On RGB-D Face Recognition using Kinect

Gaurav Goswami, Samarth Bharadwaj, Mayank Vatsa, and Richa Singh
IIIT Delhi
{gauravgs, samarthb, mayank, rsingh}@iiitd.ac.in

Abstract

Face recognition algorithms generally use 2D images for feature extraction and matching. In order to achieve better performance, 3D faces captured via specialized acquisition methods have been used to develop improved algorithms. While such 3D images remain difficult to obtain due to several issues such as cost and accessibility, RGB-D images captured by low cost sensors (e.g. Kinect) are comparatively easier to acquire. This research introduces a novel face recognition algorithm for RGB-D images. The proposed algorithm computes a descriptor based on the entropy of RGB-D faces along with the saliency feature obtained from a 2D face. The probe RGB-D descriptor is used as input to a random decision forest classifier to establish the identity. This research also presents a novel RGB-D face database pertaining to 106 individuals. The experimental results indicate that the RGB-D information obtained by Kinect can be used to achieve improved face recognition performance compared to existing 2D and 3D approaches.

1. Introduction

Face recognition is a challenging problem which suffers not only from the general object recognition challenges such as illumination and viewpoint variations but also from distortions or covariates specific to faces such as expression, accessories, and high inter-class similarity of human faces [2]. 2D images, generally used for face recognition, can only encompass a limited amount of information about a face; therefore, researchers have also considered the use of 3D information captured using special sensors [5, 20]. Although incorporating 3D with 2D has led to improvements compared to the use of only 2D images, the high cost of specialized sensors limit the applicability of such approaches in practical implementations.

With the advent of new sensing technology, 2D color image (RGB) along with the depth map (D) can now be obtained using low cost web camera style sensors such as Microsoft's Kinect. The depth map provides per pixel depth information obtained using an infrared laser projector combined with a camera. While it does not provide a *true* 3D mesh, an RGB-D image provides more information about the captured object than a traditional 2D color image. RGB-D images have been applied to a variety of tasks such as general object recognition, surface modeling and tracking, modeling indoor environments, discovering objects in scenes, and robotic vision [4, 17, 18]. Recent papers have also used RGB-D face images for face detection and gender recognition [11, 13]. Li *et al.* [15] describe an algorithm that uses RGB-D images for recognizing faces under the effect of various covariates. The algorithm combines Discriminant Color Space transform with sparse coding to perform recognition. Kinect also has it's own face recognition capabillities which is based on algorithm explained in [6].

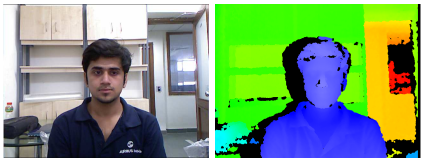

(a) RGB image (b) Depth map

Figure 1: Example RGB-D image captured by Kinect.

Figure 1 shows an RGB-D face image (RGB image and accompanying depth map) captured using a Kinect sensor. It is our assertion that RGB-D images can potentially be utilized to mitigate the effect of covariates, specifically illumination, pose and expression. Since the nature of an RGB-D image obtained from Kinect and a regular 3D map is fundamentally different, existing 3D face recognition techniques may not directly be applicable on these images. Each pixel in Kinect's depth map has a value indicating the relative distance of that pixel from the sensor at the time of image capture. As shown in Figure 2, depth maps captured from Kinect exhibit very high inter-class similarity (due to noise and holes), and therefore may not be able to differentiate among different individuals. However, it has low intra-class variation which can be utilized to increase the robustness to covariates such as expression and pose. Further, color

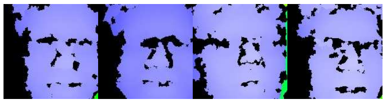


Figure 2: Depth maps of different individuals.

images can provide inter-class differentiability which depth data lacks. Therefore, it is important to utilize both RGB and depth data for feature extraction and classification.

This paper focuses on face recognition using RGB-D data and the contributions are two-fold: (1) entropy and saliency based algorithm that uses both RGB and depth information for face recognition is proposed and (2) a face database consisting of RGB-D images pertaining to 106 subjects captured exclusively using Kinect is prepared. The experiments suggest that the depth information in conjunction with RGB data improves the recognition performance.

2. Proposed Algorithm

The steps involved in the proposed algorithm are shown in Figure 3. The algorithm first computes four entropy maps corresponding to RGB and depth information with varying patch sizes and a visual saliency map of the RGB image. The Histogram of Oriented Gradients (HOG) descriptor [8] is then used to extract features from these five entropy/saliency maps. Concatenation of five HOG descriptors provides the final feature descriptor which is used as input to the trained Random Decision Forest (RDF) classifier for establishing the identity.

2.1. Extracting Entropy Maps and Visual Saliency Map

Entropy is defined as the measure of uncertainty in a random variable [19]. The entropy H of a random variable $\mathbf{x}$ is $H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i)$, where $p(x_i)$ is the value of the probability mass function for x_i. The visual entropy map of an image is a characteristic of its texture and can be used to extract meaningful information from an image. Figures 4(a) and 4(b) show an RGB face image and the corresponding entropy maps respectively. On the other hand, Figures 4(d) and 4(e) show the depth map and the depth entropy map respectively. As discussed earlier, it can also be seen that the depth map has very minor variations and therefore, the information in the depth map looks insignificant for feature extraction. However, the depth entropy map amplifies these variations, making them far more distinguishable and produces a signature of this depth map which can be used for extracting features.

Let the input image be denoted as a pair of intensity functions, $[I_{rgb}(x, y), I_d(x, y)]$, where $I_{rgb}(x, y)$ is the RGB image and $I_d(x, y)$ is the depth map, each of size $M \times N$. Let both of these be defined over the same set of (x, y) points, Z, such that $x \in [1, M]$ and $y \in [1, N]$. Let $H(I_j)$ denote the visual entropy map of the image I_j. Here, I_j can be the depth map or the RGB image or a small part of these images (i.e. restricted to a subset Z' such that $Z' \subset Z$). Two image patches are extracted for both I_{rgb} and I_d. Two patches P_1, of size $\frac{M}{2} \times \frac{N}{2}$ centered at $[\frac{M}{2}, \frac{N}{2}]$, and P_2, of size $\frac{3M}{4} \times \frac{3N}{4}$ centered at $[\frac{M}{2}, \frac{N}{2}]$, are extracted from I_{rgb}. Similarly, two patches P_3 and P_4 are extracted from I_d. Four entropy maps $E_1 - E_4$ are computed for patches $P_1 - P_4$ using Equation 2:

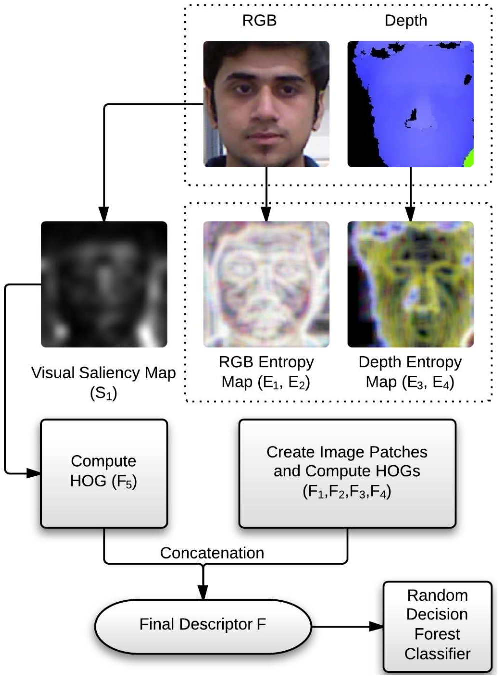
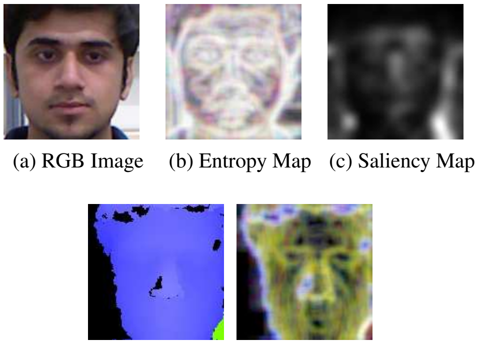

RGB
Depth
Visual Saliency Map (S1)
RGB Entropy Map (E1, E2)
Depth Entropy Map (E3, E4)
Compute HOG (F5)
Create Image Patches and Compute HOGs (F1,F2,F3,F4)
Concatenation
Final Descriptor F
Random Decision Forest Classifier

Figure 3: Illustrating the steps of the proposed algorithm.

$$E_i = H(P_i), \; where, \; i \in [1, 4] \tag{1}$$

E_1, E_2 represent the entropy of the color image (I_{rgb}) and E_3, E_4 represent the depth entropy maps.

Apart from entropy, we also utilize the *saliency* of the RGB image to compute useful face information. Visual saliency is the capability of an image region to attract a viewer's visual attention [9]. The distribution of visual attention across the entire image is termed as the visual saliency map of the image. Let the image be represented as $I(x, y)$. Its saliency map can be denoted as an intensity function $S(\cdot)$, which maps the individual pixels to an intensity value proportional to the saliency of that particular pixel. Figure 4(c) presents an example of the visual saliency map of an input image shown in Figure 4(a). There are several techniques to compute the visual saliency map of an image. In this research, the implementation based on [1, 14] is used to compute the visual saliency maps of the RGB im-



(a) RGB Image (b) Entropy Map (c) Saliency Map

(d) Depth Map (e) Depth Entropy

Figure 4: RGB-D image with corresponding entropy and saliency maps. (a) RGB image, (b) entropy map and (c) saliency map of RGB image, (d) depth map and (e) depth entropy map of (d).

ages. The nature of the quantities that a visual image and a depth map represent are fundamentally different. While an image represents intensity information, a depth map is a measure of distance. Since visual saliency methods have been designed specifically for visual images, direct application of such techniques on depth maps tends to produce irregular output. Therefore, only the color image is used for extracting saliency. The proposed algorithm extracts the visual saliency map S_1 of the color image I_{rgb} using Equation 3,

$$S_1(x, y) = S(I_{rgb}(x, y)\forall(x \in [1, M], y \in [1, N])) \quad (2)$$

2.2. Extracting Features using HOG

HOG [8] descriptor produces the histogram of a given image in which pixels are binned according to the magnitude and direction of their gradients. It is a robust descriptor with fixed lenght feature and has been used successfully in many applications such as object detection and recognition [7, 10]. In the proposed algorithm, HOG is applied on the entropy and saliency maps. The entropy maps are extracted from patches P_1 and P_2 which have the same center but different size. By extracting entropy in two patch sizes, multiple granularities of the input image can be captured. Let $D(\cdot)$ denote the HOG histogram; the proposed algorithm computes HOG of entropy maps using the following equation:

$$F_i = D(E_i), \, where, \, i \in [1, 4] \quad (3)$$

Here, F_1 represents the HOG of the entropy map E_1 defined over patch P_1 and F_2 represents the HOG of the entropy map E_2 defined over patch P_2 in I_{rgb}. Similarly, F_3 and F_4 represent the HOGs of entropy maps E_3 and E_4 defined over patch P_3 and P_4 in I_d respectively. F_1 and F_2 capture traditional texture information. Further, instead of directly using visual information, the entropy maps are used to make the descriptor robust against intra-class variations. F_3 and F_4 capture the depth information embedded in the RGB-D image. As discussed earlier, the depth information in it's original state may not be useful for feature extraction. Therefore, it is first converted into the entropy map and features are extracted from the entropy map by applying the HOG descriptor. The final descriptor F is created using an ordered concatenation of the five HOG histograms, where F_5 is the HOG descriptor of the visual saliency map S_1.

$$F_5 = D(S_1(I_{rgb})) \quad (4)$$

$$F = [F_1, F_2, F_3, F_4, F_5] \quad (5)$$

The feature vector F is provided as input to the multi-class classifier explained in the next section.

2.3. Classification

A multi-class classifier such as Nearest Neighbor (NN), Random Decision Forests (RDFs) [12], and Support Vector Machines (SVM) can be used to establish the identity of a given probe. However, the classifier should be robust for large number of classes, computationally inexpensive during probe identification, and accurate. Among several choices, RDFs being an ensemble of classifiers, can produce non-linear decision boundaries and handle the multi-class case much better than SVM. Unlike NN, RDFs are robust to outliers as well since every tree in the forest is only trained with a small subset of the data. Hence, the probability of an entire collection of trees making an incorrect decision due to a few outlier data points is usually very low. Therefore, in this research, RDF is explored for classification. In RDF training, the number of trees in the forest and the fraction of training data used to train an individual tree, control the generalizability of the forest. These parameters are obtained using the training samples and a small grid search. Here, each feature descriptor is a data point and the subject identification number is the class label, therefore, the number of classes is equal to the number of subjects. The trained RDF is then used for probe identification.

3. Experimental Results

The performance of the proposed algorithm is analyzed on two RGB-D face databases and compared with several existing algorithms.

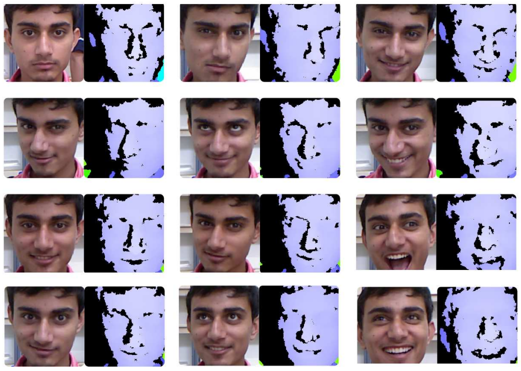


Figure 5: Sample images from the IIIT-D RGB-D face database.

3.1. RGB-D Face Databases and Experimental Protocol

There exist a few RGB-D face databases which are publicly available [11, 13, 15]. The maximum size of such databases is approximately 50 subjects. However, to perform face recognition experiments, a larger database is preferable. Therefore, to evaluate the proposed approach, we have prepared the IIIT-D RGB-D face database comprising 106 male and female subjects with multiple RGB-D images of each subject. All the images are captured using a Microsoft Kinect sensor. The OpenNI API captures the RGB image and depth map as separate 24-bit images. Since the images are unsegmented, the database can be used for both face detection and recognition in RGB-D space. The number of images per subject are variable with a minimum of 11 images and a maximum of 254. The total number of images in the database is 4605 and the size of each image is 640×480. Besides the proposed database, the EURECOM database [13] is utilized for experimentation. It has 936 images pertaining to 52 subjects and the images are captured with variations in pose, illumination, and occlusion.

First, the faces are detected using the Viola-Jones face detector [21]. Since the RGB images and their corresponding depth maps are registered, corresponding area is segmented from the depth map as well. The detected RGB-D faces include variations in expression and minor variations in illumination and pose. Figure 5 shows an example of the detected RGB-D face images of an individual. For the IIIT-D RGB-D database, four images per subject are used as gallery and for the EURECOM database the gallery size is fixed at two. The remaining images from the databases are used as probes. Five fold random subsampling based cross validation is performed and Cumulative Match Characteristic (CMC) curves for the average accuracies are computed for each experiment. The results are also compared with existing 2D and 3D face recognition algorithms.

3.2. Results and Analysis

As discussed in Section 2, the proposed algorithm has various components: *entropy*, *saliency*, and *depth* information. The experiments are performed to analyze the effect and relevance of each component. The performance of the proposed algorithm is computed in the following six cases: (a) without entropy (RGB-D image is used directly instead of entropy maps), (b) without depth (descriptor without F_3 and F_4), (c) without saliency (descriptor without F_5), (d) without entropy and depth (RGB-D image is used directly instead of entropy maps, and descriptor without F_3 and F_4), (e) without entropy and saliency (RGB-D image is used directly instead of entropy maps, and descriptor without F_5), and (f) without depth and saliency (descriptor without F_3, F_4, and F_5). These cases analyze the effect of different components on the overall performance. For example, if the descriptor performs poorly in case (a), it suggests that not using entropy maps for feature extraction is detrimental to the descriptor. Similar inferences can be drawn from the results of other five cases. Furthermore, the performance of the descriptor in these conditions can provide indications towards the utility of incorporating these components together. Finally, comparing the performance of the proposed descriptor with entropy, saliency and depth information can determine whether the proposed combination of components improves the face recognition performance with respect to the individual components.

The CMC curves Figure 6 show that removing any of the components significantly reduces the performance of the proposed algorithm. For example, case (c) shows that the contribution of including visual saliency map as an added feature is important. We observe that saliency is relevant towards stabilizing the feature descriptor and preserving intra-class similarities. Further, in cases (d) and (e), it can be noted that including depth without computing entropy performs worse than not including depth information but using entropy maps to characterize the RGB image. Intuitively, this indicates that directly using depth map results in more performance loss than not using depth at all. This is probably due to the fact that depth data from Kinect is noisy and increases intra-class variability in unaltered form. Figure 6 also shows that RDF outperforms the nearest neighbor classifier. However, it is to be noted that with the proposed descriptor, both the classifiers provide higher accuracy than HOG alone.

Next, the proposed descriptor is compared against Four Patch Local Binary Patterns (FPLBP) [22], Pyramid Histogram of Oriented Gradients (PHOG) [3], Scale Invariant Feature Transform (SIFT) [16], [24], and Sparse representation [23]. Besides these methods which utilize only 2D information, a comparison is also performed with a 3D-PCA based algorithm which computes a subspace based on depth and grayscale information. Figures 7 and 8 along with Table

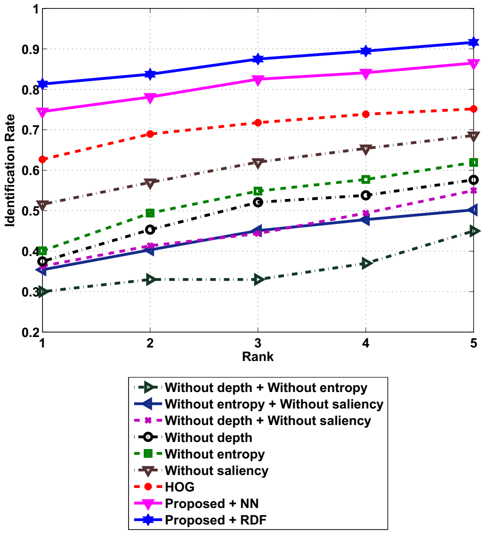


Figure 6: Analysis of the proposed algorithm and its individual components.

Modality	Approach	Rank-5 Accuracy (%)	
		IIIT-D	EURECOM
2D (Only image)	SIFT	50.1±1.4	83.8±2.1
	HOG	75.1±0.7	89.5±0.8
	PHOG	81.6±1.4	90.5±1.0
	FPLBP	85.0±0.7	94.3±1.4
	Sparse	87.2±1.9	84.8±1.7
3D (Image & depth map)	3D-PCA	83.4 ±2.1	94.1±2.7
	Proposed	91.6±1.2	98.1±1.1

Table 1: Illustrating the identification accuracy along with standard deviation of face recognition algorithms on the IIIT-D RGB-D face database and the EURECOM Kinect Face Dataset.

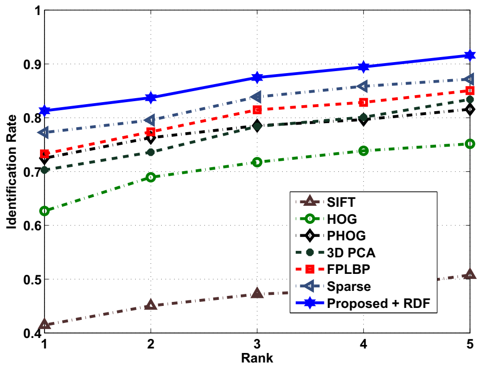


Figure 7: Comparing the proposed algorithm with existing algorithms on the IIIT-D RGB-D face database.

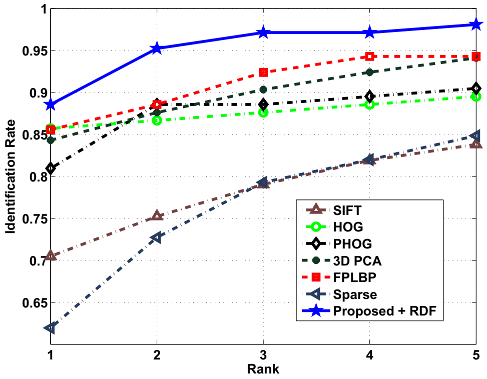


Figure 8: Comparing the proposed algorithm with existing descriptors on the EURECOM Kinect Face dataset.

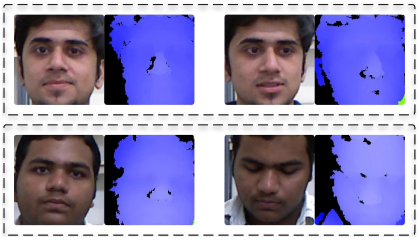


Figure 9: Analyzing the performance of the proposed algorithm. The two rows present gallery and probe images when (a) (Case 1) all the algorithms are able to recognize and (b) (Case 2) only the proposed algorithm is able to correctly identify the subject at rank-1.

1 summarize the results of this comparison. It is observed that the proposed algorithm with RDF classifier performs significantly better than other 2D and 3D algorithms over all the ranks on both the databases. This supports the hypothesis that including depth information along with RGB information can help in improving the face recognition performance. Furthermore, the length of the proposed descriptor is also shorter than the length of the descriptors produced by FPLBP, PHOG, and SIFT. Since the length of each HOG vector is 81 (9 windows of size 3×3), the final vector produced by the proposed algorithm is $81 \times 5 = 405$. In comparison, each FPLBP vector has length 560 and PHOG vector has length 680. SIFT is a keypoint based descriptor having variable number of keypoints per image with each keypoint being associated with a histogram of length 128. Therefore, we can assert that the proposed descriptor provides a robust and concise representation of face images for the purpose of face identification.

In order to further analyze the performance, we examine two types of results. Figure 9 contains two example cases of gallery and probe images. Case 1 is when all the algorithms could match the probe to the gallery image and successfully identify the subject. Case 2 is when only the proposed algorithm is able to identify the subject and other algorithms fail. As it can be seen from the example images of Case 1, when there are minor variations in expression and pose, all the algorithms are able to correctly identify. However, as shown in case 2, the proposed algorithm is able to correctly recognize even in presence of significant pose or expression variations. Thus, it can be concluded that the proposed descriptor outperforms some existing 2D and 3D approaches. This difference in performance can be attributed to the following reasons:

- The proposed descriptor uses depth information in addition to traditional color information. After amplification by visual entropy, the depth map is able to mitigate the effects of illumination and expression.
- The proposed descriptor utilizes saliency map for feature extraction which models visual attention. The saliency distribution of a face is not significantly affected by minor pose and expression variations and therefore it provides tolerance to these variations.
- Compared to existing 3D approaches, entropy and saliency maps of RGB-D images are less affected by noise such as holes in depth map and low resolution, and therefore, yield higher performance.

4. Conclusion

Existing face recognition algorithms generally utilize the 2D or 3D information for recognition. However, the performance and applicability of existing face recognition algorithms is bound by the information content or cost implications. In this paper, we have proposed a novel algorithm that utilizes the depth information along with RGB images obtained from Kinect, to improve the recognition performance. The proposed algorithm uses a combination of entropy, visual saliency, and depth information with HOG for feature extraction and random decision forest for classification. The experiments, performed on Kinect face databases, demonstrate the effectiveness of the proposed algorithm and show that it outperforms some existing 2D and 3D face recognition approaches.

5. Acknowledgement

The authors would like to thank Dr. Jean-luc Dugelay for providing the EURECOM database, Dr. Subhashis Banerjee (IIT Delhi) for his feedback, and DietY for supporting this research.

References

[1] J. Harel, a saliency implementation in MATLAB: http://www.klab.caltech.edu/ harel/share/gbvs.php.

[2] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *PRL*, 28(14):1885–1906, 2007.

[3] Y. Bai, L. Guo, L. Jin, and Q. Huang. A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In *ICIP*, pages 3305–3308, 2009.

[4] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *IROS*, pages 821–826, 2011.

[5] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches to three-dimensional face recognition. In *ICPR*, volume 1, pages 358–361, 2004.

[6] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, pages 2707 –2714, 2010.

[7] E. Corvee and F. Bremond. Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *AVSS*, pages 469–475, 2010.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[9] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *ARN*, 18(1):193–222, 1995.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[11] R. I. 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 *SITIS*, pages 42–46, 2012.

[12] T. K. Ho. Random decision forests. In *ICDAR*, pages 278–282, 1995.

[13] 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 *ACCV*, 2012.

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254 –1259, 1998.

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

[16] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157, 1999.

[17] Y. Park, V. Lepetit, and W. Woo. Texture-less object tracking with online training using an RGB-D camera. In *ISMAR*, pages 121–126, 2011.

[18] A. Ramey, V. Gonzalez-Pacheco, and M. A. Salichs. Integration of a low-cost RGB-D sensor in a social robot for gesture recognition. In *HRI*, pages 229–230, 2011.

[19] A. Rrnyi. On measures of entropy and information. In *BSMSP*, pages 547–561, 1961.

[20] A. Scheenstra, A. Ruifrok, and R. Veltkamp. A survey of 3D face recognition methods. In *AVBPA*, pages 325–345, 2005.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518, 2001.

[22] L. Wolf, T. Hassner, Y. Taigman, et al. Descriptor based methods in the wild. In *ECCV Real Faces Workshop*, 2008.

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009.

[24] L. Zhang, J. Chen, Y. Lu, and P. Wang. Face recognition using scale invariant feature transform and support vector machine. In *ICYCS*, pages 1766–1770, 2008.