ROBUST MULTIMODAL FACE RECOGNITION WITH PRE-PROCESSED KINECT RGB-D IMAGES

Nasir Ahmad[1*], Jehad Ali[2], Khalil Khan[3], Muhammad Naeem[4], Usman Ali[5]

ABSTRACT

Researchers have tried to improve the accuracy of face recognition by combining 2D and 3D images to overcome the problems of illumination, pose variation and occlusion. Although combining 2D with 3D have shown better results when compared with 2D images only, however applicability of these methods is inadequate in practical implementations due to high cost of 3D sensors, therefore we are using the low cost sensor Kinect acquired images. We do face recognition from RGB images, depth images and then we combine both RGB and depth maps i.e. concatenate different Modalities to improve the accuracy of recognition. Depth maps have holes and noise induced from camera sensors, therefore we process them to remove these distortions and then we apply the face recognition algorithm. Experimental results reveal that the accuracy of face recognition can be increased by combining RGB and depth images and applying pre-processing on depth maps which mitigate the effects of covariates such as holes and noise in the depth maps.

KEYWORDS: *Face recognition, features extraction, classification, depth images, pre-processing.*

INTRODUCTION

Face recognition or identification is a challenging task which not only undergoes from the challenges of the object recognition in general i.e. viewpoint variations and illumination; but also suffers from distortions concerning faces e.g. accessories, expression, occlusion, pose variations and high inter-class similarity among human faces[1]. Generally 2D images are used for faces recognition, which have relatively less information about a face due to lack of depth information. Consequently, the researchers are now considering the usage of 3D information acquired via special type of sensors such as Kinect[2,3]. Although, the integration of depth information has proved to improve the accuracy of face recognition as compared to using 2D images only, the high cost and noise induced by these sensors has hindered the use of these approaches in commercial application. However, with the advent of latest sensing technology i.e. Kinect, a 2D RGB colour image along with its depth map (D) can now be now obtained using its less expensive sensors. A depth map of an image gives per pixel information concerning the depth of that image, which is extracted by using a laser projector (infrared) in conjunction with a typical camera.

Concatenated RGB-D images, which are RGB images complemented with the depth maps, have been used for several tasks i.e. modelling indoor environments, object recognition, tracking, surface modelling, and robotic vision[4-6]. In recent works[7,8], there is a detailed discussion on the applicability of RGB-D concatenated face images intended for gender recognition and face detection. An algorithm which uses concatenated RGB-D images in face recognition having various covariates is described in Li et al.[9]. Depth maps obtained from Kinect show high inter-class resemblance (due to holes and noise); therefore, depth maps may not be capable to differentiate different individuals from each other. On the other hand, a depth map has a very low intra-class difference or variation that can be exploit in increasing robustness to several covariates i.e. pose and expression. Further, RGB colour images provide high inter-class dissimilarity or differentiability that is required for depth data. Consequently, it is essential to make use of both depth and RGB data in classification and feature extraction. Face recognition is affected from inter-class differentiability of RGB images and inter-class similarity of depth images. Besides this the holes and noise also adversely affect depth images (especially in the areas where the normal is almost at 900 to the view of the camera, a considerable portion of depth measurement is usually not well-represented in depth maps), which also degrades the accuracy of face recognition. Therefore there is a need to remove these covariates such as noise and holes from depth images for improving the face recognition accuracy while RGB-D

1 Loughborough University, UK.*
2 Computer Engineering Department, AJOU University, Korea.
3 The University of Poonch, Azad Kashmir, Pakistan.
4 University of Peshawar, Peshawar, Pakistan.
5 Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan

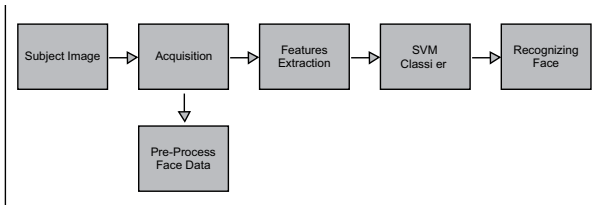

Figure 1: Proposed System for Face recognition

images are used.

Structure of Face Recognition System

A face recognition framework generally can be organized into four main steps; acquiring the face image, extraction of features from face, classification and recognition of face.

Our proposed structure of the face recognition system adds pre-processing to the existing system to combat the effect of noise and holes in Kinect depth images. Figure 1 shows our proposed structure.

Following are the steps of the proposed method for face recognition.

Acquisition of Face Data

First step for a face recognition system is getting the face data and processing the data if necessary. Face images or face data can be acquired from different sources. The sources can vary according to the needs of the user such as a camera, Kinect or readily available face image databases or datasets on the websites that deals with face recognition for commercial and research purpose. Sometimes when we process a face database it gets fix one feature but it can cause serious affects on other features which can degrade the performance of face recognition system. Therefore input image is processed and if it causes some serious affects and distort the image, then for the removal of that distortion we apply transformation methods on the input image under our consideration[10].

Pre-Processing the Data

Depth images contain many artificial distortions such as noise and areas for which no data exists. Repairing these regions of missing data is often a pre-requisite for many image processing algorithms and therefore there is a need to process this data before applying face recognition algorithm. When the acquired data is processed the accuracy and performance of the face recognition algorithm increases a lot. Therefore, we added the pre-processing step to the conventional steps to improve the accuracy of face recognition.

Extracting Face Features

The goal of the feature extraction is to take out the information or feature vectors for representing a face. Feature extraction is the process of obtaining the most related information from a facial image. A biometric reference is created by utilizing a mathematical representation which is later on stored in face database. Later on these extracted features are used in recognition process. The algorithm used for extracting these feature vectors is the Principal Component Analysis (PCA) in the proposed method.

Classification and Recognition of Faces

After the extraction and selection of features, in the next step the image is classified. In classification step, after all the face images in database, the resemblance between face images from the individual belonging to same and different classes are the similarity between faces from the same individual and different individuals are represented with relevant features. The outcome of the classification is determined by matching the user index with the user identity stored in the database for face recognition determination.

Face reduction step involves the decomposition and compression of the original extracted features but also not to demolish the utmost prominent information. Classification in the proposed method is performed with multi class Support Vector Machine (SVM).

Pre-Processing of Depth Images

Main problem with depth images of the Kinect device is the holes included in these images and the black spots present on these images. To deal with these problems, pre-processing steps are adapted which are in the following paragraphs.

Zero Elimination Median Filtering

We recursively apply the zero-elimination median filter on the given images to remove the holes from the given images followed by linear Interpolation. This hole-filling method is slightly different from the median filtering algorithm for hole-filling. The depth maps in data set we are using for testing correspond to holes as 0s. This method for hole filling applies a median filter repeatedly to the depth images, however holes (0s) in the depth images are not considered (the 0s are eliminated) when calculating the median. By applying Zero-Elimination, geometric distortion by the side of occlusion boundaries is reduced as compared to standard median filtering methodology[11].

Figure 2: Explanation of different facial images (a) smiling (b) neutral face (c) mouth open (d) glasses occlusion (e) illumination (f) hand occlusion (g) face with left profile (h) face with right profile (i) occlusion with paper.

3.2 Interpolation

A significant recent advancement in acquisition of depth map is the introduction of emerging economical fast cameras that measures the depth information[12]. Depth information is obtained at a high speed with these cameras that can be utilized in several applications. Typically, the depth maps captured from Kinect sensors have a lower resolution. Besides this, cameras for depth information sensing are normally sensitive to reflection from the objects, there is a considerable noise in the depth maps in those areas of the objects where reflectance from objects is low[13]. Moreover, depth maps obtained via Kinect devices have significant areas that have unreliable estimates of depth information; in addition, depth edges possibly will not be aligned to colour image. As a result, they need depth enhancement (DE). DE i.e. a depth map up-sampling phase (DU) in which a depth map of an image is up-sampled, further it is de-noised and then it is aligned with edges of a high resolution RGB image. The Kinect sensor captures depth maps which are usually characterize with a high noise level and, due to this reason; various approaches have been introduced for improving their level of accuracy. Techniques such as Gaussian filtering along with other approaches for smoothing show poor results particularly in areas having depth discontinuities in which the depth map appears apparently blurred; due to this reason, therefore several approaches depending on techniques in which edge is preserved are applied in literature, for avoiding the blurring phenomenon[14]. We perform re-sampling to achieve the two main objectives. Its first objective is that it smoothes the noise affected surface obtained via Kinect sensor and due to recursive filling. The second objective of re-sampling is that it fills the holes that are still there after the zero-elimination median filling. By using the smoothing factor, Surface fitting can be applied which does not let this surface to abruptly bending and therefore minimizing the noise and outliers effects. For every face image, 256 × 256 points are uniformly re-sampled from X – Y lower co-ordinates to X – Y high co-ordinates. The benefit of re-sampling, beginning from the low co-ordinates to the high co-ordinates is that the face can be aligned now on the two dimensional grid. Since the RGB image is noise free therefore there is no need for smoothing the RGB, because smoothing will bring together blurring in it. Therefore, we only resample the RGB image to the corresponding X-Y location by means of interpolation.

Experimental Setup

Dataset

We have chosen the Eurecom face dataset for experimental work. The dataset is composed of multimodal facial images of 52 persons (38 males, 14 females). These images were captured by the low cost sensor i.e. Kinect. The images from the dataset used in this work are the RGB colour images and their depth images. Facial images for every person in this dataset consist of nine states i.e. with varying facial expressions, varying lighting, occlusion, smile mouth, neutral, open mouth, left and right profile, occlusion eyes, occlusion with paper, light on and occlusion mouth as shown in Figure 2. All images in the nine categories which are present in the dataset have been used in the experiments[15]. Figure 3 show the depth maps of the corresponding RGB images.

The major components of face recognition are feature extraction and classification. In supervised learning techniques there are two main phases i.e., training and testing stages. Some part of the data is placed in training phase while remaining data is placed in testing phase. We validated our experimental work with 10-fold cross validation. In such kind of experiments 10 % data is used for testing while remaining 90 % is used for testing. By this way there is no repetition of the training data in testing phase and experiments are also validated thoroughly.

RESULTS AND DISCUSSION

In first phase we performed experiments on RGB, depth images and by concatenating the RGB and depth modalities. The results are listed below in the Figure 4. Figure depicts that by combining RGB and depth modalities the accuracy of face recognition algorithm has increased. This is because of the fact that RGB provides inter-class similarity while depth data provides intra-class similarity, so when the two modalities were combined; the accuracy increased as compared to the individual RGB and depth modalities.

In the next phase depth images were processed and by applying recursive zero-elimination median filtering the holes were filled and noise was removed from them. Then these depth images were concatenated with RGB

to obtain RGB-D images and then face recognition was performed. To improve the resolution and further smooth images, up-sampling was performed after the application of zero-elimination median filter.

Figure 4 shows the overall comparison of face recognition accuracy for isolated and concatenated modalities. From the figure it is clear that the lowest accuracy for recognition is reported for depth images alone. When RGB and depth were concatenated, increase in face recognition accuracy was noted. Similarly when pre-processed RGB-D images were tested, improvement in face recognition accuracy was noted. However, the best results were shown with up-sampled pre-processed RGB and depth concatenated images.

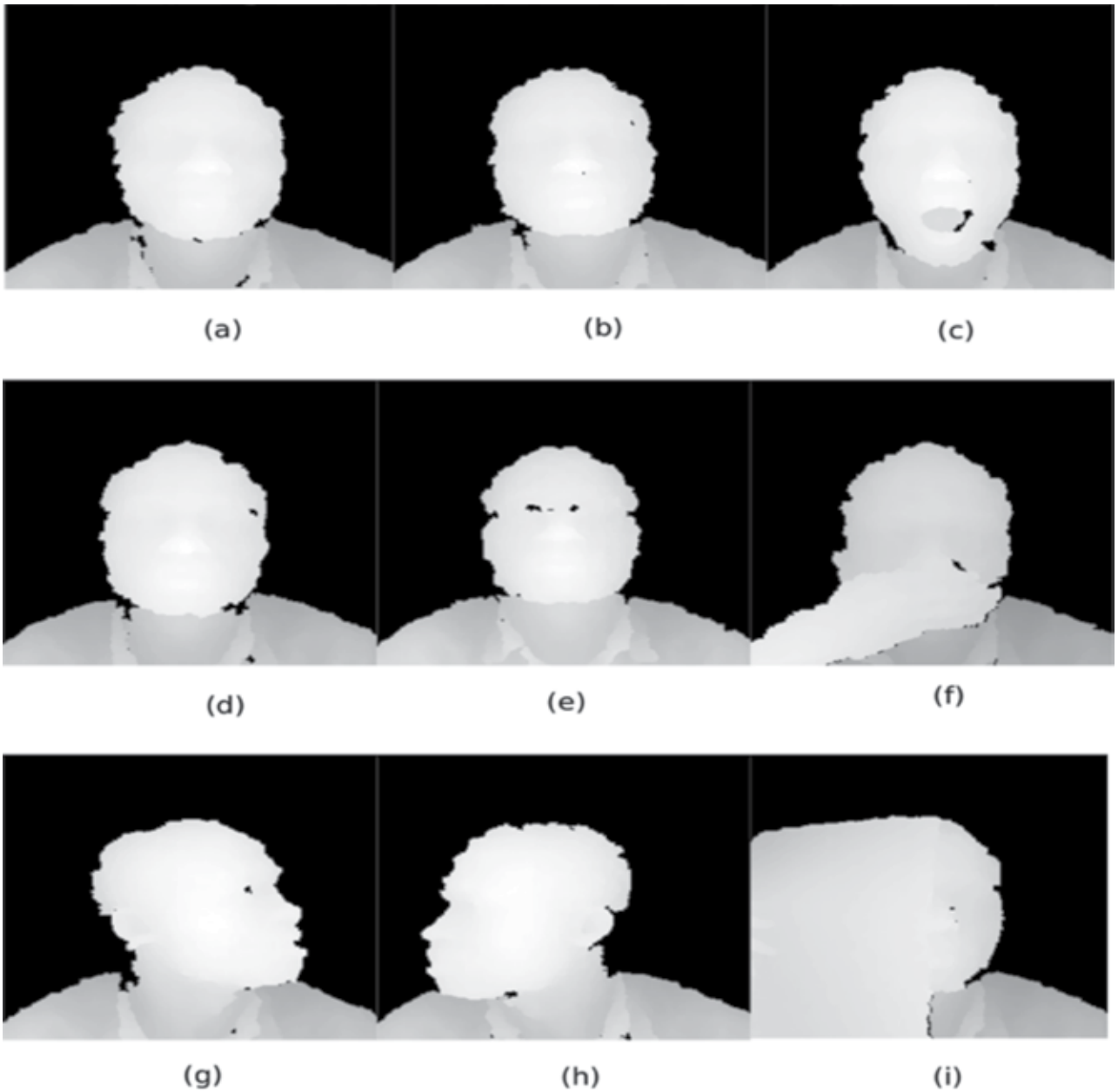

Figure 3: Explanation of depth facial images (a) smiling (b) neutral face (c) mouth open (d) glasses occlusion (e) illumination (f) hand occlusion (g) face with left profile (h) face with right profile (i) occlusion with paper

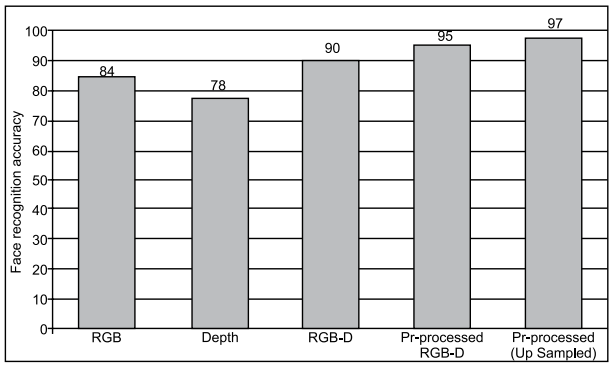

Figure 4: Face recognition accuracy in each modality

Conclusion and Future Work

A new algorithm for face recognition has been introduced in this paper which examines the face recognition accuracy of RGB, depth, RGB-D and then the pre-processed RGB-D images after filling the missing data in Kinect depth images. As RGB images provide inter-class differentiability and depth images have inter-class similarity, therefore depth images are not able to differentiate among different user face images. However, depth images have low intra class variations that provide robustness to pose and expression; therefore we combined these two modalities for feature extraction and classification. After combining the two modalities, we examined that face recognition accuracy improved. However due to holes in Kinect depth images the recognition rate noted was not optimum. When we filled the holes, removed the noise and performed up-sampling the recognition rate got improved. We concluded from our experiments that higher accuracy and robustness can be achieved by using multiple biometrics and removing holes from depth data rather than the costly and best possible sensor. In general the quality of 3D data acquired via Kinect device is perhaps not as reliable as 2D intensity data.

ACKNOWLEDGEMENT

Our special thanks to the European team for providing the Kinect face dataset.

REFERENCES

1. *Abate, A. F., Nappi, M., Riccio, D, Sabatino, G., (2007), "2D and 3D face recognition: A survey", Pattern Recognition Letters, Vol.28(14): pp.1885–1906.*

2. *Bowyer, K. W., Chang, K., Flynn, P., (2004), "A survey of approaches to three-dimensional face recognition", IEEE 17th International Conference on Pattern Recognition, Vol.1: pp.358–361.*

3. *Scheenstra, A., Ruifrok, A., Veltkamp, R. C., (2005), "A survey of 3D face recognition methods", International Conference on Audio-and Video-based Biometric Person Authentication, Springer Berlin Heidelberg, pp.891–899.*

4. *Bo, L., Ren, X., Fox, D. (2011), "Depth kernel descriptors for object recognition" IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.821–826.*

5. *Park, Y., Lepetit, V, Woo, W., (2011), "Texture-less object tracking with online training using an RGB-D camera", IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp.121–126.*

6. *Ramey, A., González-Pacheco, V., Salichs, M. A., (2011), "Integration of a low-cost RGB-D sensor in a social robot for gesture recognition" 6th international conference on Human-robot interaction, pp.229–230.*

7. *Hg, R. I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T. B., Tranchet, G., (2012), "An RGB-d database using microsoft's Kinect for windows for face detection" SITIS, pp.42-46.*

8. *Huynh, T., Min, R., Dugelay, J. L., (2012), "An efficient LBP based descriptor for facial depth images applied to gender recognition using RGB-D face data", ACCV, pp.133-145.*

9. *Li, B. Y. L., Mian, A. S., Liu, W., Krishna, A., (2013), "Using kinect for face recognition under varying poses, expressions, illumination and disguise" WACV, pp.186-192.*

10. *Shermina, J., (2011), "Illumination Invariant Face Recognition Using Discrete Cosine Transform and Principal Component Analysis", International conference on Emerging Trendes in Electrical and*

Computer Technology (ICETECT), pp.826-830

11. *Solh, M., AlRegib, G., (2012), "Hierarchical Hole Filling (HHF): Depth Image Based Rendering without Depth Map Filtering for 3DTV" International Workshop on Multimedia Signal Processing (MMSP), pp.87-92.*

12. *Kolb, A., Barth, E., Koch, R., Larsen, R., (2009), "Time-of-Flight Sensors in Computer Graphics", Eurographics (STARs), pp.119-134*

13. *Bae, K., Kyung, K. M., Kim, T. C., (2011), "Depth upsampling method using the confidence map for a fusion of a high resolution color sensor and low resolution time-of-flight depth sensor", IS&T/SPIE Electronic Imaging, pp.786406-786406.*

14. *Camplani, M., Salgado, L., (2012), "Adaptive Spatio-Temporal filter for low-cost camera depth maps", IEEE Conference on Emerging Signal Processing Applications (ESPA), pp.33-36.*

15. *Min, R., Kose, N., Dugelay, J. L., (2014), "Kinect Face DB: A Kinect Database for Face Recognition", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol.44(11): pp.1534-1548.*