality images and the reported results are already very high (over 98%). Therefore, effect of the proposed algorithm is better analyzed on Kasparov database.

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

$$\hat{V}_{shared} = \hat{X}_{depth} - mean(\hat{X}_{depth}) \qquad (7)$$

where $mean(\hat{X}_{depth})$ is the mean reconstructed depth image of the entire dataset. This visualization is depicted in Figure 4, column 3 in both (a) and (b). It can be easily observed that they are different from each other and encode 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.

4. Conclusion

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 algorithm which only needs RGB images during testing. This is accomplished using a novel representation learning model that learns the mapping and reconstruction between depth and RGB images. After training, it generates feature rich representation from RGB images which contains discriminative information from both RGB and depth images. The results show that 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.

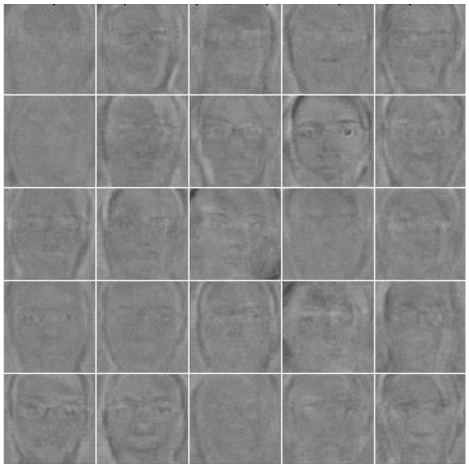


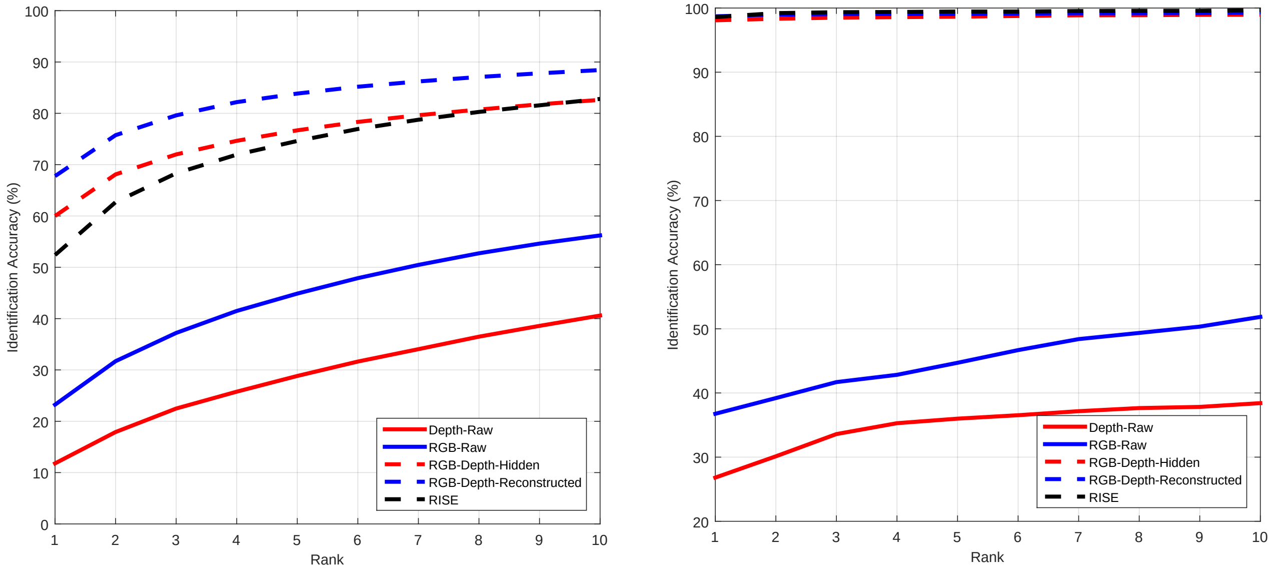
Figure 5: Visualizing the weights learnt by the hidden layer of RGB to depth mapping network.

References

[1] Kasparov kinect video dataset. `http://iab-rubric.org/resources/Kasparov.html/`, 2016.

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

[3] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IEEE BTAS*, pages 1–6, 2013.



(a)

(b)

Figure 6: CMC curves comparing the performance of the proposed and existing algorithms on the (a) KaspAROV database, (b) IIITD RGBD database.

[4] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3d face analysis. *Springer IJCV*, 101(3):437–458, 2013.

[5] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On rgb-d face recognition using kinect. In *IEEE BTAS*, pages 1–6, 2013.

[6] G. Goswami, M. Vatsa, and R. Singh. Rgb-d face recognition with texture and attribute features. *IEEE TIFS*, 9(10):1629–1640, 2014.

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

[8] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet. An rgb-d database using microsoft's kinect for windows for face detection. In *IEEE SITIS*, pages 42–46, 2012.

[9] T. Huynh, R. Min, and J.-L. Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *IEEE ACCV Workshops*, pages 133–145, 2012.

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

[11] J. Kittler, A. Hilton, M. Hamouz, and J. Illingworth. 3d assisted face recognition: A survey of 3d imaging, modelling and recognition approachest. In *IEEE CVPR Workshops*, pages 114–114, 2005.

[12] B. Y. Li, A. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE WACV Workshops*, pages 186–192, 2013.

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

[14] A. S. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE PAMI*, 28(10):1584–1601, 2006.

[15] R. Min, N. Kose, and J.-L. Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE SMC*, 44(11):1534–1548, Nov 2014.

[16] Y. Ming. Robust regional bounding spherical descriptor for 3d face recognition and emotion analysis. *Elsevier IVC*, 35:14–22, 2015.

[17] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *IEEE ECCV*, 2012.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

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

[20] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *IEEE ECCV*, pages 872–885. 2012.

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