Neurocomputing 214 (2016) 93–108

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Robust RGB-D face recognition using Kinect sensor ☆

Billy Y.L. Li [a], Mingliang Xue [b,*], Ajmal Mian [c], Wanquan Liu [a], Aneesh Krishna [a]

[a] Department of Computing, Curtin University, Kent Street, Perth, WA 6102, Australia
[b] Dalian Key Lab of Digital Technology for National Culture, Dalian Minzu University, Dalian 116600, Liaoning, China
[c] Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
16 March 2016
Accepted 3 June 2016
Communicated by Zidong Wang
Available online 17 June 2016

Keywords:
3D face recognition
Kinect
Multi-channel discriminant transform
Sparse coding

ABSTRACT

In this paper we propose a robust face recognition algorithm for low resolution RGB-D Kinect data. Many techniques are proposed for image preprocessing due to the noisy depth data. First, facial symmetry is exploited based on the 3D point cloud to obtain a canonical frontal view image irrespective of the initial pose and then depth data is converted to *XYZ* normal maps. Secondly, multi-channel Discriminant Transforms are then used to project RGB to DCS (Discriminant Color Space) and normal maps to DNM (Discriminant Normal Maps). Finally, a Multi-channel Robust Sparse Coding method is proposed that codes the multiple channels (DCS or DNM) of a test image as a sparse combination of training samples with different pixel weighting. Weights are calculated dynamically in an iterative process to achieve robustness against variations in pose, illumination, facial expressions and disguise. In contrast to existing techniques, our multi-channel approach is more robust to variations. Reconstruction errors of the test image (DCS and DNM) are normalized and fused to decide its identity. The proposed algorithm is evaluated on four public databases. It achieves 98.4% identification rate on CurtinFaces, a Kinect database with 4784 RGB-D images of 52 subjects. Using a first versus all protocol on the Bosphorus, CASIA and FRGC v2 databases, the proposed algorithm achieves 97.6%, 95.6% and 95.2% identification rates respectively. To the best of our knowledge, these are the highest identification rates reported so far for the first three databases.

1. Introduction

Three-dimensional face recognition has attracted significant research interest in the past decade due to its broad applications. Face recognition can be performed in a non-intrusive way and sometimes without the user's knowledge or explicit cooperation. However, facial images captured in an uncontrolled environment can have combinations of variations such as pose, facial expressions, illumination and disguise. Since the type of variations is unknown for a given image, it becomes critical to design a face recognition algorithm that can handle all these factors simultaneously.

Dealing with multiple types of variations simultaneously is a challenging task for face recognition. Traditional approaches have tried to tackle one challenge at a time using optical 2D images or texture. For example, the illumination cone method [1] models illumination which changes linearly. The authors prove that the images of a face under the same pose but different illuminations lie on a low dimensional convex cone which can be learned from a few training images. Although this technique can be used to generate facial images under novel illuminations, it assumes that faces are convex and requires training images to be taken with a point light source. The Spare Representation Classifier (SRC) [2] and its extension, the Robust Sparse Coding (RSC) [3], can handle face images with disguise (e.g. wearing sunglasses) and noise, by removing or correcting the outlier pixels. However, some outlier pixels may have similar texture intensity to the human face and thus cannot be identified. Some researchers have also tried to solve the pose problem using 2D images. For example, Gross et al. [4] construct the Eigen-light fields which are the 2D appearance models of a face from all viewpoints. This method requires many training images under different poses and dense correspondences between them which are difficult to achieve. Sharma and Jacobs [5] use Partial Least Squares (PLS) to linearly map facial images with different poses to a common linear subspace where they are highly correlated. However, such a linear subspace may not exist. In fact, pose variations are highly non-linear and cannot be

☆This work was partially supported by the Australian Research Council (ARC) Discovery Grant DP110102399. Portion of the research in this paper use the CASIA-3D FaceV1 collected by the Chinese Academy of Sciences' Institute of Automation (CASIA).

* Corresponding author.
E-mail addresses: y.li2@postgrad.curtin.edu.au (B.Y.L. Li), mingliangxue@gmail.com (M. Xue), ajmal@csse.uwa.edu.au (A. Mian), w.liu@curtin.edu.au (W. Liu), a.krishna@curtin.edu.au (A. Krishna).

http://dx.doi.org/10.1016/j.neucom.2016.06.012
0925-2312/

Table 1
Various 3D data acquisition devices.

Device	Speed (s)	Charge time	Size (in^3)	Price (USD)	Acc. (mm)
3dMD	0.002	10 s	423.9	>$50k	<0.2
Minolta	2.5	No	1408	>$50k	~ 0.1
Artec	0.063	No	160.8	>$20k	~0.5
BLITZ	0.9	No	n/a	>$14k	~0.2
3D3-R1	1.3	No	n/a	>$10k	>0.3
SR4000	0.02	No	17.53	>$5k	~10
David	2.4	No	n/a	>$2k	~0.5
Kinect	0.033	No	41.25	<$200	>1.5

modelled by linear methods. This is why the performance of the above methods drops dramatically with extreme pose variations.

Limitations of 2D face recognition, especially sensitivity to pose and illumination variations, can be overcome by using 3D face data. Facial geometry is invariant to illumination whereas 2D images are a direct function of the lighting conditions (direction and spectrum). Though the 3D imaging process can be influenced by lighting, the 3D data itself is illumination invariant. Furthermore, 3D face model can be used to render facial images under different illumination conditions [6], and to correct the facial pose or to generate infinite novel poses. Although existing methods of 3D face recognition [7–12] can achieve very high accuracy even under challenging experiments such as the Face Recognition Grand Challenge (FRGC) [13], they all assume the availability of high resolution 3D face scanners. Such scanners are costly, bulky and have slow acquisition speed, which limit their applications.

1.1. Commercial 3D acquisition devices

Table 1 compares the properties and price of several three-dimensional acquisition devices, which can be purchased off the shelf. The 3dMD scanner is designed specifically for instant 3D face acquisition, which has been used for medical applications. It can capture a single scan in 2 ms, but requires 10 s to charge prior to every scan. In contrast, the Minolta scanner, which was used to acquire 3D face data for the well-known FRGC experiments [13], does not need to charge. However, it takes over 2 s for one scan. The long scanning time will cause misalignment problems if the subject moves, since 3D data and texture are acquired sequentially by Minolta.

SwissRanger SR4000 is a time of flight 3D scanner, which has faster capture speed when compared to the Kinect Sensor. However, its cost is relatively higher and its accuracy is much lower than Kinect at medium range. In fact, most devices that have a depth accuracy lower than 0.5 mm will require more than 1 s to acquire one 3D sample. Consequently, subjects must keep still in front of the sensor for the duration of scanning which implicitly means that the subjects are cooperative. Therefore, the advantage of non-intrusiveness for face recognition is compromised. Although low cost and high speed acquisition devices are available in the market such as the Kinect sensor, they usually have low depth resolution and accuracy.

1.2. Challenges of Kinect data

It can be seen from Table 1 that the Kinect sensor is low cost, compact in size and has a very fast acquisition speed. These properties are more appealing for non-intrusive face recognition applications. Technically, Kinect integrates a standard RGB camera, an infra-red projector and an infra-red camera. The infra-red projector projects a static infra-red pattern on the scene (face in our case) which is sensed by the infra-red camera. This pattern is used to resolve correspondence between the projector and the camera, and depth is calculated using stereopsis [15]. Kinect can provide RGB-D (Red, Green, Blue, Depth) data of 640×480 spatial resolution at 30 frames per second. Although this spatial resolution is similar to other 3D scanners, the depth resolution and accuracy of Kinect decreases dramatically from 1.5 mm to 50 mm as the object moves further away form the camera. Since we use Kinect at a distance of 1 m, we are operating at a depth resolution of about 1.5 mm. However, this resolution is still quite low and leads to very noisy 3D data. Sample data acquired by Kinect is shown in Fig. 1. Compared to high resolution 3D scans, such as from Minolta, the Kinect 3D face model is hardly recognizable as a human face and most of the popular face landmarks such as eye or mouth corners are not precisely locatable even manually. In this paper, we look into the feasibility of using the Kinect depth data alone as well as in combination with texture for face recognition under varying pose, expressions, illumination and disguise. We report results for our proposed technique (details in Section 3) and four existing face recognition algorithms.

1.3. Contributions

In this paper, we look into the feasibility of using Kinect depth

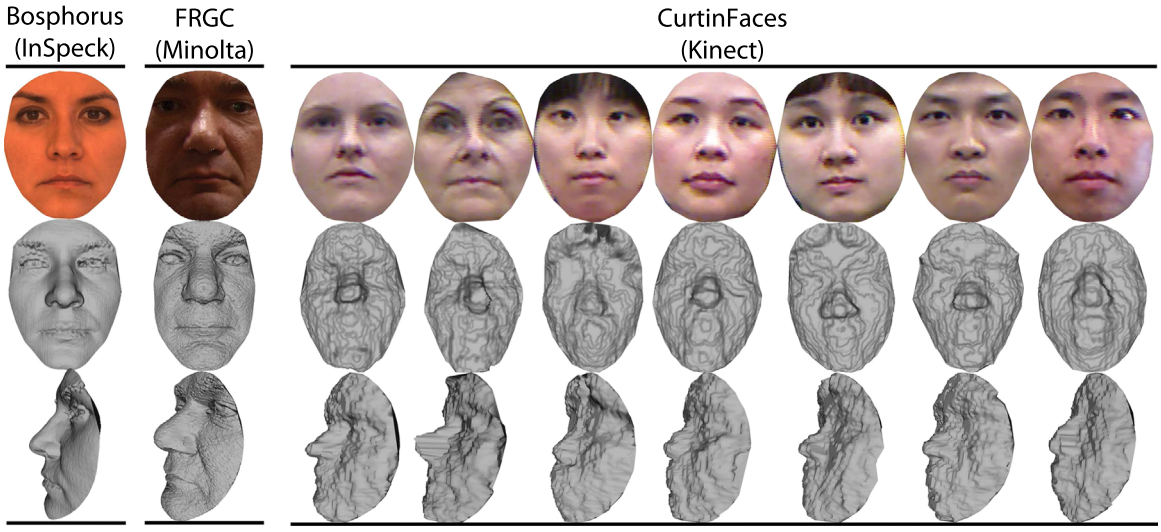


Fig. 1. Sample texture and 3D face models acquired with Minolta [13], InSpeck [14] and Kinect sensors. 3D faces are rendered as smooth shaded surfaces.

data for face recognition in uncontrolled conditions. We propose a face recognition algorithm that performs equally well on low and high resolution data. The proposed approach requires only the nose tip position and is composed of the following novel components. Firstly, a preprocessing algorithm is proposed which exploits the facial symmetry at the 3D point cloud level to obtain a canonical frontal view, shape and texture, of the faces irrespective of their initial pose. Secondly, a Multi-channel Discriminant Transform (MDT) is proposed for discriminative representation of color texture and 3D normal map images. Thirdly, a Multi-channel Weighted Sparse Coding (MWSC) method is proposed that computes a weight mask using multiple data channels which is more invariant to imaging conditions. The sparse coding part considers linear combinations of multi-gallery images, effectively utilizing the advantage of Kinect's high acquisition speed. Finally, the same algorithm can recognize faces under different and unknown poses, expressions, illumination and disguise.

To the best of our knowledge, this is the first work that formulates sparse coding for RGB-D data. An earlier version of our work appeared in [16]. In this paper, we extend the algorithm by including discriminant transforms and multi-channel weighting. We also perform more thorough experiments on multiple databases and comparisons to existing techniques.

Due to the lack of public Kinect based face database, we evaluate our algorithm on a new dataset collected using the Kinect sensor in our laboratory. This database, namely CurtinFaces, is available to the research community [16]. High recognition rates were achieved under challenging experiments, which justify the feasibility of performing non-intrusive face recognition using a low-cost sensor. We have also compared our algorithm to four existing algorithms on CurtinFaces, Bosphorus [14] and CASIA [17] databases. The results show that our algorithm outperforms existing methods on all three databases. Finally, the proposed algorithm is tested on the FRGC v2 database, and achieves comparable performance to the state-of-the-art.

2. Related work

Although many methods have been proposed for 3D face recognition with increasing performance and sophistication, they are not designed for noisy data such as from Kinect. It would be interesting to see the performance of existing 3D face recognition techniques on Kinect data. Bowyer et al. [18] gave a comprehensive survey of 3D face recognition methods in 2006 and recent developments until 2012 are covered in the literature review section of a recent paper by [10]. Here, we complement these surveys and discuss the limitations of some representative techniques, especially in the context of uncontrolled face images acquired with a noisy sensor such as the Kinect.

The Iterative Closest Point (ICP) algorithm was proposed by [19] for the registration and comparison of rigid surfaces. It has been used by many researchers for pose normalization and comparison of 3D faces. It finds the optimal rigid transformation that minimizes the distance between the corresponding (nearest) points of two 3D datasets. The final registration error is generally used as a classification criterion. ICP and its variants have been used for 3D face recognition [7,20]. The point-to-point error of ICP is sensitive to expression variations and incorrect point-to-point correspondences may lead to a local minimum. These two problems are more likely to occur when ICP is applied on a pair of noisy 3D faces acquired by Kinect and therefore, the result can be highly inaccurate.

Bronstein et al. [21] proposed an expression-invariant representation of the facial surfaces based on isometric deformations. Matching was done by computing distance between the canonical forms of two faces in their embedded subspace. This algorithm assumes that all faces are frontal and does not perform any pose correction. Imaging artifacts caused by disguise changing the facial surface deformation can also affect the canonical representation.

Mian et al. [7] proposed multi-modal (2D + 3D) hybrid (holistic + part-base) approach for robust face recognition. An iterative PCA based algorithm was used for pose correction. Matching of two faces was done using several heuristics: similarity of the SIFT and spherical feature on the holistic 2D and 3D faces respectively, and ICP registration errors of the segmented 3D nose and eyes–forehead components. They have shown that these segmented parts are robust against expression variations. However, the segmentation requires automatic detection of the inflection points around the nose, which may not be achievable on the noisy Kinect data. Faltemier et al. [20] divided a face into 28 overlapping regions in order to improve recognition robustness. ICP method is applied on each region and matching of two faces was performed by fusing the regional registration errors. The subdivision idea is very effective for high resolution scans. Over 80% matching accuracy was reported on some of the small individual region of just 25–45 mm spherical radius. However, small regions in a low resolution data may not be sufficiently discriminative for face recognition.

Queirolo et al. [8] proposed a Simulated Annealing (SA) approach for range image registration and Surface Interpenetration Measure (SIM) for similarity. Matching was done by fusing the SIM for elliptical regions around the nose, forehead and the holistic face. Although impressive results were reported on the FRGC database, their algorithm requires six landmark points including eye and nose corners, which cannot be detected, even manually, on the noisy Kinect 3D face data. Kakadiaris et al. [22] proposed the Annotated Face Model (AFM) to register the input 3D face to an expression-invariant deformable model. After fitting onto the AFM, several features were extracted on the geometric and normal map for matching. Recently, Passalis et al. [23] further extended the AFM method with facial symmetry to handle missing data caused by self-occlusion in non-frontal poses. Two fitted AFM were generated for matching by mirroring the AFM external forces from one side to the other. As a result, their method can handle pose and expression variations at the same time. However, the fitting of AFM requires eight landmarks over the 3D face, which is even more difficult than Queirolo et al.'s [8] method to be applicable on Kinect data. Even on high resolution scans, some landmark detection errors can be greater than 10 mm, especially on 60° side scans.

Wang et al. [24] proposed a novel representation namely the Signed Shape Difference Map (SSDM). They fused several features extracted on the SSDM based on a boosting algorithm for face recognition. Pose correction was done by aligning the normal of the symmetry plane, nose tip and direction of nose bridge. The symmetry plane was determined, by registering the 3D face with its mirrored version using ICP, and fitting a plane to the resulting registered faces. This pose correction method is more efficient than ICP since it avoids registration to every gallery face. The proposed SSDM was created by taking the direct pixel differences between two depth images after pose correction. However, searching for the symmetric plane on a profile view leads to non-convergence of ICP. Additionally, the SSDM may not be effective for Kinect data where depth map differences can also be a problem due to noise.

Alyuz et al. [25] is one of the very few works that address the disguise problem in 3D face recognition. They registered the probe face to a generic face model by applying ICP to the area around the nose. After registration, points that were far away from the generic face model were treated as outliers and removed. Missing data

Table 2
Summary of some 3D methods on their required resolution, landmarks and the main variation they addressed, i.e. Pose (P), Illumination (I), Expression (E) and Disguise (D).

Methods	Res.	Landmark	Var.[a]
A. Bronstein [21]	High	2	E
A. Mian [7]	High	5	E
T. Faltemier [20]	High	Nose tip	E
C. Queirolo [8]	High	6	E
Y. Wang [24]	High	Nose tip	E
G. Passalis [23]	High	8	E, P
N. Alyüz [25]	High	Nose tip	D
H. Drira [26]	High	4	E,P,D
N. Alyüz [27]	High	22	E,D
S. Berretti [28]	High	N/A	E,P,D
D. Smeets [29]	High	22	E,D
N. Alyüz [30]	High	9	E
A. Colombo [31]	High	3	D
O. Ocegueda [32]	High	N/A	E
D. Huang [33]	High	N/A	E,D,P
This paper	Low	Nose tip	P, I, E, D[b]

[a] Main variation addressed.
[b] We also consider some combinations of variations.

was then restored using "Gappy PCA". Matching was done by dividing the restored face into 30 regions and fusing the regional similarity scores obtained from multiple local Linear Discriminant Analysis (LDA) classifiers. Although they achieved 76–94% in their experiments on the Bosphorus 3D face dataset [14], they have only considered frontal views with neutral expression and have not compared with other 3D methods.

A summary of the aforementioned methods is presented in Table 2. The main limitations of these techniques are as follows.

- All methods assume the availability of high resolution 3D scans and therefore, may not work well for low resolution data.
- Some methods require more than one landmarks which may not be accurately detectable on low resolution 3D data.
- Most techniques rely on face segmentation or region sub-division to ensure robustness. This idea works well for high resolution scans because these smaller parts themselves are discriminative enough to separate different identities. However, such an approach is not feasible for noisy Kinect data, where even the completed 3D face is hardly recognizable.
- Most algorithms focus on 3D data alone and ignore the 2D texture. However, 3D data alone is insufficient for robust face recognition especially when acquired with a low resolution sensor. Although pure 3D techniques can be extended to multi-modal (2D+3D) in a straight forward way using score level fusion, a more sophisticated approach that considers interaction between 2D and 3D data will be more reliable and accurate.
- Most methods are optimized for the FRGC data alone and have not been tested on other datasets. Their high performance could well be due to overfitting on this data since faces in the FRGC data are all near-frontal and without disguise. In realistic applications, uncontrolled images can be acquired in arbitrary variations of pose, illumination, expression and disguise. None of the existing methods are evaluated against all of these factors.

3. Overview of the proposed method

The proposed method is designed specifically for robust face recognition using Kinect and when the query image is acquired without the user's explicit cooperation. The following imaging conditions can be expected in this context:

- Multiple training images of the enrolled subjects are available. This is possible with Kinect's high acquisition speed.
- The query image is uncontrolled since it can be acquired under arbitrary poses, illuminations, expressions, and possibly with disguise.
- Both 2D (RGB) and 3D (*XYZ*) data are available. The 3D data are in low resolution and noisy.

An overview of our proposed algorithm is shown in Fig. 2. It consists of three blocks, namely canonical preprocessing, discriminant transform, and multi-channel weighted sparse coding. The details of the three blocks are given in Section 4, 5 and 6 respectively.

Our results show that the proposed algorithm performs equally well on high and low resolution data. Moreover, it can use a single gallery image per subject or exploit the presence of multiple gallery images to perform recognition. Although multiple query images can also be acquired during recognition, we restrict our experiments to recognition based on a single query image to make our results consistent with standard face recognition protocols.

4. Canonical preprocessing

The input to our canonical preprocessing algorithm is a 6D (*XYZ*-RGB) point cloud and the output is a canonicalized depth map and registered RGB texture image of the face. Unlike common

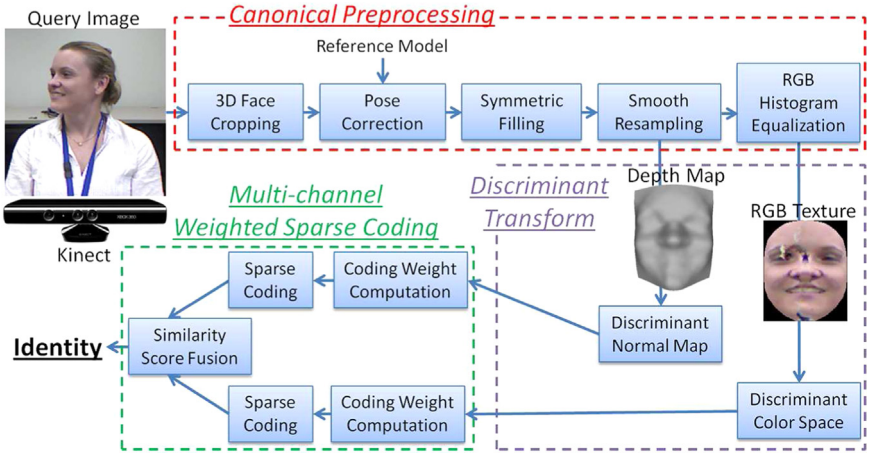


Fig. 2. Overview of the proposed method.

range data preprocessing which only removes spikes and fills holes, the proposed algorithm additionally corrects the facial pose so that it is view-point invariant and completes missing data due to self-occlusion. In fact, most data obtained from the Kinect sensor do not have spikes.[1] Holes are filled during a resampling step. Details of each preprocessing step are given below.

4.1. 3D face cropping and pose correction

Due to the level of noise in Kinect depth data (as illustrated in Fig. 1), the nose tip is the most reliable landmark that can be located on the 3D face. In this work, we assume that the approximate nose tip location has been detected. Since the nose tip is required only for rough alignment and face cropping, the algorithm works as long as the detected nose tip is close enough to the true location. Given the nose tip position, we first translate the point cloud such that the nose tip is at the origin. Then a sphere of 80 mm radius centered at the nose tip is used to crop the face. As a result, a 6D point cloud (*XYZ*-RGB) of only the face area is obtained.

The Iterative Closest Point (ICP) algorithm [19] is an accurate technique for aligning two 3D point clouds. However, it is known to be computationally expensive, and hence registering the query face to every frontal gallery face in search of the best alignment is not feasible. Instead, we register the query (*XYZ* only) to a reference model. Since different subjects have different face shape, the reference face model must be a reliable representation of common 3D faces. Such a reference face cannot be constructed from the noisy Kinect data. Therefore, we build the reference using face models (with no expression) from the FRGC [13] and the UWA database [34]. The reference face is constructed by aligning the scans, resampling them on a uniform 128×128 grid and then taking their mean. The reference face has 64 points between the centers of the eyes. The number of points from the center of the lip to the line joining the eyes is also 64. Fig. 3 shows the reference face used in our experiments.

In pose correction, both training and test data are registered to the reference face by ICP, which needs a coarse initialization. However, the pose variations may up to $\pm 90^\circ$, which makes certain part of the face are rather far from reference face. Therefore, we do not consider point correspondences further than 16 mm apart in the beginning of ICP iteration. Instead, only the points around given nose tip of uncorrected are cropped out by a sphere of 16 mm radius centered at nose top, and then fed to ICP algorithm to build correspondences to reference face. Such a setting allows us to correct poses up to $\pm 90^\circ$ since ICP does not require the isometric of two point cloud. Once the correspondence is established based on the points around nose in first few iterations, the whole uncorrected face are then used to find more correspondences until the two faces are correctly aligned. An example registration of a profile face to our reference face is shown in Fig. 4.

4.2. Symmetric filling and resampling

After pose correction, some data may be missing due to self-occlusion for non-frontal views. This missing data can be estimated based on facial symmetry. Despite the fact that human faces are not perfectly symmetric, the variations caused by facial asymmetry are less than the variations caused by different identities [23]. Unlike the work in [23] which mirrors the AFM external forces from one side to the other and then generates two different fitted AFMs for recognition, we utilize facial symmetry in the preprocessing stage at the point cloud level. Specifically, a mirrored point cloud is created by replacing the X values in the original point cloud by their opposite numbers ($-X$). However, not all the mirrored points are useful as we only want to fill in the missing data. Ideally, no point should be added on a frontal face, while all points should be mirrored on a profile view. To this end, for each mirrored point, we compute its Euclidean distance using (XY values only) to the closest point in the original point cloud. If this distance is less than δ, the mirrored point is removed. The idea is to add the mirrored point only if there is no neighboring point at that location. Note that Z is not used when calculating the distance, because the difference in Z is usually caused by facial asymmetry rather than missing data. The remaining mirrored points are then combined with the original point cloud before resampling. A sample symmetric filling can be seen in Fig. 5. The threshold δ can be chosen based on the spatial resolution of the sensor or the point cloud itself. In our experiments, it was user defined. Depending on the original sample density, high values of δ will lead to a noisy surfaces while values too low will not help in symmetric filling. We empirically found that a good balance can be achieved with $\delta = 2$ mm, however, the performance is not affected much when setting δ to values between 1 and 5 mm.

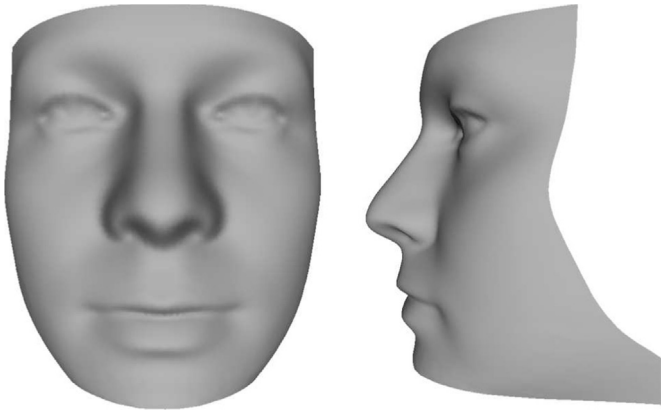


Fig. 3. The reference face model.

Resampling is the final step in our preprocessing algorithm, which is done by fitting a smooth surface to the point cloud (*XYZ*) using an approximation approach.[2] This algorithm fits a surface to the points with a smoothing (or stiffness) constrain that does not allow it to bend abruptly and thereby alleviating the effects of noise and outliers. Since surface fitting is done after symmetric filling, the added mirror points will also contribute to the surface. This is especially helpful to stabilize the noisy Kinect data. For each face, 161×161 points are re-sampled uniformly from its minimum to the maximum X and Y values. The advantage of re-sampling from min to max is that it aligns faces on a 2D grid. Notice that we do not smooth the RGB texture since it is not noisy and smoothing will only blur it. Instead, we just re-sample it to the same XY location with interpolation. After re-sampling, the X and Y grids are discarded and the Z depth map is converted to its surface normal map. Finally, six 161×161 matrices (RGBN_x N_y N_z) are obtained. These matrices are down-sampled to 32×32 for further processing. Some sample output images of the proposed canonical preprocessing algorithm are shown in Fig. 10b.

4.3. RGB histogram equalization

The above preprocessing steps also give the RGB texture of the face that is registered with the depth map. However, texture is easily affected by illumination. Therefore, the last step of our proposed preprocessing algorithm aims to enhance the reliability

[1] It is possible that filtering is done inside the Kinect hardware or API.

[2] mathworks.com/matlabcentral/fileexchange/8998.

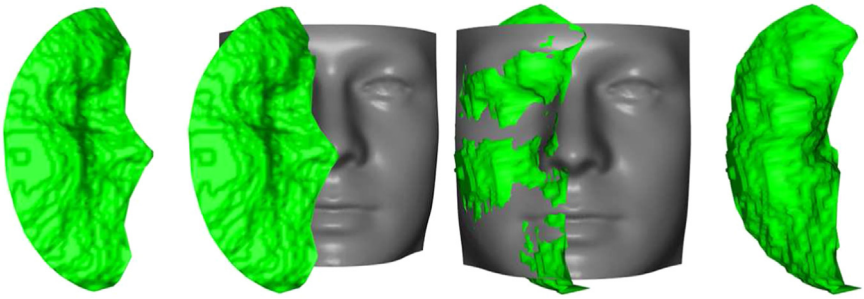


Fig. 4. First column shows the profile face before ICP and last two columns show the result after ICP converge.

of the texture map. Since we only want to enhance the illumination contrast, the RGB color image is first converted to the CIE $L^*a^*b^*$ color space, where the three channels representing the lightness of color (L^*), position between red and green (a^*) and position between yellow and blue (b^*), respectively. The Contrast-Limited Adaptive Histogram Equalization (CLAHE) method [35] is applied on the L^* component only, using a 2×2 window, to enhance local contrast. This method applies histogram equalization on each patch of the image with a specified contrast limit to avoid amplification of noise. Neighboring patches are then combined using bilinear interpolation to eliminate artificially induced boundaries. The standard histogram equalization method is then applied again on L^* channel of the holistic image. Finally, the resulting image is converted back to RGB space. Some examples are shown in Fig. 6, in which the first row is the original image and the second row is the pre-processed images. Compared with the original images, the facial details are more apparent after preprocessing.

5. Multi-channel discriminant transform

Since both the color map (RGB) and normal map (N_x N_y N_z) images consist of multiple channels, we can improve their discriminative power by applying a transformation that is learned from labeled training data similar to the idea of Discriminant Color Space (DCS) [36]. Most existing multi-modal (2D+3D) face recognition methods convert the color image to gray-scale first [7,37]. However, color information is proven to be useful especially when the shape cue is noisy [38]. Therefore, color cue is likely to be very useful in the case of Kinect where the 3D data is noisy and low resolution.

Color images are usually modeled in the RGB space which is not a discriminant space due to high inter-component correlation [39]. Although the authors of [39] have proposed Color Space Normalization (CSN) to reduce correlation, CSN does not consider class separability. Recently, optimal color spaces, that are learned from the training data to maximize class separability, are proposed. The Discriminant Color Space (DCS) [36] method finds a set of linear combinations for the R, G and B components in order to maximize class separability similar to the idea of LDA. The Color Image Discriminant Model (CID) [40] seeks the optimal color space and feature subspace simultaneously. The Tensor Discriminant Color Space (TDCS) [41] method models the color image as tensor and seeks the color space transformation and two feature subspaces respectively along the row and column directions. However, both CID and TDCS methods do not have closed-form solutions and are solved iteratively which can lead to local minima or non-convergence. Our experience shows that DCS is a reliable color space for face recognition.

Similar to RGB image, the normal map also has three channels (N_x N_y N_z). A discriminant transform similar to DCS can be derived to increase its discriminative power. To this end, we propose a Multi-channel Discriminant Transform (MDT) method, which is a generalization of the DCS method to work on multi-channel data of any order.

Suppose there are a total of M training samples of C classes. Each d dimensional training sample with h channels is denoted by

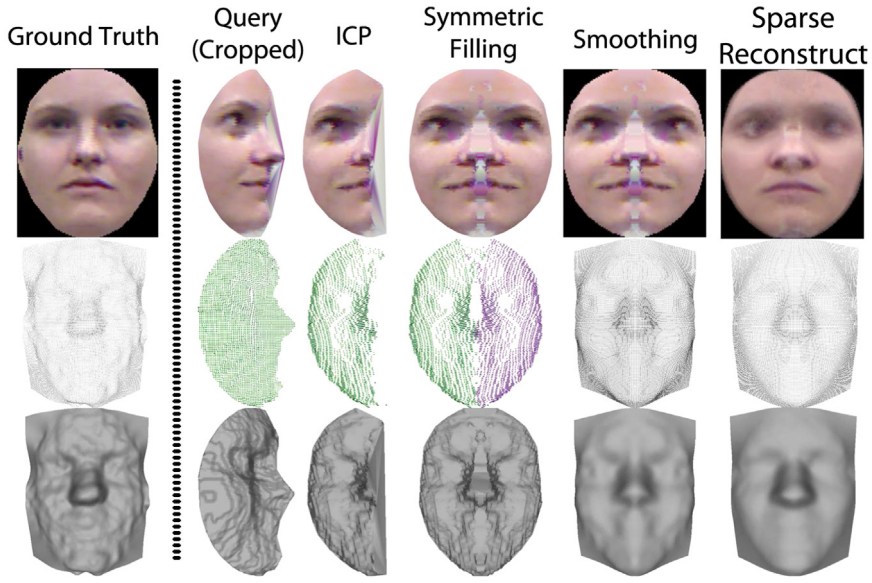


Fig. 5. Example canonical preprocessing and sparse reconstruction on profile view probe image.

Fig. 6. Result of RGB histogram equalization.

$U_j = [u_j^1, u_j^2, \ldots, u_j^h] \in R^{d\times h}$, where $j = 1, 2, \ldots, M$. We can define the following linear transform:

$$V = a_1u^1 + a_2u^2 + \cdots + a_3u^h = U_j \cdot A, \tag{1}$$

where $A = [a_1, a_2, \ldots, a_h]^T$ is a vector of the transformation coefficients. A good transformation vector should point to the direction that maximizes class separability. Let S_b^v and S_w^v be the scatter matrices to describe the between-class and within-class variance in the transformed V -space, then A can be found by maximizing the following objective function:

$$J(A) = \frac{\mathrm{tr}(S_b^v)}{\mathrm{tr}(S_w^v)}, \tag{2}$$

Let P_i be the label of the ith class (i=1,2,…,C), $\overline{U_i}$ be the mean sample of class P_i, $\overline{U}$ be the grand mean for all data, and M_i be the number of samples in class P_i, we can derive the following:

$$\mathrm{tr}(S_b^v) = A^T\left(\sum_{i=1}^{C} M_i(\overline{U_i} - \overline{U})^T(\overline{U_i} - \overline{U})\right)A = A^T S_b^u A, \tag{3}$$

$$\mathrm{tr}(S_w^v) = A^T\left(\sum_{i=1}^{C}\sum_{U_j\in P_i} (U_j - \overline{U_i})^T(U_j - \overline{U_i})\right)A = A^T S_w^u A. \tag{4}$$

The objective junction in Eq. (2) can be re-written as:

$$J(A) = \frac{A^T S_b^u A}{A^T S_w^u A}. \tag{5}$$

Since S_b^u and S_w^u are both $h \times h$ nonnegative definite matrices in general (where h is usually small), the optimal solution can be obtained by solving the equivalent generalized eigenvalue problem:

$$S_b^u A = \lambda S_w^u A, \tag{6}$$

such that $A = [A^1, A^2, \ldots, A^h] \in R^{h\times h}$ are the eigenvectors arranged in descending order of their corresponding eigenvalues.

In the proposed approach, the RGB map is ordered as a 3-channel sample (i.e. $[R, G, B] \in R^{d\times 3}$). We apply MDT to obtain the corresponding discriminative transformation matrix $A = [A^1, A^2, A^3] \in R^{3\times 3}$. Although A^1 is the most discriminative projection, usually all three eigenvectors are required to achieve maximum recognition performance. After the transformation, the texture map is converted from RGB space to Discriminant Color Space (DCS). Similarly, by applying MDT on the normal map which consist of three channels (i.e. $[N_x, N_y, N_z] \in R^{d\times 3}$), a Discriminant Normal Map (DNM) is obtained. Both DCS and DNM consist of three discriminant channels. Each channel is then normalized to zero mean and unit standard deviation to avoid magnitude domination of one channel over the others.

Fig. 7 shows some sample images after transformation and a plot of Euclidean distances between images of different subjects of the FRGC dataset. It can be observed that the images exhibit a greater color contrast after MDT. Similarly, distances between images of different subjects also increase i.e. the red circles shift upwards.

6. Multi-channel weighted sparse coding

Given multiple gallery images per subject, one way to utilize them effectively is by allowing their sparse linear combinations to be matched with a query image [2]. Although our preprocessing algorithm removes some level of noise and completes missing data due to occlusions, it cannot perfectly reconstruct a frontal view from profile views. This is because missing data cannot be estimated when there are no reference points for mirroring. See Fig. 5 as an example, which shows an error line in the middle of the resulting canonical face image. A robust approach such as weighted sparse coding [3,42], which are shown to be robust against outliers and missing data, can be employed to overcome this problem. In this paper, we propose a novel face recognition method namely Multi-channel Weighted Sparse Coding (MWSC), which extends Yang et al.'s Robust Sparse Coding [3] method to work effectively on multiple channels (e.g. R, G and B) and multiple modalities (2D + 3D). First we formulate the sparse coding as the weighted Lasso problem with ℓ_1 penalty:

$$x = \mathrm{argmin}\ \|W(Ax - y)\|_2 + \lambda\ \|x\|_1 \tag{7}$$

where A is the dictionary i.e. the training samples in our case, y is the query face, x is the coding parameters vector, λ is a constant that controls the coding sparsity and W is a vector consisting of weights for each variable in A. Yang et al. [3] have shown that a robust W can be estimated by this function:

$$W = \frac{\exp(\mu\delta - \mu(e)^2)}{(1 + \exp(\mu\delta - \mu(e)^2))} \tag{8}$$

where $e = Ax - y$ is a vector of reconstruction residuals, μ and δ are user defined parameters controlling the rate of decrease and the location of demarcation point respectively. $W^{(1)}$ is initialized as the residual to the mean dictionary atom $e^{(1)} = \overline{A} - y$. Eq. (7) is then iteratively solved for x and W. The iterations stop at the tth iteration when the change in W is smaller than ε, i.e.

$$\| W^{(t)} - W^{(t-1)} \|_2 / \| W^{(t-1)} \|_2 < \varepsilon \tag{9}$$

In our case, the d-pixels query image $Y = [c^1, c^2, c^3] \in R^{d\times 3}$ consists of three channels: c^1, c^2 and c^3 (which can be either the three channels of DCS or DNM). In order to apply Eq. (7), we

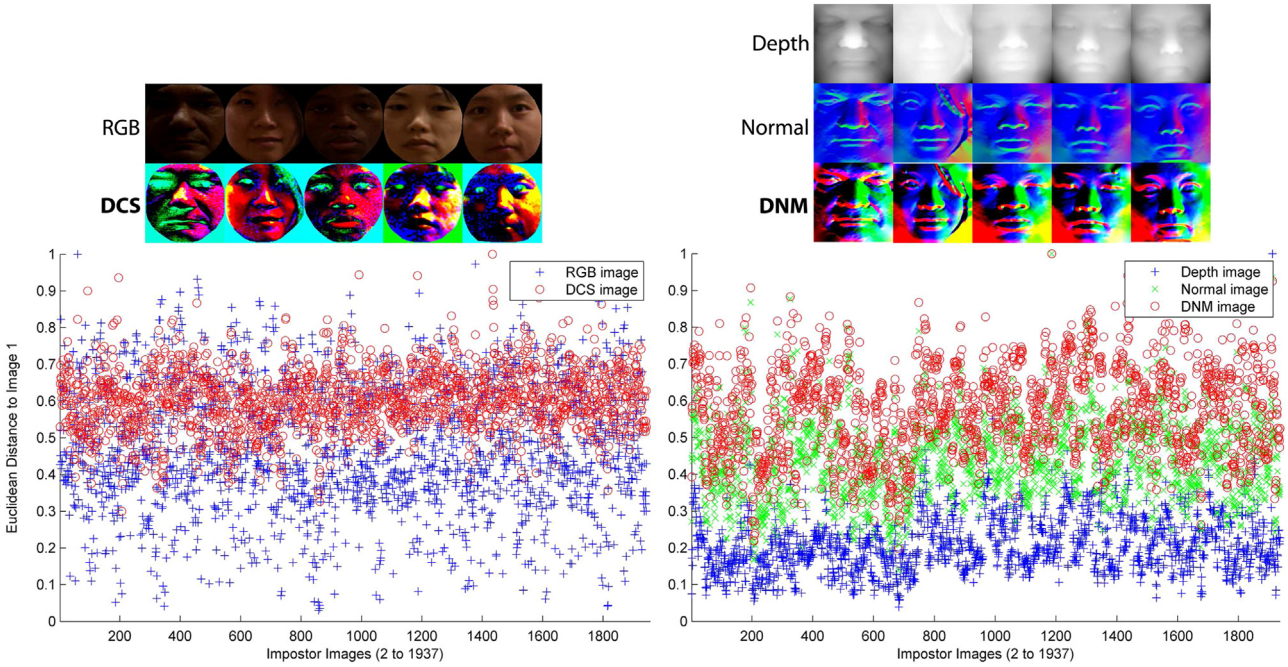


Fig. 7. Some sample images from the FRGC dataset. Different subjects are more discriminative after Multi-channel Discriminant Transform (MDT). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

convert Y to a column vector $y = [c^1 c^2 c^3]^T \in R^{3d}$ by stacking the three channels. The dictionary $A \in R^{3d \times m}$ with m training samples is also arranged in a similar way. While this is a standard data-level fusion strategy, we propose a more effective way to compute the weights W.

We use three channels instead of gray scale images because outliers have more chances of getting detected with multi-channels compared to a single channel. Consider a toy example where a pure red face (RGB1: [1 0 0]) is wearing a pure blue scarf (RGB2: [0 0 1]). Although the facial part and scarf have different colors, they have the same gray-scale intensity of 1/3. Since the residual vector e is computed by direct pixel difference, both red and blue pixels are likely to contribute equally in a grayscale image. Based on this observation, we treat each multi-channel pixel as a h-dimensional vector ($h=3$ in our case), and compute the residual for each pixel using Euclidean distance in the h-dimensional space.

In each iteration, after the coding vector x is computed using Eq. (7), the reconstructed image vector $\tilde{y} = Ax \in R^{3d}$ is rearranged back to image matrix $\tilde{Y} = [\tilde{c}^1, \tilde{c}^2, \tilde{c}^3] \in R^{d \times 3}$. The query image vector y is also rearranged back to image matrix $Y = [c^1, c^2, c^3] \in R^{d \times 3}$. The d-dimensional residual vector is computed pixel-wise by

$$e_j = \| \tilde{Y}_j - Y_j \|_2 \quad (j = 1, \ldots, d). \tag{10}$$

In our approach, two sets of weights (W_{tex} and W_{dep}) are computed for the DCS and DNM images respectively. Similarly, separate coefficient vectors (x_{tex} and x_{dep}) are computed by sparse coding the DCS and DNM images using Eq. (7). For C classes, two sets of class-wise similarity scores (S_{tex} and S_{dep}) are computed based on the class-wise weighted reconstruction residual:

$$\begin{aligned} S^i_{tex} &= -\| W_{tex}(A^{\in P_i}_{tex} x^{\in P_i}_{tex} - y_{tex}) \|_2, \\ S^i_{dep} &= -\| W_{dep}(A^{\in P_i}_{dep} x^{\in P_i}_{dep} - y_{dep}) \|_2 \end{aligned} \tag{11}$$

where $i=1,\ldots,C$, and P_i is the label for class i. The two scores, S_{tex}

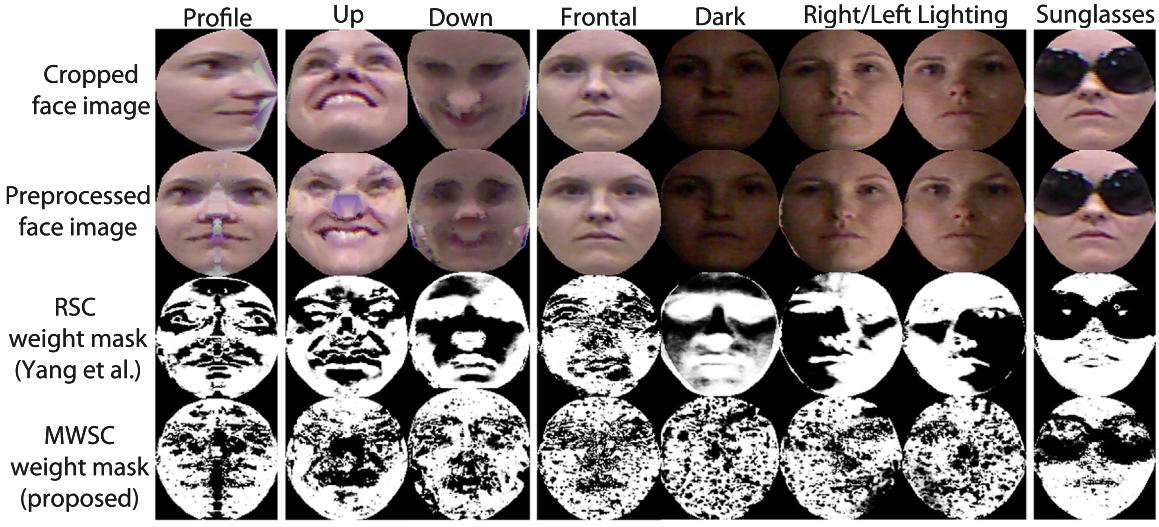


Fig. 8. Weight masks computed for some probes in CurtinFaces (without histogram equalization). The proposed MWSC works better in terms of masking out outlier pixels.

Table 3
Some publicly available 3D databases.

Name	Year	Res.	Acq.	Variation[a]
CASIA [17]	2004	High	4624	P, I, E[b]
FRGCv2 [13]	2005	High	4007	E, I
BU-3DFE [45]	2006	High	2500	E
ND2006 [46]	2007	High	13450	E
Bosphorus [14]	2008	High	4652	P, I, E, D
CurtinFaces	2012	Low	5044	P × E, P × I, D
KinectFaceDB [47]	2013	Low	N.A.	P, I, E, D

[a] Main ones only: {P(pose), I(illum.), E(exp.), D(Disguise)}.
[b] CASIA also contains some E+I and P+E.

and S_{dep}, are then individually normalized using z-score technique [43] and then summed before final decision. The query Y is recognized as the person with the highest final similarity score.

Fig. 8 shows some sample weight matrices obtained using [3]'s Robust Sparse Coding (RSC) method and our proposed MWSC approach. The weight matrices are computed using the texture images in CurtinFaces and are shown as 256-level intensity images where brighter pixels represent higher weights. One can see that, for the preprocessed profile face (column 1), a few error lines appear in the middle caused by missing data. These lines are masked out better by our MWSC approach compared to RSC. For a smiling face that is looking up or down (columns 2 and 3), MWSC assigns lower weights to the mouth area, which is the non-rigid part of the face that easily deforms under facial expression. The most apparent advantage of MWSC over RSC is that it is more robust against illumination. RSC assigns low weights to shaded pixels under non-frontal lighting, while MWSC computes a sparse weight mask consistently for frontal faces (columns 4–7). In the last column, both RSC and MWSC correctly mask out the pixels associated with sunglasses.

For the choice of parameters, we set λ to 0.001 and ε to 0.05. Following [3], μ is set as c/δ and δ is chosen as the ℓth smallest pixel residual, where ℓ is computed by $\lfloor \tau n \rfloor$. We set c to 8 and τ to 0.6. We empirically found out that this setting yields the best trade-off between robustness, accuracy and speed. To show that these parameters are not tuned to overfit a specific situation, this setting is used through out this paper in all the experiments using different datasets.

7. Experimental results

Face recognition can be divided into verification and identification. The former verifies if two face images are from the same subject, while the latter finds the identity of a query face image within a database. Verification problems usually assume certain level of user cooperation (such as displaying the passport photo and looking at the camera). However, as discussed previously, this work focuses on applications where user cooperation may not be achievable. Therefore, we evaluate our system under face identification protocol. The proposed method applies canonical preprocessing on both training and test data as detailed in Section 3. The multi-channel Discriminant Transform (mDT) is applied afterwards. The multi-channel Weighted Sparse Coding (MWSC) method is then used to identify the probe. We also compare the performance of the proposed method with the following methods from existing literature:

- [7] (2D+3D)
- RSC [3] (RGB-D)
- SRC [2] (RGB-D)
- SVM-rbf [44] (RGB-D)

Although some of these methods were proposed for 2D gray-scale images, we extend them and tune them for RGB-D data for a fair comparison. First, we scale and translate every face image such that the eyes and mouths are aligned on the same pixels. Then a bounding box is used to crop the face area. These procedures are applied on both the RGB and depth data. Afterwards, they are resized to 32×32 and converted to two vectors by stacking their columns. The resulting 1024D depth vector and 3072D RGB vector are input into the aforementioned 2D algorithms separately, except for SVM, where PCA is applied first to reduce the data dimension retaining 99% energy (around 350D). Two different similarity scores obtained from RGB and depth are then normalized using z-score [43] and summed for final decision.

7.1. Performance on CurtinFaces: a Kinect face database

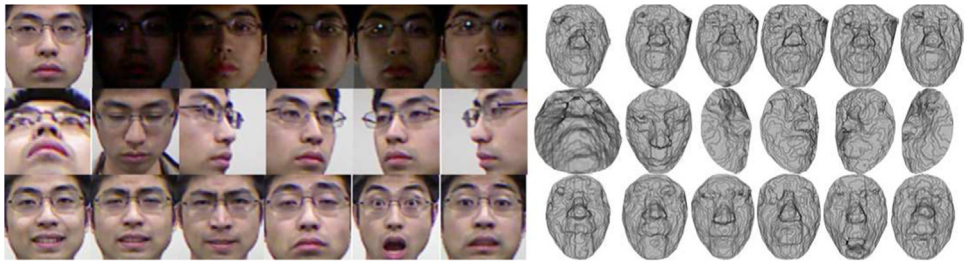
Table 3 lists some publicly available datasets. To the best of our knowledge, most of the existing 3D databases are acquired using high resolution scanners. In 2013, KinectFaceDB [47] is collected for 3D face analysis, providing 2D, 2.5D, 3D, and video-based face data. However, none of them consider extensive combination of variation factors. Therefore, we construct our own dataset namely CurtinFaces which is available to the research community.[3] This dataset contains over 5000 images of 52 subjects acquired using Kinect. In this work, we use a subset which consists of 4784 images of 52 individuals with variations in poses (P), illumination (I), facial expressions (E) and sunglasses disguise. The database contains facial images with and without glasses. For each subject, three images in the front, left and right profile view are without glasses. Additionally, for each subject, there are 49 images at $7E \times 7P$ and 35 images at $7E \times 5I$ i.e. combinations of 7 expressions with 7 poses and 5 illuminations. Images with sunglasses are under five conditions (i.e. $3P$ and $2L$). The full set of images per person is 92.

Out of the 92, 18 images per subject (see Fig. 9) contain only one type of the three variations (I, P or E). These images are used as the training/gallery set after preprocessing, since it is more suitable to learn the DCS and DNM transformations and as the coding dictionary. As shown in Fig. 9(a), the images in first row only contain illumination changes, the second row mainly reflect the variation of pose, and the last row shows the images under different expression. The 74 remaining images per subject are used as test images (see Fig. 10). For all images, the nose tip is manually detected.

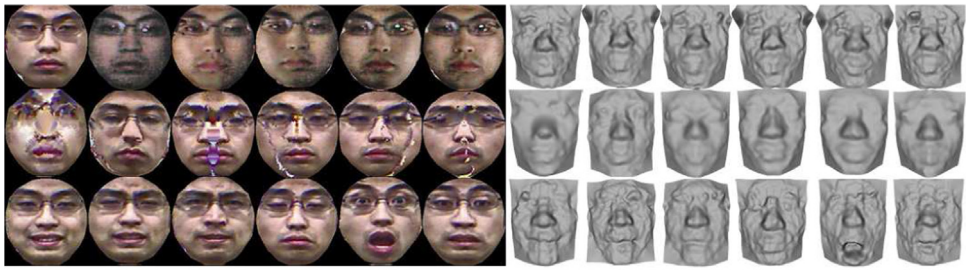
Identification results are reported in Fig. 11 and labels are defined in Fig. 10a. Fig. 12 shows the CMC curves of several methods on CurtinFaces database, and Table 4 records the rank-1 identification rates. It can be seen that the proposed method achieves the best performance under pose variations (yaw and pitch), illumination changes and occlusions (sunglasses). For simultaneous variation in pose and expression (P × E, top two plots of Fig. 11), the performances of other methods decrease dramatically with larger pose variations (both yaw and pitch) to the extent that accuracy is <15% on profile query faces ($\pm 90°$). Although Main et al.'s MMH method performs pose correction using 3D data, their method is not designed to handle missing data caused by self-occlusion on non-frontal views. Thus, the proposed method is more robust to variations.

It can be observed that (bottom left plot of Fig. 11) most methods for simultaneous variation in illumination and expression (I × E) are not affected by illumination. This is because of the fusion of depth data which is less sensitive to illumination. Also notice that RSC performs the worst in this case. The reason is that,

[3] Curtin Dataset is available by sending a request to w.liu@curtin.edu.au.



(a) Before preprocessing.



(b) After preprocessing.

Fig. 9. Sample enrollment images of one subject.

as illustrated in Fig. 8, the RSC's weight mask computation is sensitive to extreme outliers such as texture pixels that are in cast shadows. Although our method also uses texture data, it is not affected by illumination and is able to maintain consistent performance across different lighting conditions. In Table 5, we list the intermediate results of the proposed method under different face poses. The experiments are conducted based on RGB texture, depth, and fusion data respectively. The average recognition rate is improved from 77.0% to 96.3% with the help of symmetric face based on depth information. This is a significant improvement for low-resolution face recognition under severe pose variation. In symmetric scenario, the recognition rates of faces with severe pose variation can be boosted when depth information is fused with RGB texture, which shows the significance of the depth information. On the contrary, the depth information does not help if the non-frontal face are used without symmetric. That is to say, the depth information can not be used directly, which reveals the necessity and effectiveness of the proposed canonical preprocessing. Although the method in [48] achieves a comparable result on CurtinFaces data, it uses image sets of different head poses, which requires more information for training.

The results of sunglasses disguise are presented in the bottom right plot of Fig. 11. As expected, SVM achieves very low performance because it is not designed to handel occlusion. RSC performs better than SRC on the average, as it can correctly mask out the sunglasses pixels. Mian et al.'s method can handle disguise to some extends as it segments the face into two parts and the nose part is completely un-occluded by sunglasses. However, our method achieves the best performance. The main reason is that the small nose region used by MMH is not discriminative enough due to Kinect's low resolution.

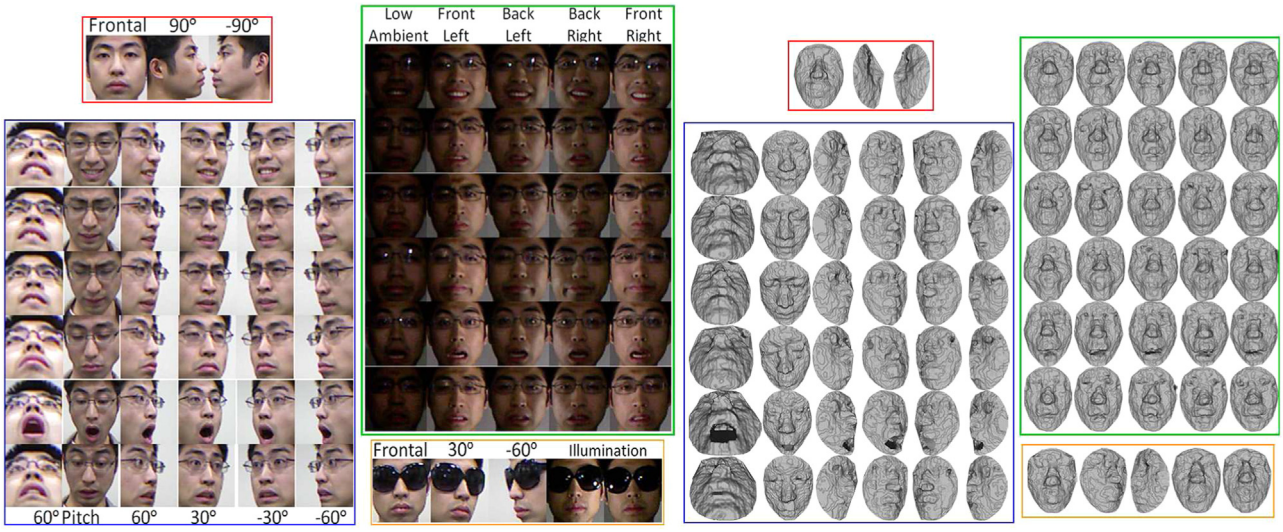
The main reason why methods like MMH that have been tested on the FRGC data are not suitable for Kinect is because they require face segmentation or landmark identification which can not be accurately performed on low resolution data. Observe that MMH only achieves 94.2% on frontal views with expressions. This suggests that the idea of using rigid face segments to deal with expressions is not very effective on low resolution data. This is due to the fact that smaller regions themselves are not sufficiently discriminative in low resolution. Moreover, some errors are also caused by the failure of landmark detection leading to incorrect face segmentation.

Furthermore, MMH achieves only 85.5% for test images under illumination variations (frontal views with expressions). This significant drop in performance is caused by the false rejection in the first stage of MMH where candidates are eliminated by a low cost rejection classifier which is partly based on the SIFT features extracted from the texture image. Removing the rejection step or setting a higher threshold to reject less faces may not be feasible for identification problem due to the fact that the matching engine of MMH is based on ICP which is computationally very expensive. Matching a single query to a large gallery may take several hours.

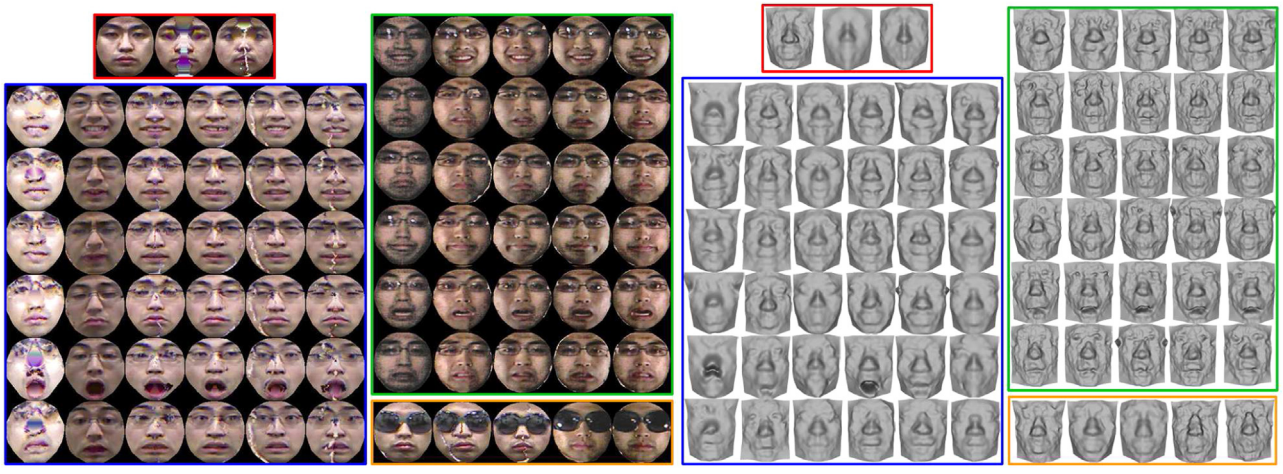
In summary, our proposed algorithm achieves an overall average of 96.5% and 94.6% rank-1 identification rates respectively using only the DCS texture and DNM depth images alone. Considering the high levels of noise in Kinect 3D data, 94.6% accuracy is a significant achievement. Combining DCS and DNM, our proposed approach is able to achieve an average of 98.4% identification rate. Even under the extreme case of profile view, our accuracy is 87.5%. To the best of our knowledge, these are the best identification rates reported for low resolution 3D data acquired under challenging conditions. Our results justify that noisy Kinect data is useful for face recognition.

7.2. Performance on Bosphorus database

To show that the proposed algorithm works equally well on high resolution scans, we evaluate it on the Bosphorus 3D face database [14] which was acquired with a high resolution scanner namely InSpeck. This database contains 4666 scans of 105 individuals. Each subject has up to 54 scans out of which up to 35 are with expression variations. There are six standard emotional expressions and 28 expressions are performed according to the Facial Action Coding System (FACS) [49]. Besides expressions, the database also contains 13 poses with different degrees of yaw, pitch and a combination of both up to 90°. Each person also has up to four types of occlusions or disguises. Some sample images are shown in Fig. 13. The amount of uncontrolled factors in this



(a) Before preprocessing.



(b) After preprocessing.

Fig. 10. Sample test images of one subject.

database makes it very suitable for our study on non-intrusive face recognition.

All images in the Bosphorus database are labeled with up to 24 landmarks. However, our algorithm requires only the nose tip location. For training, we use the FRGC dataset to compute the DCS and DNM transformation matrices. For testing, we follow the first versus all experimental protocol similar to Li et al. [50]. This protocol uses the first scan of each subject (105) as the gallery and the rest (4561) for testing. In fact, this setting is not favorable for sparse coding which exploits linear combinations of multiple gallery images per subject. To overcome this limitation, we generate multiple samples from each gallery image. Since our approach works on images as low in resolution as 32×32, multiple independent (to some extent) samples can be generated by downsampling the original high-resolution image. More specifically, we downsample the original preprocessed image from 161×161 to 64×64, and then generate four 32×32 images by taking its alternate (even or odd) rows and columns. The fifth image is obtained by directly downsampling the 64×64 image to 32×32 using interpolation. These five samples mimic images taken by five independent low resolution cameras with slight translational shifts. Our claim is backed up by the increased identification rates we achieved using this approach.

Furthermore, these generated samples also increase the tolerance to minor translational errors caused by misalignments. Note that the use of synthesized samples do not violate Li's protocol since they are generated from a single gallery image. Therefore, our performance comparison to Li's work in the first versus all scenario is fair. A comparison of rank-1 identification rates is reported in Table 6. Note that Mian's approach [7] does not achieve good performance in this dataset because it was not designed to handle pose and occlusion variations. Our proposed method outperforms its competitors in all cases except with occlusion. This is mainly because of the point-to-point ICP registration which tries to register the hand-occluded surface to the reference face. In fact, our performance for occlusion can be increased if ICP step is skipped since all occluded faces are frontal. However, we assume we do not have previous knowledge about the query face and the same algorithm is applied on every database. The 91.1% identification rate shows that our system does not fail even in the difficult case of hand ling occlusion.

Interestingly, without the use of synthesized multiple enrollments, our system still achieves an overall average 96.6% identification rate on the Bosphorus database. This is because sparse coding is performed collaboratively over all gallery images. Faces of different subjects share some parts in common which can help

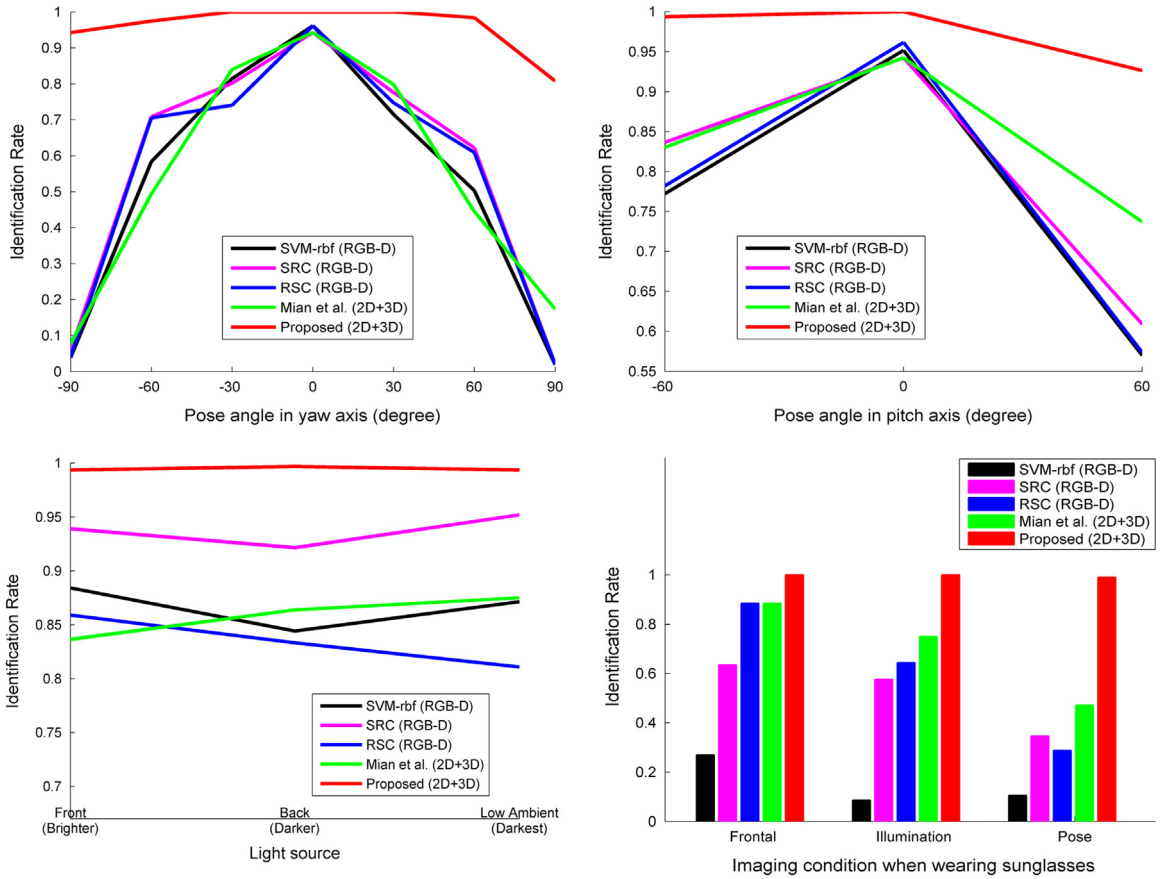


Fig. 11. Identification results on *CurtinFaces*. The top two plots show that the proposed method outperforms all others and is robust to pose variations in yaw and pitch. The lower two plots show that the proposed algorithm is robust to illumination and occlusions.

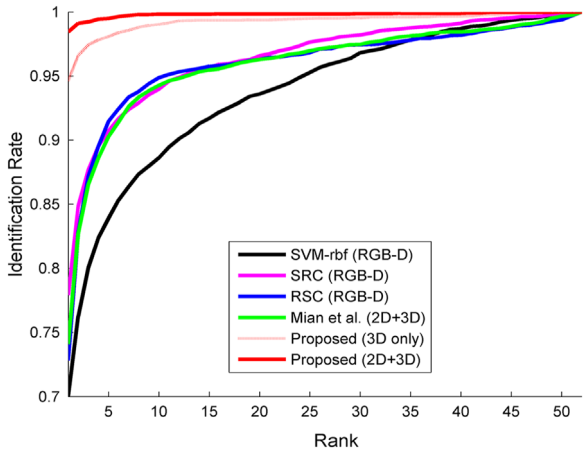


Fig. 12. CMC curves on CurtinFaces database.

Table 4
Rank-1 identification rates are summarized in the table. Note that the proposed method assumes the nose tip of is given.

Face variations	SVM	SRC	RSC	Mian	**Prop.**
Frontal (52)	96.2	94.2	96.2	94.2	**100**
Yaw 30° (624)	76.4	78.9	74.4	81.9	**100**
Yaw 60° (624)	54.3	66.5	65.7	46.7	**97.9**
Yaw 90° (104)	2.9	3.9	2.9	12.5	**87.5**
Pitch (624)	68.1	72.3	67.8	78.4	**96.0**
Illum. (1560)	87.6	93.5	83.9	85.5	**99.5**
Sun. (260)	13.1	49.6	55.0	62.3	**99.6**
All (3848)	70.0	77.9	72.8	74.1	**98.4**

Table 5
Performance of different face poses based on depth, texture and fusion data.

Pose	Without symmetric			With symmetric		
	Depth	Texture	Fusion	Depth	Texture	Fusion
Frontal (52)	100	100	100	100	100	100
Yaw 30° (624)	49.5	98.1	93.6	88.3	99.8	99.4
Yaw 60° (624)	14.9	80.4	55.1	87.0	97.4	98.2
Yaw 90° (104)	1.0	39.4	14.4	74.0	83.7	84.6
Pitch 60° (624)	77.2	91.3	90.9	81.6	89.1	92.8
Average	46.2	87.6	77.0	85.4	95.0	96.3

stabilize the sparse coding results (also see [52]). Due to the high discriminativeness of DCS and DNM feature, the sparse coding solution always assign a large coefficient to the image of correct identity, hence recognizing most of the query faces correctly.

In summary, our proposed approach achieves the highest average performance of 97.6% which is, to the best of our knowledge, the highest identification rate reported for the Bosphorus database. It is important to emphasize that the performances reported in the table are obtained by employing exactly the same algorithm and parameters as those used for CurtinFaces in Section 7.1. Our proposed method performs consistently and robustly across different datasets with arbitrary uncontrolled conditions.

7.3. Performance on CASIA

The CASIA dataset [17] is also acquired with a high resolution 3D scanner namely Minolta. It contains a total of 4624 scans of 123 individuals. We consider this dataset suitable for our experiments

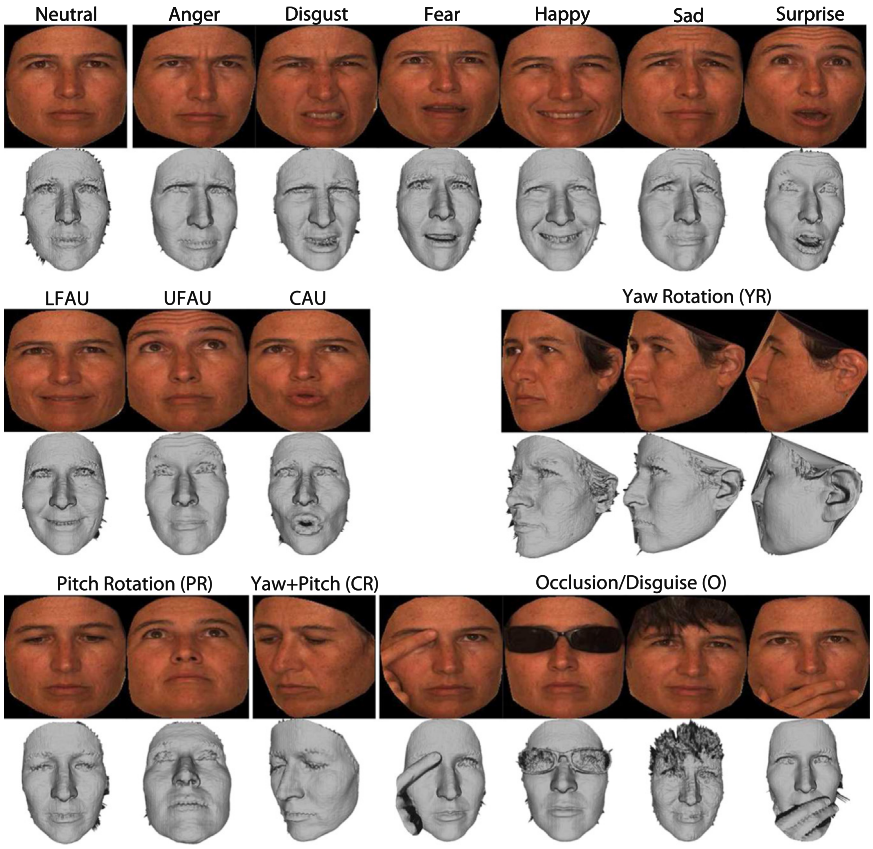


Fig. 13. Sample images of *Bosphorus*.

Table 6
Results on *Bosphorus* using first-neutral (105) vs. all (4561) protocol. Note that the proposed method assumes the nose tip is given.

Face variations	SVM	SRC	RSC	[7]	[50]	[51]	**Prop.**
Expression by emotion							
Neu. (194)	91.8	89.7	90.7	100	100	100	**100**
Anger (71)	67.6	64.8	81.7	93.0	88.7	97.2	**98.6**
Disg. (69)	52.2	75.4	88.0	88.4	76.8	87.0	**100**
Fear (70)	40.0	58.6	81.4	95.7	92.9	98.6	**100**
Happy (106)	38.7	73.6	88.7	87.7	95.3	**98.1**	97.2
Sad (66)	68.2	77.3	89.4	98.5	95.5	100	**100**
Surp. (71)	31.0	66.2	90.1	95.8	98.6	98.6	**100**
Expression by facial action unit							
LFAU (1549)	72.8	83.5	91.7	94.3	97.2	98.8	**99.5**
UFAU (432)	81.9	85.7	92.8	98.8	99.1	100	**100**
CAU (169)	60.4	81.1	94.1	96.5	98.8	100	**100**
Poses in yaw, pitch and combination							
YR (735)	7.8	18.2	23.8	50.9	78.0	84.1	**92.4**
PR (419)	50.8	59.2	73.5	98.1	98.8	99.5	**99.5**
CR (419)	4.7	13.7	19.0	62.6	94.3	–	**96.7**
Disguise/occlusion							
O (381)	28.9	51.70	74.0	77.7	**99.2**	99.2	91.1
All probes							
All (4561)	52.3	63.9	73.9	87.9	94.1	96.6	**97.6**

because of two reasons. Firstly, it contains separate variations in expressions (E), poses (P) up to 90° and illumination (I). Secondly, it contains two types of combined variations: expression variations with illumination from the right side (E+I) and pose variations with a smiling expression (E+P). Some sample images are shown in Fig. 14. Notice that due to the use of fast mode of Minolta, the 3D models are not as accurate as those in Bosphorus or FRGC databases. However, they are far better than those acquired by Kinect.

All images in the CASIA database with ≤60° pose and without glasses are labeled with nose tip position using their nose tip detection algorithm. We manually label the nose tip in the remanning images. We follow Xu et al.'s [53] experimental protocol and use 759 images from the last 23 subjects for training and the first images of the remaining 100 subjects for gallery. Unlike Xu et al., we use all the remaining images as probes whereas Xu et al. exclude probes with >60° pose and those wearing glasses. We use the training set to derive DCS and DNM transforms and the gallery set to form the coding dictionary. Similar to the case of Bosphorus database, we synthesize five samples from each gallery image as described in Section 7.2. The rank-1 identification rates are reported in Table 7 and a similar pattern to former experimental results can be observed.

The proposed algorithm outperforms all other methods and achieves an average of 95.6% identification rate using all the probes and 97.5% when using Xu et al.'s limited probe set. No obvious drop in performance is observed across most of the variations. Even under the challenging case of profile view, we can maintain an accuracy of 81% without expression and 77% with a smiling expression. To the best of our knowledge, these are the best results reported for the CASIA databases under the first versus all protocol.

7.4. Performance on face recognition grand challenge database

This database contains 4007 images of 465 subjects[4] captured across multiple sessions and with various expressions. The defacto standard in FRGC for identification is the first versus all protocol. Fig. 15 shows the CMC curve of our algorithm. Although

[4] Confirmed in [8] and by the FRGC organizers that subject ID 04643 and 04783 are the same person.

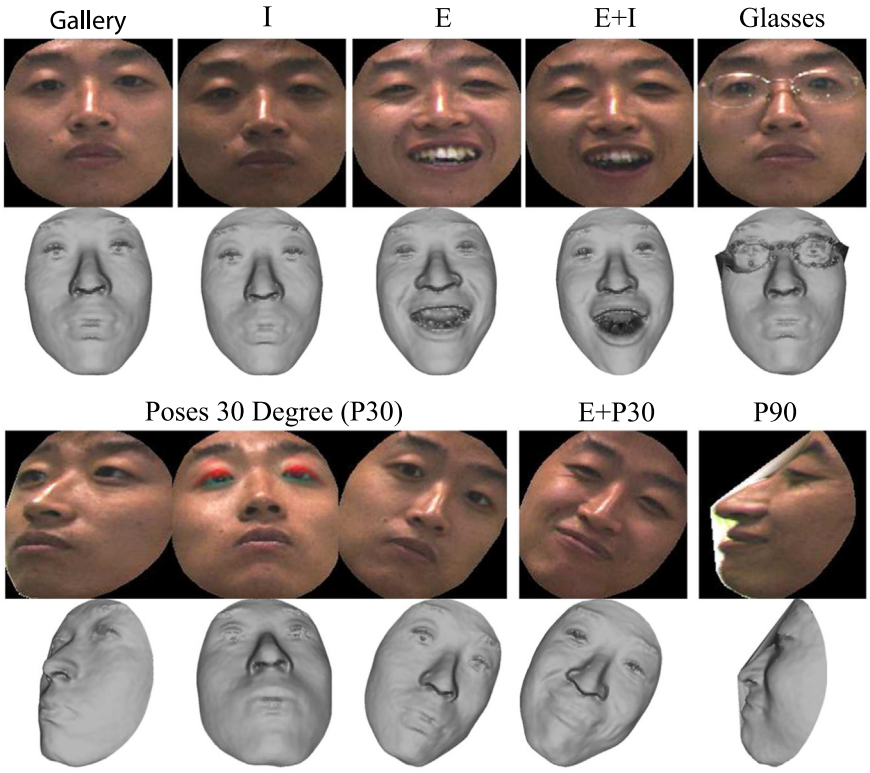


Fig. 14. Sample images of *CASIA*.

Table 7
Results on **CASIA** using first-neutral (100) vs. all (3663) protocol. Note the proposed method assumes that the nose tip is given.

Face variations	SVM	SRC	RSC	[7]	[53]	**Prop.**
I (400)	80.5	86.0	88.5	99.8	98.3	**99.8**
E (500)	57.6	74.4	76.6	94.2	90.0	**98.0**
E+I (500)	58.0	73.8	77.8	93.2	93.3	**97.8**
P30° (700)	20.3	26.7	36.3	96.1	91.0	**98.3**
P60° (200)	1.0	4	6	47	91.0	**96.5**
P90° (200)	1.0	0.5	1.5	5	–	**81.0**
E+P30° (700)	19.0	26.4	34.9	93.7	87.9	**96.0**
E+P60° (200)	1.0	4	5.5	45.5	79.0	**95.0**
E+P90° (200)	1.0	2.5	2.0	4.5	–	**77.0**
Glasses (63)	54.0	79.4	82.5	95.3	–	**100**
All (3663)	33.2	41.7	46.6	80.0	–	**95.6**
All Xu's(3200)	36.8	46.0	51.5	89.1	90.7	**97.5**

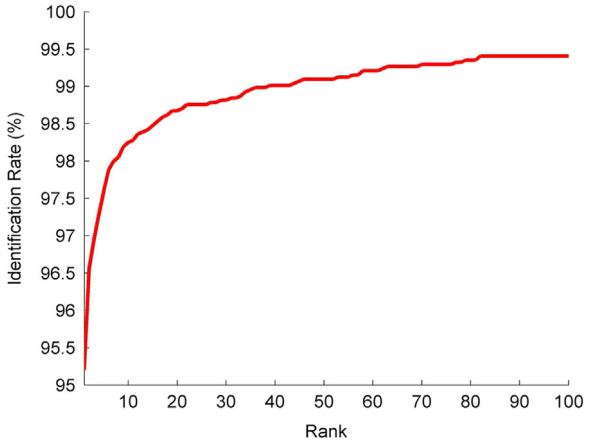


Fig. 15. CMC curve for first-neutral (465) vs. all (3542) protocol on *FRGC*.

synthesized gallery images are generated in a similar way (see Section 7.2), we only use them in the last iteration of our multi-channel weighted sparse coding to avoid the computational cost associated with large dictionary ($465 \times 5 = 2325$).

The proposed algorithm achieves a rank-1 identification rate of 95.2% which is comparable to the state-of-the-art [7–9,20]. Although some existing 3D methods in the literature report higher performance compared to our algorithm, most of them are evaluated only on the FRGC dataset. These methods may have been optimized for high resolution data and it is difficult to say how they will scale to low resolution data. Moreover, the FRGC database mostly contain expression variations, no occlusions and only minor pose variations. Therefore, it is difficult to perceive the robustness of these algorithms to occlusions and pose variations. Furthermore, the de-facto standard protocol allowing only a single gallery image per subject does not favor our proposed algorithm. In practical applications, multiple 3D scans of the same subject can easily be obtained during enrollment. Nevertheless, our results show that our method does not fail even under the unfavorable situation of single gallery image per subject.

7.5. Time complexity

The proposed algorithm was implemented on an Intel Core2 Quad 3 GHz CPU with 4 GB RAM using a 64-bit Matlab. It takes a total of 15 s to identify one query face from a gallery of 100 images. The time increases to 25 s for a larger dictionary of 2325 images. The average time per match is less than 0.5 s. Most of the time is taken by ICP based registration. We used up to 30 ICP iterations to achieve fine registration in our experiments which takes 10 s on the average. This can be speeded up to just 2 s by using only 5 ICP iterations which causes little drop in accuracy and can correct the pose even in the case of profile faces. Symmetric filling step requires searching for nearest neighbor points which takes 0.5 s. The histogram equalization takes 2 s. The DCS and DNM are linear transformations and take less than a millisecond.

Sparse coding is performed using the SPAMS [54] package with Matlab interface. Despite the overhead of Matlab function calls, SPAMS is able to return the solution in 0.05 s on a dictionary of 100 images with 3072 ($32 \times 32 \times 3$) feature dimension. Note that our Multi-channel Weighted Sparse Coding (MWSC) computes a mask of weights for each pixel iteratively. With our proposed parameter setting, about 40% of pixels receive a weight $< = 0.001$ and 10% of the pixels are located outside of the face after resampling to a square grid. These pixels are removed and therefore, the effective feature dimension is around 1100. Furthermore, about 4.5 iterations are required on the average for MWSC to converge and we apply MWSC on both 2D and 3D images. Taking all these into account, a total of 0.3 s are required to recognize one face in the Bosphorus and CASIA databases with a gallery size of 105 and 100 respectively. When synthesized images are generated, the dictionary size increases five times and the algorithm converges in 2 s. For CurtinFaces with 936 gallery images (dictionary size), a total of 3 s are required. For FRGC, as mentioned in Section 7.4, the synthesized gallery images are used only in last iteration. A total of 9.5 s are needed because the last iteration on a gallery of 2325 images consumes most of the time (about 8 s). Lastly, computation of the class-wise distance requires less than 0.05 s in all cases.

Note that excluding the sparse coding step, our algorithm is implemented fully in Matlab which makes it slow. Moreover, for consistency and comparison with existing 3D face recognition techniques, we focused on achieving higher accuracy at the cost of computational complexity. Implementation in a faster programming language will considerably increase the speed of our algorithm given the iterative nature of many components of our algorithm. Further, higher speed can also be achieved by algorithmic changes such as fewer ICP iterations and smaller dictionary sizes albeit at the cost of minor drop in accuracy.

8. Conclusion

We proposed a practical algorithm for robust face recognition that works equally well on low and high resolution RGB-D data. The proposed canonical preprocessing algorithm can correct poses and estimate the full frontal view from even the profile view of a face. The proposed multi-channel Discriminant Transform (mDT) can increase the discriminative power of any multi-channel data. Our results show that the proposed multi-channel Weighted Sparse Coding (MWSC) method, which computes variable weightings for sparse coding using multiple channels, is better than single channel in terms of robustness to variation in imaging conditions.

Although Kinect is low cost and high speed, the depth data it provides is in very low resolution. We analyzed and experimentally showed that existing 3D face recognition methods designed for high resolution 3D data are not suitable for such low resolution data. Using our proposed approach, accurate face recognition can still be performed when using low resolution Kinect data. Furthermore, our method outperforms existing techniques even on high resolution data under simultaneous variations in imaging conditions. The proposed face recognition algorithm can maintain consistent performance under simultaneous variations in pose, expression, illumination and disguise. Our method can also handle faces in challenging profile view condition to some extend. More importantly, no parameter tuning is required to achieve satisfactory performance in arbitrary cases. Finally, we have reported state-of-the-art results on the CurtinFaces, Bosphours and CASIA databases.

In future, we will investigate how each component in the proposed approach can affect the performance as well as how the training sample selection can have a significant impact on the performance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neucom.2016.06.012.

References

[1] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.

[3] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 625–632.

[4] R. Gross, I. Matthews, S. Baker, Appearance-based face recognition and light-fields, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4) (2004) 449–465.

[5] A. Sharma, D. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 593–600.

[6] G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, I. Kakadiaris, Bidirectional relighting for 3d-aided 2d face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2721–2728.

[7] A. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2d-3d hybrid approach to automatic face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1927–1943.

[8] C. Queirolo, L. Silva, O. Bellon, M. Segundo, 3d face recognition using simulated annealing and the surface interpenetration measure, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 206–219.

[9] L. Spreeuwers, Fast and accurate 3d face recognition, Int. J. Comput. Vis. 93 (3) (2011) 389–414.

[10] Y. Lei, M. Bennamoun, A. El-Sallam, An efficient 3d face recognition approach based on the fusion of novel local low-level features, Pattern Recognit. 46 (1) (2013) 24–37.

[11] H. Li, D. Huang, J.-M. Morvan, L. Chen, Y. Wang, Expression-robust 3d face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns, Neurocomputing 133 (2014) 179–193.

[12] Y. Ming, Rigid-area orthogonal spectral regression for efficient 3d face recognition, Neurocomputing 129 (2014) 445–457.

[13] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 947–954.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, Bosphorus database for 3d face analysis, Biomet. Identity Manag. (2008) 47–56.

[15] K. Khoshelham, S. Elberink, Accuracy and resolution of kinect depth data for indoor mapping applications, Sensors 12 (2) (2012) 1437–1454.

[16] B.Y. Li, A. Mian, W. Liu, A. Krishna, Using kinect for face recognition under varying poses, expressions, illumination and disguise, in: IEEE Workshop on Applications of Computer Vision, 2013, pp. 345 –352.

[17] Casia-3d facev1, ⟨http://biometrics.idealtest.org/⟩, 2004.

[18] K. Bowyer, K. Chang, P. Flynn, A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition, Comput. Vis. Image Underst. 101 (1) (2006) 1–15.

[19] P. Besl, N. McKay, A method for registration of 3-d shapes, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (1992) 239–256.

[20] T. Faltemier, K. Bowyer, P. Flynn, A region ensemble for 3-d face recognition, IEEE Trans. Inf. Forens. Security 3 (1) (2008) 62–73.

[21] A. Bronstein, M. Bronstein, R. Kimmel, Expression-invariant representations of faces, IEEE Trans. Image Process. 16 (1) (2007) 188–197.

[22] I. Kakadiaris, G. Passalis, G. Toderici, M. Murtuza, Y. Lu, N. Karampatziakis, T. Theoharis, Three-dimensional face recognition in the presence of facial expressions: an annotated deformable model approach, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 640–649.

[23] G. Passalis, P. Perakis, T. Theoharis, I. Kakadiaris, Using facial symmetry to handle pose variations in real-world 3d face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (10) (2011) 1938–1951.

[24] Y. Wang, J. Liu, X. Tang, Robust 3d face recognition by local shape difference boosting, IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1858–1870.

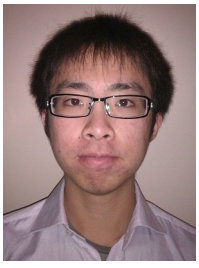
[25] N. Alyuz, B. Gokberk, L. Spreeuwers, R. Veldhuis, L. Akarun, Robust 3d face recognition in the presence of realistic occlusions, in: IAPR International Conference on Biometrics, 2012, pp. 111–118.

[26] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3d face recognition under expressions, occlusions, and pose variations, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2270–2283.

[27] N. Alyuz, B. Gokberk, L. Akarun, A 3d face recognition system for expression and occlusion invariance, in: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008, BTAS 2008, IEEE, Crystal City, Washington DC, USA, 2008, pp. 1–7.

[28] S. Berretti, N. Werghi, A. Del Bimbo, P. Pala, Matching 3d face scans using

interest points and local histogram descriptors, Comput. Graph. 37 (5) (2013) 509–525.
[29] D. Smeets, J. Keustermans, D. Vandermeulen, P. Suetens, meshSIFT: local surface features for 3d face recognition under expression variations and partial data, Comput. Vis. Image Underst. 117 (2) (2013) 158–169.
[30] N. Alyüz, B. Gökberk, L. Akarun, Regional registration for expression resistant 3-d face recognition, IEEE Trans. Inf. Forens. Security 5 (3) (2010) 425–440.
[31] A. Colombo, C. Cusano, R. Schettini, Three-dimensional occlusion detection and restoration of partially occluded faces, J. Math. Imaging Vis. 40 (1) (2011) 105–119.
[32] O. Ocegueda, G. Passalis, T. Theoharis, S. K. Shah, I. Kakadiaris, et al., Ur3d-c: Linear dimensionality reduction for efficient 3d face recognition, in: 2011 International Joint Conference on Biometrics (IJCB), IEEE, Crystal City, Washington DC, USA, 2011, pp. 1–6.
[33] D. Huang, M. Ardabilian, Y. Wang, L. Chen, 3-d face recognition using elbp-based facial description and local feature hybrid matching, IEEE Trans. Inf. Forens. Security 7 (5) (2012) 1551–1565.
[34] A. Mian, Illumination invariant recognition and 3d reconstruction of faces using desktop optics, Opt. Express 19 (8) (2011) 7491–7506.
[35] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics Gems IV, Academic Press Professional, Inc., San Diego, CA, USA, 1994, pp. 474–485.
[36] J. Yang, C. Liu, J. Yu Yang, What kind of color spaces is suitable for color face recognition? Neurocomputing 73 (10–12) (2010) 2140–2146.
[37] F. Al-Osaimi, M. Bennamoun, A. Mian, Spatially optimized data-level fusion of texture and shape for face recognition, IEEE Trans. Image Process. 21 (2) (2012) 859–872.
[38] A. Yip, P. Sinha, Role of color in face recognition, J. Vis. 2 (7) (2001).
[39] J. Yang, C. Liu, L. Zhang, Color space normalization: enhancing the discriminating power of color spaces for face recognition, Pattern Recognit. 43 (4) (2010) 1454–1466.
[40] J. Yang, C. Liu, Color image discriminant models and algorithms for face recognition, IEEE Trans. Neural Netw. 19 (12) (2008) 2088–2098.
[41] S.-J. Wang, J. Yang, N. Zhang, C.-G. Zhou, Tensor discriminant color space for face recognition, IEEE Trans. Image Process. 20 (9) (2011) 2490–2501.
[42] R. He, W. Zheng, B. Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2010) 1561–1576.
[43] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognit. 38 (12) (2005) 2270–2285.
[44] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27.
[45] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3d facial expression database for facial behavior research, in: International Conference on Automatic Face and Gesture Recognition, 2006, pp. 211–216.
[46] T. Faltemier, K. Bowyer, P. Flynn, Using a multi-instance enrollment representation to improve 3d face recognition, in: IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–6.
[47] R. Min, N. Kose, J.-L. Dugelay, Kinectfacedb: a kinect database for face recognition, IEEE Trans. Syst. Man Cybern.: Syst. 44 (11) (2014) 1534–1548.
[48] M. Hayat, M. Bennamoun, A.A. El-Sallam, An rgb-d based image set classification for robust face recognition from kinect data, Neurocomputing 171 (2016) 889–900.
[49] A. Savran, B. Sankur, M. Taha Bilge, Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units, Pattern Recognit. 45 (2) (2012) 767–782.
[50] H. Li, D. Huang, P. Lemaire, J. Morvan, L. Chen, Expression robust 3d face recognition via mesh-based histograms of multiple order surface differential quantities, in: IEEE International Conference on Image Processing, 2011, pp. 3053–3056.
[51] H. Li, D. Huang, J.-M. Morvan, Y. Wang, L. Chen, Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors, Int. J. Comput. Vis. 113 (2) (2014) 128–142.
[52] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, 2011, pp. 471–478.
[53] C. Xu, S. Li, T. Tan, L. Quan, Automatic 3d face recognition from depth and intensity Gabor features, Pattern Recognit. 42 (9) (2009) 1895–1905.
[54] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, J. Mach. Learn. Res. 11 (2010) 19–60.



Billy YL Li received the Bachelor of Science degree with first class honors in Computer Science from Curtin University in 2009 and he obtained his Ph.D. degree in 2013. His research area is pattern recognition and face recognition.

Mingliang Xue received the B.Sc. degree in Automation from Dalian University of Technology, PR China, in 2008, the M.Sc. degree in Control theory and Control Engineering from Dalian University of Technology, PR China, in 2011, and the Ph.D. degree in Computing from Curtin University, Australia, in 2015. His current research interests include pattern recognition, image processing, 3D facial expression analysis, computer vision.

Ajmal Mian completed his Ph.D. degree from The University of Western Australia in 2006 with distinction and received the Australasian Distinguished Doctoral Dissertation Award from Computing Research and Education Association of Australasia. He received two prestigious nationally competitive fellowships namely the Australian Postdoctoral Fellowship in 2008 and the Australian Research Fellowship in 2011. He received the UWA Outstanding Young Investigator Award in 2011, the West Australian Early Career Scientist of the Year award in 2012 and the Vice-Chancellor's Mid-Career Research Award in 2014. He has secured seven Australian Research Council (ARC) grants and one National Health and Medical Research (NHMRC) grant. He is currently with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include computer vision, machine learning, action recognition, 3D shape analysis, 3D facial morphometrics, hyperspectral image analysis and biometrics.

Wanquan Liu received the B.Sc. degree in Applied Mathematics from Qufu Normal University, PR China, in 1985, the M.Sc. degree in Control Theory and Operation Research from Chinese Academy of Science in 1988, and the Ph.D. degree in Electrical Engineering from Shanghai Jiaotong University, in 1993. He once holds the ARC Fellowship and JSPS Fellowship and attracted research funds from different resources. He is currently an Associate Professor in the Department of Computing at Curtin University. His research interests include large scale pattern recognition, control systems, signal processing, machine learning, and intelligent systems.

Aneesh Krishna is currently a Senior Lecturer of Software Engineering with the Department of Computing, Curtin University, Australia, since July 2009. He holds a Ph.D. in Software Engineering from the University of Wollongong, Australia, an M.Sc. (Engg.) in Electronics Engineering from Aligarh Muslim University, India, and a B.E. degree in Electronics Engineering from Bangalore University, India. He was a Lecturer in Software Engineering at the School of Computer Science & Software Engineering, University of Wollongong, Australia (from February 2006 to June 2009). His research interests include Software Engineering, Requirements Engineering, Conceptual Modeling, Agent Systems, Formal Methods and Service-Oriented Computing.