RGB-D Face Recognition with Texture and Attribute Features

Gaurav Goswami, *Student Member, IEEE,* Mayank Vatsa, *Senior Member, IEEE*, and Richa Singh, *Senior Member, IEEE*

***Abstract*—Face recognition algorithms generally utilize 2D images for feature extraction and matching. To achieve higher resilience towards covariates such as expression, illumination and pose, 3D face recognition algorithms are developed. While it is highly challenging to use specialized 3D sensors due to high cost, RGB-D images can be captured by low cost sensors such as Kinect. This research introduces a novel face recognition algorithm using 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. Geometric facial attributes are also extracted from the depth image and face recognition is performed by fusing both the descriptor and attribute match scores. The experimental results indicate that the proposed algorithm achieves high face recognition accuracy on RGB-D images obtained using Kinect compared to existing 2D and 3D approaches.**

***Index Terms*—Face Recognition, Saliency, Entropy, RGB-D, Kinect.**

I. Introduction

FACE recognition with 2D images is a challenging problem especially in the presence of covariates such as pose, illumination, expression, disguise, and plastic surgery. These covariates introduce high degree of variation in two 2D images of the same person thereby reducing the performance of recognition algorithms [3], [9], [20]. Therefore, it is desirable to perform face recognition using a representation which is less susceptible to such distortions. While 2D images are not robust to these covariates, 3D images offer a comparatively resilient representation of a face. 3D images can capture more information about a face, thus enabling higher preservation of facial detail under varying conditions. 3D face recognition has been explored in literature and several algorithms have been developed [5]. While it is advantageous to utilize 3D images for face recognition, the high cost of specialized 3D sensors limits their usage in large scale applications.

With advancements in sensor technology, low cost sensors have been developed that provide (pseudo) 3D information in the form of RGB-D images. As shown in Fig. 1, an RGB-D image consists of a 2D color image (RGB) along with a depth map (D). RGB image provides the texture and appearance information whereas depth map provides the distance of each pixel from the sensor. The depth map is a characterization of the geometry of the face with grayscale values representing the distance of each point from the sensor. While a RGB-D image does not provide highly accurate 3D information, it captures more information compared to a 2D image alone.

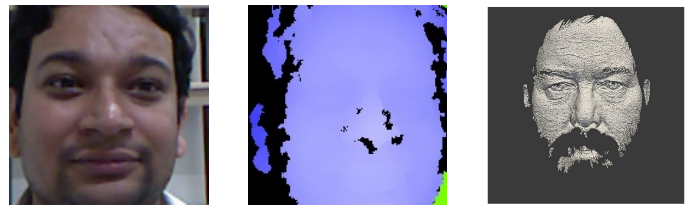

Fig. 1. Different modes of capture: (a) RGB image, (b) depth map captured using Kinect, and (c) Range image from 3D TEC dataset [30] obtained using a 3D scanner.

An RGB-D image captured using consumer devices such as Kinect is fundamentally different from a 3D image captured using range sensors due to the manner in which they capture the target. Kinect captures RGB-D image by utilizing an infrared laser projector combined with a monochrome CMOS sensor. 3D sensors on the other hand utilize specialized high quality sensors to obtain accurate range and texture image. 3D face recognition approaches utilize techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to characterize a 3D face model. Some approaches also utilize facial landmarks identified in a 3D face model to extract local features. However, 3D face recognition algorithms generally rely on accurate 3D data. Since the depth map returned by RGB-D Kinect sensor is not as precise as a 3D sensor and contains noise in the form of holes and spikes, existing 3D face recognition approaches may not be directly applied to RGB-D images. While RGB-D images have been used for several computer vision tasks such as object tracking, face detection, gender recognition, and robot vision [11], [14], [15], [17], [18], [26], there exists relatively limited work in face recognition. Li et al. [24] proposed a face recognition framework based on RGB-D images. The RGB-D face image obtained from Kinect is cropped using the nose tip which is reliably detectable via the depth map. The face is then transformed into a canonical frontal representation and pose correction is performed using a reference face model. The missing data is filled by symmetric filling which utilizes the symmetry of human faces to approximate one side of the face with corresponding points from the other side. Smooth resampling is then performed to account for holes and spikes. The image is converted using Discriminant Color Space (DCS) transform [32], [35] and the three channels are stacked into

This manuscript has been accepted for publication in the IEEE Transactions on Information Forensics and Security, 2014. A shorter version of the manuscript received the best poster award at the IEEE 6^{th} International Conference on Biometrics: Theory, Applications and Systems, 2013 [13].

G. Goswami, M. Vatsa and R. Singh are with IIIT-Delhi, India. Email: {gauravgs, mayank, rsingh}@iiitd.ac.in.

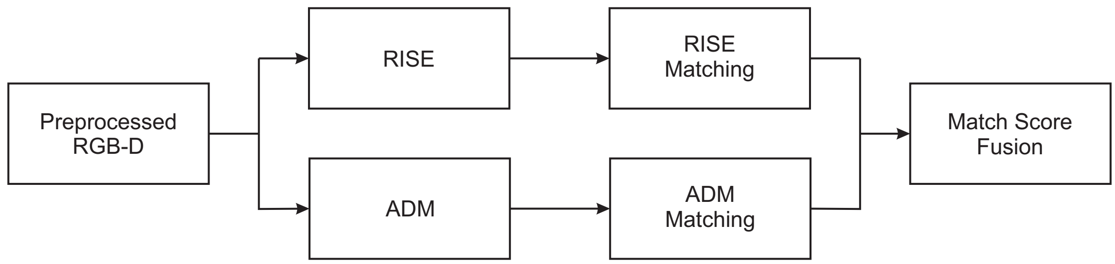

Fig. 2. Illustrating the steps involved in the proposed RGB-D face recognition algorithm.

one augmented vector. This vector and the depth map are individually matched via Sparse Representation Classifier [34] and the scores are combined. Experimental results indicate that using both depth and color information yields around 6% higher identification accuracy compared to color image based algorithms. Segundo et al. [29] proposed a continuous face authentication algorithm which utilizes Kinect as the RGB-D sensor. The detected face image is aligned to an average face image using the Iterative Closest Point (ICP) algorithm [2] and a region of interest (ROI) is extracted. The ROI is then characterized using Histogram of Oriented Gradients (HOG) approach and utilized for matching with stored user template for authentication. Kinect also has its own algorithm for face recognition, the details of which are not publicly available.

While there are few algorithms that utilize RGB-D images obtained from consumer devices for face recognition, this research presents a different perspective. As mentioned previously, the depth maps obtained using Kinect are noisy and of low resolution. Therefore, instead of using the depth information to generate a 3D face model for recognition, we utilize noise tolerant features for extracting discriminatory information. We propose a novel face recognition algorithm that operates on a combination of entropy and saliency features extracted from the RGB image and depth entropy features extracted from the depth map. The proposed algorithm also utilizes geometric attributes of the human face to extract geometric features. These geometric features are utilized in conjunction with the entropy and saliency features to perform RGB-D face recognition. The key contributions of this research are:

- A novel algorithm is developed that uses both texture (oriented gradient descriptor based on saliency and entropy features) and geometric attribute features for identifying RGB-D faces.
- IIIT-D RGB-D face database of 106 individuals is prepared and shared with the research community to promote further research in this area. A detailed experimental protocol along with train-test splits are also shared to encourage other researchers to report comparative results.

II. Proposed RGB-D Face Recognition Algorithm

The steps involved in the proposed algorithm are shown in Fig. 2. The algorithm is comprised of four major steps: (a) preprocessing, (b) computing texture descriptor from both color image and depth map using entropy, saliency, and HOG [7], (c) extracting geometric facial features from depth map, and (d) combining texture and geometric features for classification. These steps are explained in the following subsections.

A. Preprocessing

First, an automatic face detector (Viola-Jones face detector) is applied on the RGB image to obtain the face region. The corresponding region is also extracted from the depth map to crop the face region in depth space. While texture feature descriptor does not require image size normalization, the images are resized to 100×100 to compute depth features. Depth map is then preprocessed to remove noise (holes and spikes). Depth map of a face is divided into 25×25 blocks and each block is examined for existence of holes and spikes. Depth values identified as the hole/spike are rectified using linear interpolation, i.e. assigned the average value of their 3×3 neighborhood.

B. RISE: RGB-D Image descriptor based on Saliency and Entropy

The motivation of the proposed *RGB-D Image descriptor based on Saliency and Entropy* (termed as RISE descriptor) lies in the nature of the RGB-D images produced by Kinect. Specifically, as shown in Fig. 3, depth information obtained from Kinect has high inter-class similarity and may not be directly useful for face recognition. It is our assertion that 3D reconstruction based approaches may not be optimal in this scenario. However, due to low intra-class variability, depth data obtained from Kinect can be utilized to increase robustness towards covariates such as expression and pose after relevant processing/feature extraction. On the other hand, 2D color images can provide inter-class differentiability which depth data lacks. Since the color images contain visible texture properties of a face and the depth maps contain facial geometry, it is important to utilize both RGB and depth data for feature extraction and classification. As shown in Fig. 4, four entropy maps corresponding to both RGB and depth information and a visual saliency map of the RGB image are computed. The HOG descriptor [7] is then used to extract features from these five entropy/saliency maps. The concatenation of five HOG descriptors provides the texture feature descriptor which is used as input to the trained Random Decision Forest (RDF) classifier to obtain the match score.

1) Entropy and Saliency: Entropy is defined as the measure of uncertainty in a random variable [28]. Similarly, the entropy of an image characterizes the variance in the grayscale

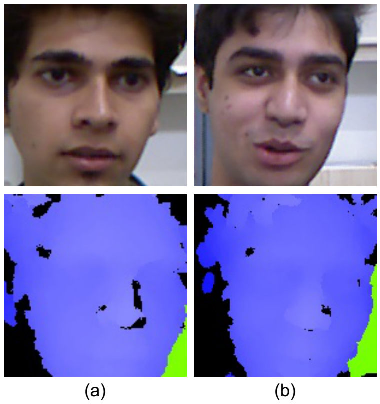

Fig. 3. RGB-D images of two subjects illustrating the inter-class similarities of RGB images and depth maps.

levels in a local neighborhood. The entropy H of an image neighborhood $\mathbf{x}$ is given by Equation 1,

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \tag{1}$$

where $p(x_i)$ is the value of the probability mass function for x_i. In the case of images, $p(x_i)$ signifies the probability that grayscale x_i appears in the neighborhood and n is the total number of possible grayscale values, i.e., 255. If $\mathbf{x}$ is a $M_H \times N_H$ neighborhood then

$$p(x_i) = \frac{n_{x_i}}{M_H \times N_H} \tag{2}$$

Here, n_{x_i} denotes the number of pixels in the neighborhood with value x_i. $M_H \times N_H$ is the total number of pixels in the neighborhood. By controlling the size of neighborhood, entropy computation can be performed at a fine or coarse level. In the current research, the neighborhood size for entropy map computation is fixed at 5×5 and RGB input images are converted to grayscale. The visual entropy map of an image is a characteristic of its texture and can be used to extract meaningful information from an image. Examples of entropy and depth entropy maps are presented in Fig. 4. The absolute values of the depth entropy map do not vary abruptly in adjacent regions except in special regions such as near the eye sockets, nose tip, mouth, and chin. The local entropy of an image neighborhood measures the amount of randomness in texture (in local region). Higher local entropy represents higher prominence and therefore, it can be viewed as a texture feature map that encodes the uniqueness of the face image locally and allows for a robust feature extraction.

Apart from entropy, we also utilize *visual saliency* of the RGB image to compute useful facial information. It measures the capability of local regions to attract the viewer's visual attention [8]. The distribution of visual attention across the entire image is termed as visual saliency map of the image. There are several approaches to compute the visual saliency map of an image. This research utilizes the approach proposed by Itti et al. [19]. Let the image be represented as an intensity function which maps a set of co-ordinates (x, y) to intensity values. The approach preprocesses a color image to normalize

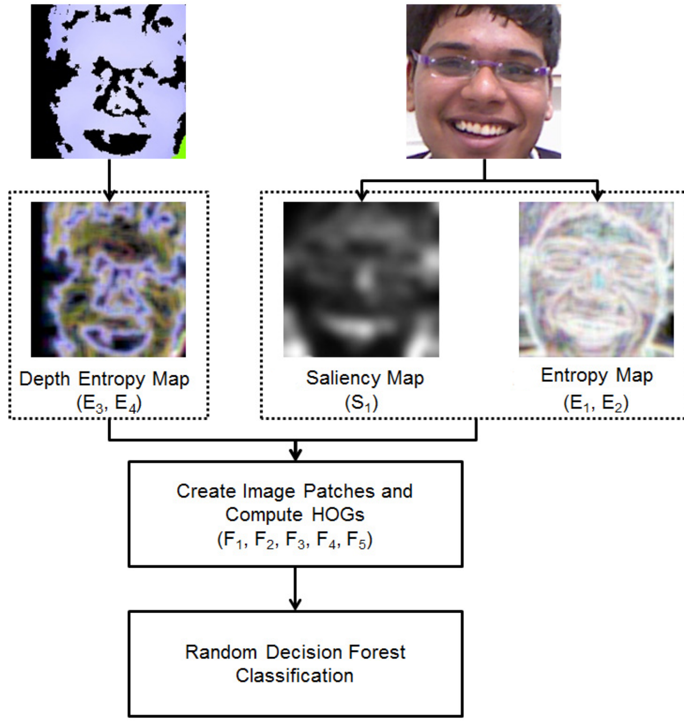

Fig. 4. Illustrating the steps of the proposed RISE algorithm.

the color channels and de-couple hue from intensity. After normalization, center-surround differences are utilized to yield the feature maps [19]. 42 feature maps are extracted from the image in accordance with the visual cortex processing in mammals. Six of these maps are computed for intensity, 12 for color, and 24 for orientation across multiple scales. Intensity and orientation feature maps are denoted by I and O respectively. The color feature maps are represented by RG and BY which are created to account for color double opponency in the human primary visual cortex [10]. Based on these maps, the saliency map of the image is computed by cumulating the individual feature maps obtained at different scales to one common scale ($= 4$) of the saliency map. This is achieved after inhibiting the feature maps which are globally homogeneous and promoting the maps which comprise of few unique activation spots (global maxima) via a normalization function $N(\cdot)$. The feature maps for color, intensity and orientation are combined in separate groups to create three feature maps C_{final}, I_{final}, and O_{final} corresponding to color, intensity, and orientation respectively.

$$C_{final} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N(RG(c,s)) + N(BY(c,s))] \tag{3}$$

$$I_{final} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \tag{4}$$

$$O_{final} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \right) \tag{5}$$

Here, c and s denote the center and surround scales respectively and the $\bigoplus$ operator denotes across-scale addition which is defined to consist of reduction of each map to the common scale and point-wise addition [19]. These maps are then combined into the final visual saliency map S according to equation 6:

$$S = \frac{1}{3}[N(C_{final}) + N(I_{final}) + N(O_{final})] \tag{6}$$

Fig. 4 presents an example of the visual saliency map, S, of an input face image. It models the image regions with high feature activation in accordance with the visual processing that occurs in the visual cortex of mammals. It is observed that gradient orientations of this saliency map provide discriminative information which aids in improving the recognition performance, specifically in reducing the intra-class discrepancies. Therefore, orientation histogram of the saliency map of a color image (obtained using HOG approach) is utilized as an additional feature. It is to be noted that saliency is computed only for RGB image and not depth map because the depth map lacks salient information and therefore, the saliency of depth map does not provide discriminating information.

2) Extracting Entropy Map and Visual Saliency Map: Let the input RGB-D image be denoted as $[I_{rgb}(x,y), I_d(x,y)]$, where $I_{rgb}(x,y)$ is the RGB image and $I_d(x,y)$ is the depth map, both of size $M \times N$. Let both of these be defined over the same set of (x,y) points such that $x \in [1,M]$ and $y \in [1,N]$. Let $H(I_j)$ denote the visual entropy map of image I_j. Here, I_j can be the depth map or the RGB image or a small part of these images. 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 7:

$$E_i = H(P_i),\ where\, i \in [1,4] \tag{7}$$

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

The proposed RISE algorithm also extracts visual saliency map S_1 of the color image I_{rgb} using Equation 8.

$$S_1(x,y) = S(I_{rgb}(x,y) \forall (x \in [1,M], y \in [1,N])) \tag{8}$$

3) Extracting Features using HOG: HOG [7] descriptor produces the histogram of a given image in which pixels are binned according to the magnitude and direction of their gradients. HOG has been successfully used as a feature and texture descriptor in many applications related to object detection, recognition, and other computer vision problems [6], [12], [31]. HOG of an entropy map or saliency map encodes the gradient direction and magnitude of the image variances in a fixed length feature vector. The information contained in the entropy/saliency map can therefore be represented compactly with a HOG histogram. Further, histogram based feature encoding enables non-rigid matching of the entropy/saliency characteristics which may not be possible otherwise.

In the proposed RISE algorithm, HOG is applied on the entropy and saliency maps. The entropy maps are extracted from patches P_i which allows capturing multiple granularities of the input image. Let $D(\cdot)$ denote the HOG histogram; the proposed algorithm computes HOG of entropy maps using the following equation:

$$F_i = D(E_i),\ where\, i \in [1,4] \tag{9}$$

Here, F_1 represents the HOG of entropy map E_1 defined over patch P_1 and F_2 represents the HOG of entropy map E_2 defined over patch P_2 of I_{rgb}. Similarly, F_3 and F_4 represent the HOG of entropy maps E_3 and E_4 defined over patches P_3 and P_4 of I_d respectively. F_1 and F_2 capture traditional texture information but instead of directly using visual information, 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.

Next, HOG descriptor of visual saliency map, S_1 is computed using Equation 10. The final descriptor F is created using an ordered concatenation of the five HOG histograms as shown in Equation 11.

$$F_5 = D(S_1(I_{rgb})) \tag{10}$$

$$F = [F_1, F_2, F_3, F_4, F_5] \tag{11}$$

Concatenation is used to facilitate training by reducing five vectors to a single feature vector. Since each HOG vector is small, the resulting concatenated vector has a small size which helps in reducing the computational requirement. The feature vector F is provided as input to a multi-class classifier.

4) Classification: To establish the identity of a given probe, a multi-class classifier such as Nearest Neighbor (NN), Random Decision Forests (RDFs) [16], and Support Vector Machines (SVM) can be used. 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 multi-class classification. RDFs are also robust towards outliers compared to the Nearest Neighbor algorithm, since every tree in the forest is only trained with a small subset of data. Therefore, the probability of an entire collection of trees making an incorrect decision due to a few outlier data points is very low. Moreover, as per the experimental results in the preliminary research, RDF is found to perform better than NN [13]. Other classifiers such as SVM require significantly more training data per class. Therefore, in this research, RDF is used 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 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. A probe feature vector is input to the trained RDF which provides a probabilistic match score for each

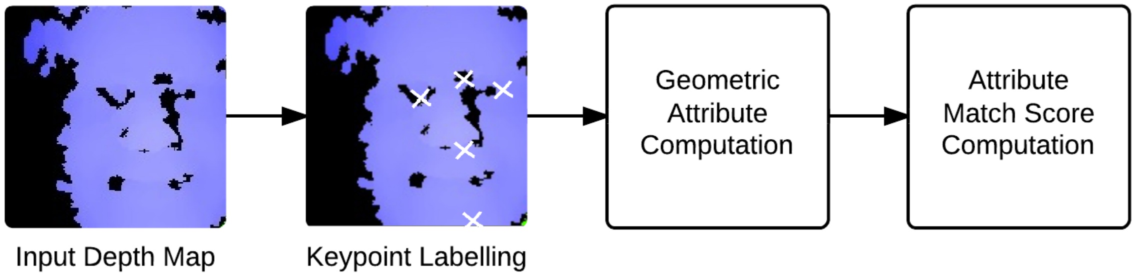

Fig. 5. Steps involved in the proposed ADM approach.

class. This match score denotes the probability with which the feature vector belongs to a particular class. To summarize, the RISE algorithm is presented in Algorithm 1.

Data: Preprocessed RGB-D image, I_{rgb}, denotes the color image and I_d denotes the depth map
Result: The RISE descriptor for the given RGB-D image F
for $i \leftarrow 1$ **to** 2 **do**
  E_i = Entropy map of patch P_i of $grayscale(I_{rgb})$;
end
for $i \leftarrow 3$ **to** 4 **do**
  E_i = Entropy map of patch P_i of I_d;
end
S = Saliency map of I_{rgb};
E_5 = Entropy map of S;
for $i \leftarrow 1$ **to** 5 **do**
  F_i = HOG of E_i;
  F = Concatenation of H_i;
end

Algorithm 1: The RISE algorithm

C. ADM: Attributes based on Depth Map

Attribute based methodologies have been applied successfully in image retrieval [21], [23] and face verification [22]. In RGB-D face recognition, it can be a useful additional feature. However, instead of qualitative or descriptive attributes such as gender, age, and complexion, the proposed **A**ttributes based on **D**epth **M**ap (ADM) algorithm extracts geometric attributes. Multiple geometric attributes can be utilized to describe a face such as the distances between various key facial features such as eyes, nose, and chin. By exploiting the uniform nature of a human face, key facial landmarks can be located and utilized to extract geometric attributes that can be used for face recognition in conjunction with the entropy and saliency features. An overview of the ADM approach is illustrated in Fig. 5. The ADM approach consists of the following steps.

1) Keypoint Labeling: To extract geometric attributes, first a few facial key points are located with the help of depth map. The points such as nose tip, eye sockets, and chin can be extracted by using a ”rule template”. In a detected face depth map, the nose tip is closest point from the sensor, the two eye sockets are always located above the nose tip and at a higher distance than their local surrounding regions (due to cheek bones and eyebrows being at a lesser distance), the chin can be detected as the closest point to the sensor below the nose tip. Utilizing these key points, some other landmarks such as the nose bridge and eyebrow coordinates can also be located. By using a standard set of landmarks for all faces, a consistent way to compute geometric measurements of the face is possible.

2) Geometric Attribute Computation: To obtain the geometric attributes, various distances between these landmark points are computed: inter-eye distance, eye to nose bridge distance, nose bridge to nose tip distance, nose tip to chin distance, nose bridge to chin distance, chin to eye distance, eyebrow length, nose tip distance to both ends of both eyebrows, and overall length of the face. Since the measured value of these parameters may vary across pose and expression, multiple gallery images are utilized to extract the facial features. Attributes are computed individually for each gallery image and the distances are averaged. In this manner, a consistent set of attributes is computed for a subject. These contribute towards the attribute feature vector for the RGB-D face image.

3) Attribute Match Score Computation: The attributes for a probe are computed similar to gallery images. Once the attributes are computed for a probe, the match score Φ is computed for each subject in the gallery using Equation 12.

$$\Phi = \sum_{i=1}^{N} w_i \times (A_i - a_i)^2 \tag{12}$$

Here, A_i and a_i are the i^{th} attributes of the probe image and the gallery image respectively. w_i is the weight of the i^{th} attribute and N is the total number of attributes. w_i is used to assign different weights to different attributes depending upon how reliably they can be computed. In this research, w_i is optimized using grid search for efficient identification performance on the training dataset. After computation, the match scores from each subject can be utilized for identification. However, in the proposed approach it is combined with the match score obtained by RISE algorithm for taking the final decision.

D. Combining RISE and ADM

The match scores obtained by RISE and ADM algorithms can be combined in various ways. In this research, we explore two types of fusion:

1) Match Score Level Fusion: Match score level fusion is performed using the weighted sum rule [27]. Let Φ_{RISE} be the match score obtained using the RISE approach and Φ_{ADM} be the match score obtained by the ADM approach. The fused match score Φ_{fused} is computed as,

$$\Phi_{final} = w_{RISE} \times \Phi_{RISE} + w_{ADM} \times \Phi_{ADM} \qquad (13)$$

where w_{RISE} and w_{ADM} are the weights assigned to the RISE and ADM match scores respectively.

2) Rank Level Fusion: Rank level fusion is performed using Weighted Borda Count approach [27]. Weighted Borda count allocates a score to a subject depending on its rank in both the ranked lists and then creates a new ranked list for identification based on these scores. The ranked list of subjects is created using both RISE and ADM match scores individually. These ranked lists are then combined by computing a new match score for each subject based on these ranked lists according to Equation 14.

$$Rf_{subj} = \sum_{i=RISE,ADM} \sum_{j=1}^{R_{max}} \begin{cases} w_i(R_{max} - j) & if\ R_{ij} = rank(subj) \\ 0 & otherwise \end{cases} \qquad (14)$$

Here, R_{max} denotes the maximum (worst) possible rank value. w_{RISE} and w_{ADM} denote the weights for RISE and ADM respectively. Similarly, R_{RISE} and R_{ADM} denote the ranked lists of RISE and ADM respectively. The weights w_{RISE} and w_{ADM} can be used to control the number of points that the ranked lists of RISE and ADM can provide to the subject. $R_{ij} = rank(subj)$ signifies the condition that the $subject$ has rank j in the i^{th} ranked list.

III. Experimental Results

The performance of the proposed approach is analyzed via two types of experiments. First, the experiments are conducted on the IIIT-D RGB-D dataset to analyze the performance of the proposed approach with various combinations of constituent components and their parameters. Thereafter, the performance is compared with existing 2D and 3D approaches on an extended dataset.

A. Database and Experimental Protocol

There are a few existing RGB-D databases in literature. The EURECOM [18] database has 936 images pertaining to 52 subjects and the images are captured in two sessions with variations in pose, illumination, view, and occlusion. The VAP RGB-D [15] face database contains 153 images pertaining to 31 individuals. The dataset has 51 images for each individual with variations in pose. However, both of these datasets contain images pertaining to a relatively small number of individuals. To evaluate the performance of face recognition, a larger dataset is preferable. Therefore, the IIIT-D RGB-D database[1] is prepared for the experiments. This database contains 4605 RGB-D images pertaining to 106 individuals captured in two sessions using Kinect sensor and OpenNI SDK. The resolution of both the color image and the depth map is 640×480. The number of images per individual is variable with a minimum of 11 images and a maximum of 254 images. In this database, the images are captured in normal illumination with variations in pose and expression (in some cases, there are variations due to eye-glasses as well). Some sample images for a subject in the IIIT-D database are presented in Fig. 6. Using these three datasets, two types of experiments are performed. The initial experiments are performed on the IIIT-D RGB-D dataset to analyze the component-wise performance of the proposed RISE approach as well as to study the impact of weights and gallery size on the identification performance. Thereafter, the IIIT-D RGB-D dataset is merged with the EURECOM [18] and VAP [15] datasets to create an extended dataset of 189 individuals. The extended dataset is used to compare the performance of the proposed algorithm with existing 2D and 3D approaches.

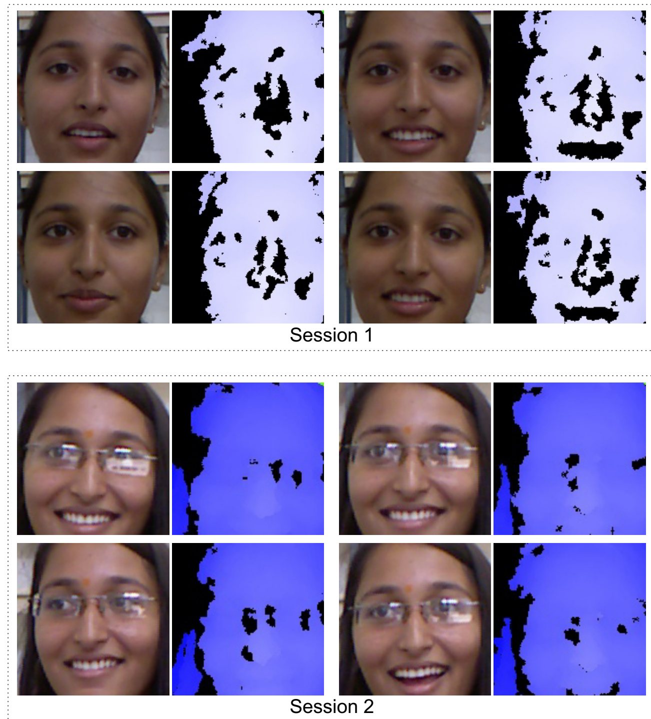

Fig. 6. Sample images of a subject in two sessions from the IIIT-D RGB-D database.

The experimental protocol for each experiment is detailed below:

- **Experiment 1:** 40% of the IIIT-D Kinect RGB-D database is used for training and validation. The training dataset is utilized to compute the weights involved in ADM approach, RDF classifier parameters, and weights

[1]The database and ground truth information is available via https://research.iiitd.edu.in/groups/iab/rgbd.html

TABLE I
EXPERIMENTAL PROTOCOL FOR BOTH INITIAL AND EXTENDED EXPERIMENTS.

Experiment	Database	No. of Images	No. of Subjects	
			Training	Testing
Experiment 1	IIIT-D RGB-D	4605	42	64
Experiment 2	IIIT-D RGB-D + VAP + EURECOM	5694	75	114

for fusion. Note that RDF classifier is separately trained for the initial and extended experiments by utilizing the respective training datasets. After training and parameter optimization, the remaining 60% dataset (unseen subjects) is used for testing. The results are computed with five times random subsampling. In each iteration of the subsampling, the subjects chosen for training/testing as well as the gallery images selected for each subject are randomly selected. Gallery size is fixed at four images per subject.

- **Experiment 2:** The extended database of 189 subjects is used for this experiment. Images pertaining to 40% individuals from the extended database are used for training and the remaining 60% unseen subjects are used for testing. To create the complete subject list for the extended dataset, the subjects are randomly subsampled within the three datasets according to 40/60 partitioning and then merged together to form one extended training/testing partition. Therefore, the extended training dataset has proportionate (40%) representation from each of the three datasets. The number of images available per individual varies across the three datasets and therefore the gallery size for the extended dataset experiment is fixed at two gallery images per individual. The remaining images of the subject are used as probe.

Cumulative Match Characteristics (CMC) curves are computed for each experiment and the average accuracy values are presented along with standard deviations across random subsamples. The experimental protocol for all the experiments are summarized in Table I.

The performance of the proposed algorithm is compared with several existing algorithms namely: Four Patch Local Binary Patterns (FPLBP) [33], Pyramid Histogram of Oriented Gradients (PHOG) [1], Scale Invariant Feature Transform (SIFT) [25], and Sparse representation [34]. Besides these methods which utilize only 2D information, a comparison is also performed with 3D-PCA based algorithm [5] which computes a subspace based on depth and grayscale information.

B. Results and Analysis: Experiment 1

Component-wise analysis: As discussed in Section II, the proposed RISE 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:

- **Case (a)** RGB-D and saliency without entropy: RGB image and depth map are used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_3, F_4, F_5]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 4]$.
- **Case (b)** RGB only: Only the RGB image is used to extract entropy and saliency features, i.e., $F = [F_1, F_2, F_5]$
- **Case (c)** RGB-D only: Only the entropy maps are used, saliency is not used, i.e., $F = [F_1, F_2, F_3, F_4]$
- **Case (d)** RGB and saliency without entropy: RGB information is used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_5]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 2]$.
- **Case (e)** RGB-D only without entropy: RGB-D information is used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_3, F_4]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 4]$.
- **Case (f)** RGB only without saliency: $F = [F_1, F_2]$

These cases analyze the effect of different components of the proposed algorithm 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 potentially be drawn from the results of other five cases. Comparing the performance of the proposed descriptor with entropy, saliency and depth information can also determine whether the proposed combination of components improves the face recognition performance with respect to the individual components.

The results of individual experiments are shown in Fig. 7. It is observed that removing any of the components significantly reduces the performance of the proposed algorithm. For example, the CMC curve corresponding to case (c) shows that the contribution of including visual saliency map as an added feature is important. It is observed that saliency is relevant towards stabilizing the feature descriptor and preserving intra-class similarities. Further, in cases (d) and (e), it is observed 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 raw form. Overall, the proposed algorithm yields 95% rank 5 accuracy on IIIT-D database. Further, Table II shows the comparison of the proposed algorithm with existing algorithms. The results indicate that, on the IIIT-D database, the proposed algorithm is about 8% better than the second best algorithm (in this case, Sparse representation [34]). Compared with 3D-PCA algorithm, the proposed algorithm is able to yield about 12% improvement.

Fusion of algorithms: Experiments are performed with various combinations of the proposed RISE and ADM approaches as well as 3D-PCA [5]. In order to fuse 3D-PCA with RISE and ADM, both weighted sum rule and weighted Borda count can be utilized. The results of this experiment are presented in Fig. 8. W.B.C. refers to rank level fusion using Weighted Borda Count and W.S. refers to match score level fusion using Weighted Sum rule. The key analysis are explained below:

- The proposed RISE + ADM with weighted sum rule

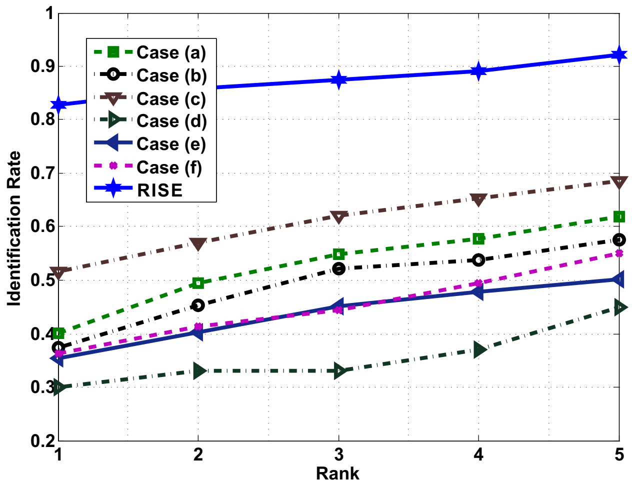

Fig. 7. Analyzing the proposed RISE algorithm and its individual components on the IIIT-D RGB-D face database.

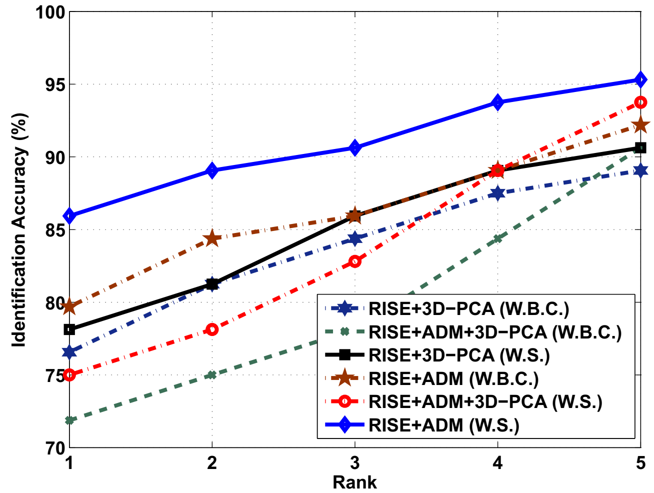

Fig. 8. Analyzing the performance of different combinations of the proposed algorithm with 3D PCA and fusion algorithms on the IIIT-D RGB-D face database.

yields the best rank 5 identification accuracy of 95.3%. RISE+ADM approach using weighted Borda count also performs well providing an accuracy of 79.7% which is better than the remaining combinations at rank 1.

- Even though RISE+ADM+3D-PCA performs second best with rank 5 identification accuracy of 93.7%, the difference in performance at rank 1 is 10.9% lower than RISE+ADM (W.S.) and the use of 3D-PCA also adds to the computational complexity.
- The weighted sum variants of the combinations perform consistently better than their weighted Borda count variants. This indicates that match score level fusion performs better than rank level fusion. However, it is also notable that the difference in performance for all approaches reduces at rank 5 compared to rank 1. This implies that any advantage gained by utilizing one approach over the other diminishes at higher ranks as the criteria for successful identification is relaxed.

Since weights are involved in both weighted Borda count and weighted sum approaches, it is interesting to observe how the performance of the proposed algorithm varies with the variation in weights. The results of this experiment are presented in Figs. 9 and 10 for weighted sum rule and weighted Borda count respectively. The number in parenthesis after the algorithm indicates their weight in the approach. For example, RISE (0.5) + ADM (0.5) implies that both RISE and ADM are allocated equal weights. Based on these results, the following analysis can be performed:

- The best performance is achieved with RISE (0.7) + ADM (0.3) for both the fusion algorithms. This indicates that texture features extracted by RISE are more informative for identification and therefore must be assigned higher weight. However, the geometric features from ADM also contribute towards the identification performance after fusion, thereby increasing the rank 5 accuracy from 92.2% (RISE only) to 95.3% (RISE + ADM) when weighted sum rule is utilized.
- The performance of weighted Borda count is lower than weighted sum possibly because of the loss of information

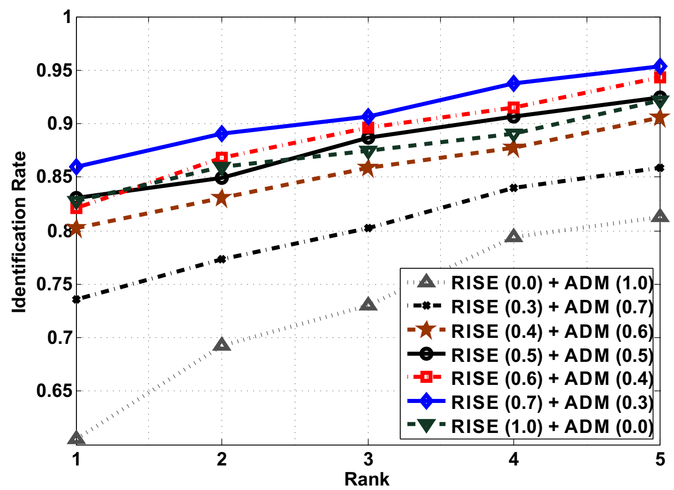

Fig. 9. Analyzing the effect of weights in match score level fusion using weighted sum rule on the IIIT-D RGB-D face database.

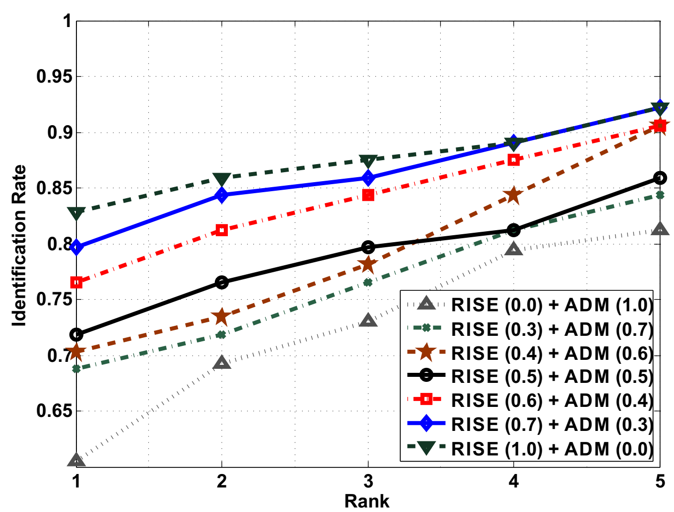

Fig. 10. Analyzing the effect of weights in rank level fusion using weighted Borda count on the IIIT-D RGB-D face database.

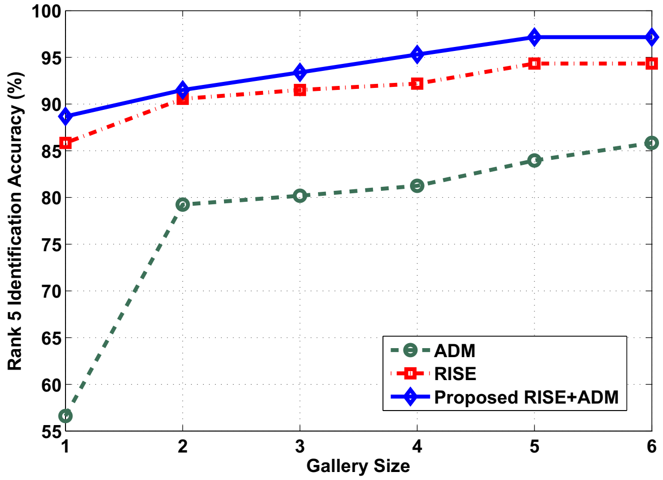

Fig. 11. Analyzing the effect of gallery size on the identification performance on the IIIT-D RGB-D face database.

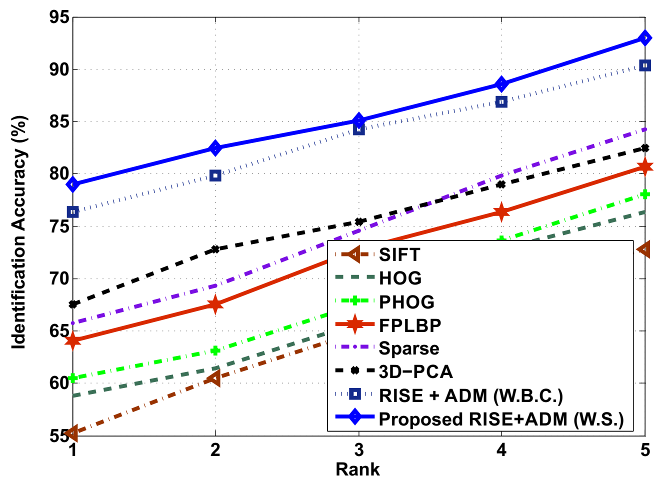

Fig. 12. Comparing the performance of the proposed approach with existing 2D and 3D approaches on the extended database.

that occurs in using the ranked list for fusion instead of the match scores.

- Experiments have been conducted to assess the performance with all other combinations of weights as well, but none of these combinations perform better than RISE (0.7) + ADM (0.3).

Analysis with gallery size: All the experiments described above on the IIIT-D RGB-D database are performed with a gallery size of four. To analyze the effect of gallery size on the identification performance, additional experiments are performed by varying the number of images in the gallery. The results of this experiment are presented in Fig. 11 and the analysis is presented below.

- The curve indicates that the performance of RISE, ADM and the proposed RISE+ADM approach increases with increase in gallery size. However, the maximum increment is observed from gallery size 1 to gallery size 2 in the ADM approach. This major performance increment of 22.6% can be credited to the possibility that using only single gallery image yields approximate geometric attributes. As soon as more than one sample becomes available, the averaging process increases the reliability of the geometric attributes and hence there is a significant increase in performance.
- With the above discussed exception, the performance of each approach increases consistently but in small amounts with increase in gallery size. Therefore, after a certain point, increasing gallery size does not provide high returns in terms of the performance. It is notable that even with single gallery image, the proposed algorithm yields the rank 5 identification accuracy of 89%.

TABLE II
IDENTIFICATION ACCURACY (%) FOR THE RECOGNITION EXPERIMENTS ON IIIT-D RGB-D FACE DATABASE AND EURECOM DATABASE INDIVIDUALLY. THE MEAN ACCURACY VALUES ARE REPORTED ALONG WITH STANDARD DEVIATION.

Modality	Descriptor	Rank 5 Identification Accuracy (%)	
		IIIT-D RGB-D	EURECOM
2D	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	3D-PCA	83.4±2.1	94.1±2.7
	RISE + ADM	**95.3±1.7**	**98.5±1.6**

Assessing the accuracy of ADM keypoint labeling: The performance of ADM approach is dependent on the keypoint labeling phase. In order to determine the accuracy of this phase, manual keypoint labels are collected via crowdsourcing. Human volunteers are requested to label the keypoints (nose, left eye, right eye and chin) on 10 images of every subject. The average of human-annotated keypoint coordinates is computed and compared with the automatically obtained keypoints. An automatic keypoint is considered to be correct if it lies within a small local neighborhood of the average human-annotated keypoint. It is observed that the overall accuracy of automated keypoint labeling, using manual annotations as ground truth, on the IIIT-D Kinect RGB-D database is 90.1% with a 5×5 neighborhood and 93.6% with a neighborhood size of 7×7. Based on the performance of ADM on individual frames, it can be noted that it performs the best on near frontal frames and semi-frontal frames.

Performance on EURECOM: Performance of the proposed algorithm is also compared with existing algorithms on the EURECOM dataset. In order to perform this recognition experiment, the gallery sizes for the EURECOM dataset is fixed at 2 images per subject. The results of this experiment are presented in Table II. The analysis is similar to the IIIT-D database and the proposed algorithm yields an accuracy of 98.5% rank-5 identification accuracy which is around 4% better than existing algorithms. Note that the EURECOM database is relatively smaller than IIIT-D database and therefore, near perfect rank 5 accuracy is achieved.

C. Results and Analysis: Experiment 2

The proposed RISE + ADM approach is compared with some existing 2D and 3D approaches on the extended dataset

TABLE III
IDENTIFICATION ACCURACY (%) FOR THE EXTENDED EXPERIMENTS. THE MEAN ACCURACY VALUES ARE REPORTED ALONG WITH STANDARD DEVIATION.

Modality	Descriptor	Rank 1	Rank 5
2D	SIFT	55.3 ± 1.7	72.8 ± 2.1
	HOG	58.8 ± 1.4	76.3 ± 1.8
	PHOG	60.5 ± 1.6	78.1 ± 1.1
	FPLBP	64.0 ± 1.1	80.7 ± 2.0
	Sparse	65.8 ± 0.6	84.2 ± 0.8
3D	3D-PCA	67.5 ± 1.2	82.5 ± 1.9
	RISE+ADM (W.B.C.)	76.3 ± 1.0	90.3 ± 1.1
	RISE+ADM (W.S.)	**78.9 ± 1.7**	**92.9 ± 1.3**

TABLE IV
A DETAILED COMPARATIVE ANALYSIS OF THE PROPOSED ALGORITHM WITH 3D-PCA, FPLBP, AND SPARSE APPROACHES. T AND F REPRESENT TRUE AND FALSE RESPECTIVELY. TRUE GROUND TRUTH REFERS TO GENUINE CASES AND FALSE GROUND TRUTH REFERS TO THE IMPOSTOR CASES.

Algorithm Results	Ground Truth	
	True	False
3D-PCA=T, Proposed=T	61.9%	5.3%
3D-PCA=F, Proposed=T	21.3%	5.4%
3D-PCA=T, Proposed=F	10.0%	24.3%
3D-PCA=F, Proposed=F	6.8%	65.0%
FPLBP=T, Proposed=T	61.8%	6.8%
FPLBP=F, Proposed=T	27.6%	3.4%
FPLBP=T, Proposed=F	6.3%	25.3%
FPLBP=F, Proposed=F	4.3%	64.5%
Sparse=T, Proposed=T	68.6%	3.2%
Sparse=F, Proposed=T	18.7%	11.4%
Sparse=T, Proposed=F	8.0%	26.0%
Sparse=F, Proposed=F	4.7%	59.4%

(Experiment 2). The identification performance of these approaches is presented in Fig. 12 and summarized in Table III. The results indicate that the proposed RISE+ADM algorithm (both weighted sum and weighted Borda count versions) outperforms the existing approaches by a difference of around 8% in terms of the rank 5 identification performance. The proposed algorithm yields the best results at rank 1 with an accuracy of 78.9% which is at least 11.4% better than second best algorithm, 3D-PCA.

Detailed comparison with other algorithms: In order to compare the performance of the proposed algorithm with other top performing algorithms, a comparative study is performed. The details of this study are presented in Table IV. As is evident from the results presented, the proposed algorithm is able to correctly determine ground truth in the case of a wrong decision by another algorithm more often than the reverse case, i.e., when another algorithm is correct and the proposed algorithm is incorrect. For example, the percentage of impostor cases when 3D-PCA is incorrect and the proposed algorithm is correct is 24.30% whereas the percentage of impostor cases where the proposed algorithm is incorrect and 3D-PCA is correct is only 5.38%.

In order to further analyze the performance, we examine two types of results. Fig. 13 contains two samples 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

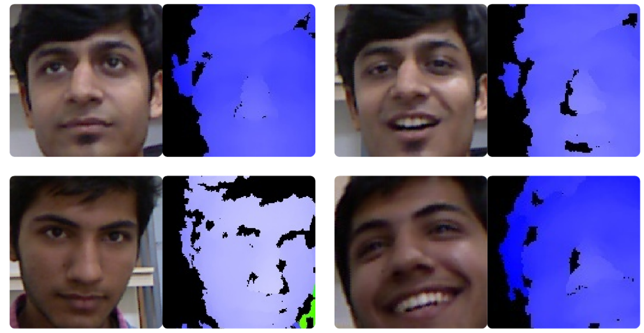

Fig. 13. Analyzing the performance of the proposed algorithm. The first row (Case 1) presents sample gallery and probe images when all the algorithms are able to recognize. The second row (Case 2) presents example gallery and probe images when only the proposed algorithm is able to correctly identify the subject at rank-1.

to identify the subject and other algorithms provide incorrect results. 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 recognize even when there are high pose and expression variations. Thus, it can be concluded that the proposed descriptor outperforms these existing 2D and 3D approaches. In summary, this difference in performance can be attributed to the following reasons:

- The RISE descriptor uses depth information in addition to traditional color information which allows it to utilize additional sources for feature extraction. After characterization by local entropy, the depth map is able to mitigate the effect of illumination and expression. The geometrical attributes obtained from the ADM approach further contribute towards resilient identification.
- The proposed descriptor utilizes saliency map for feature extraction to model visual attention. The saliency distribution of a face is not significantly affected by pose variations and therefore it provides tolerance to minor pose variations.
- Compared to existing approaches, entropy and saliency maps of RGB-D images are not highly affected by noise such as holes in depth map and low resolution, and therefore, yield higher performance. The additional geometric attributes are another source of noise tolerant features as they are averaged across multiple gallery images.

D. Experiments on 3D TEC dataset

In order to evaluate the performance of the proposed RGB-D recognition algorithm on other 3D databases, identification results are also presented on the 3D-Twins Expression Challenge (3D TEC) dataset [30]. The database contains images pertaining to 107 pairs of twins acquired using a Minolta VIVID 910 3D scanner in controlled illumination and background. The range and texture images are of 480 × 640 resolution. The dataset provides four sets for performing identification experiments between two twins, A and B. Each set defines

TABLE V
RANK-ONE IDENTIFICATION ACCURACIES ON THE 3D TEC [30] DATASET. THE RESULTS OF OTHER ALGORITHMS ARE PRESENTED AS REPORTED IN [30]. THE PROPOSED ALGORITHM ACHIEVES CLOSE TO STATE-OF-THE-ART PERFORMANCE.

Algorithm	**Rank 1 Identification Accuracy**			
	I	**II**	**III**	**IV**
Alg. 1 (E_{pkn})	93.5%	93.0%	72.0%	72.4%
Alg. 1 (E_{minmax})	94.4%	93.5%	72.4%	72.9%
Alg. 2 (SI)	92.1%	93.0%	83.2%	83.2%
Alg. 2 (eLBP)	91.1%	93.5%	77.1%	78.5%
Alg. 2 (Range PFI)	91.6%	93.9%	68.7%	71.0%
Alg. 2 (Text, PFI)	95.8%	96.3%	91.6%	92.1%
Alg. 3	62.6%	63.6%	54.2%	59.4%
Alg. 4	98.1%	98.1%	91.6%	93.5%
Proposed	95.8%	94.3%	90.1%	92.4%

the gallery and probe images for each twin according to the expression variations (smile or neutral). Further details of these sets are provided in [30].

Along with the proposed algorithm, we also compare the results with four existing algorithms that participated in the Twin Expression Challenge, 2012. The existing algorithms, Alg. 1 to Alg. 4, are designed to utilize rich 3D maps and/or texture information captured using telephoto lens equipped Minolta scanner. The details of these algorithms and their results are available in [30]. Table V presents the results of the proposed and four existing algorithms on the 3D TEC dataset. As shown in Table V, even though the proposed algorithm does not fully utilize rich depth maps, it achieves the second best performance on two of the four sets and is able to yield close to state-of-the-art performance with more than 90% rank 1 accuracy on all four sets.

IV. CONCLUSION

Existing face recognition algorithms generally utilize 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. This research proposes a novel RISE 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. Further, the ADM algorithm is proposed to extract and match geometric attributes. ADM is then combined with the RISE algorithm for identification. The experiments performed on the RGB-D databases demonstrate the effectiveness of the proposed algorithm and show that it performs better than some existing 2D and 3D approaches of face recognition. Future research can be directed towards incorporating depth data in video face recognition [4].

APPENDIX

The performance of the proposed algorithm is also compared with a commercial system (3D-COTS)[2]. COTS employs a 3D model reconstruction for each subject using the gallery

[2]Name of the commercial system is suppressed due to the constraints in the license agreement.

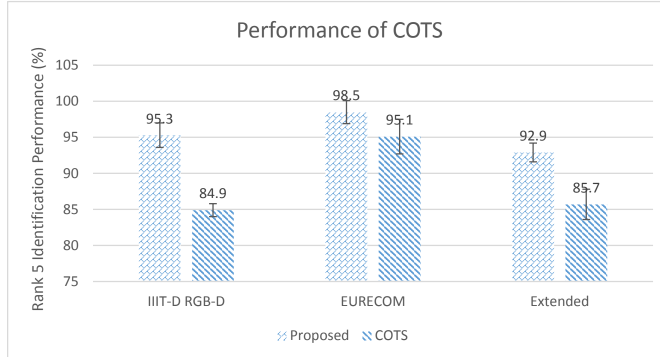

Fig. 14. Comparing the identification performance of the proposed algorithm with COTS on all three databases.

RGB images. RGB probe image is also converted to 3D model for matching. The details of reconstruction algorithm are not available. Fig. 14 presents a comparison of identification performance between COTS and the proposed algorithm. It is evident that the proposed algorithm is able to consistently achieve better performance. The failure of COTS can be attributed to the 3D reconstruction method which possibly suffers from low spatial resolution of RGB images.

ACKNOWLEDGEMENT

The authors would like to acknowledge Dr. Jean-Luc Dugelay for providing the EURECOM database, Dr. Thomas B. Moeslund for the VAP database, CVRL, University of Notre Dame for providing the 3D TEC dataset, and Samarth Bharadwaj for his help in the IIIT-D Kinect RGB-D database collection. The research is partly supported through a grant from DIT, Government of India. The authors also acknowledge reviewers and associate editor for helpful and constructive feedback on the paper.

REFERENCES

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

[2] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Recognizing surgically altered face images using multiobjective evolutionary algorithm. *IEEE Transactions on Information Forensics and Security*, 8(1):89–100, 2013.

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

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

[6] E. Corvee and F. Bremond. Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *Advanced Video and Signal-Based Surveillance*, pages 469–475, 2010.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

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

[9] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar. Recognizing disguised faces: Human and machine evaluation. *PLoS ONE*, 9(7):e99212, 2014.

[10] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997.

[11] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard. Real-time 3D visual slam with a hand-held RGB-D camera. In *RGB-D Workshop, European Robotics Forum*, 2011.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[13] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On RGB-D face recognition using kinect. In *Biometrics: Theory, Applications and Systems*, 2013.

[14] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In *International Symposium on Experimental Robotics*, volume 20, pages 22–25, 2010.

[15] 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 *International Conference on Signal Image Technology and Internet Based Systems*, pages 42–46, 2012.

[16] T. K. Ho. Random decision forests. In *International Conference on Document Analysis and Recognition*, pages 278–282, 1995.

[17] D. Holz, S. Holzer, R. Rusu, and S. Behnke. Real-time plane segmentation using RGB-D cameras. *RoboCup 2011*, pages 306–317, 2012.

[18] 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 *Asian Conference on Computer Vision*, 2012.

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254 –1259, 1998.

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

[21] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.

[22] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *International Conference on Computer Vision*, 2009.

[23] D. Kun, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition*, pages 3474–3481, 2012.

[24] 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 *Workshop on the Applications of Computer Vision*, pages 186–192, 2013.

[25] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[26] 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 *Human-Robot Interaction*, pages 229–230, 2011.

[27] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of multibiometrics*, volume 6. Springer Science and Business Media, 2006.

[28] A. Rrnyi. On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.

[29] M. P. Segundo, S. Sarkar, D. Goldgof, and L. S. O. Bellon. Continuous 3d face authentication using rgb-d cameras. In *Computer Vision and Pattern Recognition Biometrics Workshop*, 2013.

[30] V. Vijayan, K. W. Bowyer, P. J. Flynn, D. Huang, L. Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris. Twins 3D face recognition challenge. In *International Joint Conference on Biometrics*, pages 1–7, 2011.

[31] C. Wang and L. Guan. Graph cut video object segmentation using histogram of oriented gradients. In *International Symposium on Circuits and Systems*, pages 2590–2593, 2008.

[32] S.-J. Wang, J. Yang, N. Zhang, and C.-G. Zhou. Tensor discriminant color space for face recognition. *IEEE Transactions on Image Processing*, 20(9):2490–2501, 2011.

[33] L. Wolf, T. Hassner, Y. Taigman, et al. Descriptor based methods in the wild. In *European Conference on Computer Vision Real Faces Workshop*, 2008.

[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[35] J. Yang and C. Liu. Color image discriminant models and algorithms for face recognition. *IEEE Transactions on Neural Networks*, 19(12):2088–2098, 2008.

Gaurav Goswami (S'12) received his Bachelor in Technology degree in Information Technology in 2012 from the Indraprastha Institute of Information Technology (IIIT) Delhi, India where he is currently pursuing a PhD. His main areas of interest are image processing, computer vision and their application in biometrics. He is the recipient of best poster award in BTAS 2013.

Mayank Vatsa (S'04-M'09, SM'14) received the M.S. and Ph.D. degrees in computer science in 2005 and 2008, respectively from the West Virginia University, Morgantown, USA. He is currently an Assistant Professor at the Indraprastha Institute of Information Technology (IIIT) Delhi, India. He has more than 125 publications in refereed journals, book chapters, and conferences. His research has been funded by the UIDAI and DIT. He is the recipient of FAST award by DST, India. His areas of interest are biometrics, image processing, computer vision, and information fusion. Dr. Vatsa is a member of the IEEE, Computer Society and Association for Computing Machinery. He is the recipient of several best paper and best poster awards in international conferences. He is also an area editor of IEEE Biometric Compendium, area chair of Information Fusion, Elsevier, and PC Co-Chair of ICB-2013 and IJCB-2014.

Richa Singh (S'04-M'09-SM'14) received the M.S. and Ph.D. degrees in computer science in 2005 and 2008, respectively from the West Virginia University, Morgantown, USA. She is currently an Assistant Professor at the Indraprastha Institute of Information Technology (IIIT) Delhi, India. Her research has been funded by the UIDAI and DIT, India. She is a recipient of FAST award by DST, India. Her areas of interest are biometrics, pattern recognition, and machine learning. She has more than 125 publications in refereed journals, book chapters, and conferences. She is also an editorial board member of Information Fusion, Elsevier and EURASIP Journal of Image and Video Processing, Springer. Dr. Singh is a member of the CDEFFS, IEEE, Computer Society and the Association for Computing Machinery. She is the recipient of several best paper and best poster awards in international conferences.