Unconstrained Face Detection and Recognition Based on RGB-D Camera for the Visually Impaired

Xiangdong Zhao, Kaiwei Wang, Kailun Yang, Weijian Hu
College of Optical Science and Engineering, Zhejiang University, Hangzhou 310027, P. R. China

ABSTRACT

It is highly important for visually impaired people (VIP) to be aware of human beings around themselves, so correctly recognizing people in VIP assisting apparatus provide great convenience. However, in classical face recognition technology, faces used in training and prediction procedures are usually frontal, and the procedures of acquiring face images require subjects to get close to the camera so that frontal face and illumination guaranteed. Meanwhile, labels of faces are defined manually rather than automatically. Most of the time, labels belonging to different classes need to be input one by one. It prevents assisting application for VIP with these constraints in practice. In this article, a face recognition system under unconstrained environment is proposed. Specifically, it doesn't require frontal pose or uniform illumination as required by previous algorithms. The attributes of this work lie in three aspects. First, a real time frontal-face synthesizing enhancement is implemented, and frontal faces help to increase recognition rate, which is proved with experiment results. Secondly, RGB-D camera plays a significant role in our system, from which both color and depth information are utilized to achieve real time face tracking which not only raises the detection rate but also gives an access to label faces automatically. Finally, we propose to use neural networks to train a face recognition system, and Principal Component Analysis (PCA) is applied to pre-refine the input data. This system is expected to provide convenient help for VIP to get familiar with others, and make an access for them to recognize people when the system is trained enough.

Keywords: Visually impaired assisting, face tracking, face recognition, Mean Shift, Neural Networks, RGB-D camera.

1. INTRODUCTION

Face recognition technology has become more and more applicable although both 2D and 3D methods require some inevitable conditions to perform well. According to WHO, there are 285 million VIP in the world. Some of them cannot name an approaching person, because of their poor sight, even if the person is one of their very familiar friends or relatives. It is worth trying to introduce face recognition technology into VIP assisting as it brings great convenience that helps the blind identify acquaintances and strangers.

Like most pattern recognition techniques which require the quality of train data, face recognition depends on head pose, illumination, complex training procedure, etc. Labels of train data mostly are defined manually. In this paper, we use Intel RGB-D camera RealSense [1] to get color and depth information. Color image is firstly processed to generate a primary face detection result, then both color and depth image are processed to achieve tracking face in the image field. Here we use Mean Shift method [2] for tracking. Thus, not only the detection rate of face is enhanced but also an access to labelling faces automatically is obtained. A class of face images are saved after face tracking. After that, these face images undergo frontal transformation enhanced from [3] which uniforms a non-frontal face to a frontal one, which is also uniformly illuminated. Intra-class differences caused by pose and illumination are significantly diminished after this transformation. Then, we propose to use PCA for dimensional reduction to these optimized face images in training data set.

Lower dimensional data serves as the input layer in an Artificial Neural Network (ANN) [4], while label data serves as the output layer. Back propagation algorithm [5] is used to train ANN. After the ANN training, if a new face image from the result of tracking need to be recognized, it will undergo frontal transformation, dimensional reduction, and then be input to ANN. Output data of ANN indicates the recognition result, that is, which trained class the face belongs to. It will also be decided that none of trained class a face belongs to if the face is stranger's.

A friendly interaction approach is proposed. We synthesis 3D sound to make the blind be aware of someone around them, and to tell them who this person is.

Eighth International Conference on Graphic and Image Processing (ICGIP 2016), edited by Tuan D. Pham, Vit Vozenilek, Zhu Zeng, Proc. of SPIE Vol. 10225, 1022509 · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2266122

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term

Naïve face recognition methods involve PCA Eigen-face [6], LDA Fisher-face [7], LBPH face [8]. These methods perform good effects although they are simple. Recently, as the development of deep learning, face recognition methods tend to be more complicated, such as ANN and CNN. Rabia Jafri and Hamid R. Arabnia [9] summarized face recognition technology as three classes: face recognition from intensity images, face recognition from video sequences, and face recognition from other sensory inputs. Face recognition from intensity images achieves high recognition accuracy but it is only suitable for face images that have been captured already. For our real-time oriented application, face recognition from video sequences is no doubt a better choice. Tracking method refines the detection rate and recognition accuracy as detailed in Section 2. Another recognition method whose data is captured from other sensory inputs, which saves more information such as depth information and thermal infrared image of a face. However, it suffers from high complexity and low frame rate (FPS). In this article, PCA and neural networks are combined to achieve high recognition accuracy in video sequences.

For face images in training set, the ideal condition is that intra-class differences are small and inter-class differences are large. Unfortunately, most of time faces detected from images tend to be profile and have non-uniform illumination. Thus, what dominate the classifier are not only inter-class differences among different people but also intra-class differences caused by non-frontal head pose or bad illumination. As a result, recognition given by classifier degrade to be not reliable. Several approaches have been proposed to solve this problem. Active Shape Model (ASM) [10] and Active Appearance Model (AAM) [11] are classical algorithms for modeling a face and marking key-points of a face, that is, face alignment. M. Haghighat et al. [12] proposed to synthesize frontal face with AAM and piece-wise affine warping. However, deformation of synthesized faces is aggravated while head pose becomes more non-frontal. L. Ding et al. [13] proposed an algorithm that densely maps each pixel in the face image onto the 3-D face model and rotates it to the frontal view, which is still not suitable for real time application because of large computational cost. T. Hassner et al. [3] proposed to synthesize frontal face images from a uniform 3D face model, and face alignment is applied to relate 2D image and 3D model. It is concluded that the textures of face are more important than the shape for recognition. Real time frontal face transformation enhanced from [3] is implemented in this work and will be presented in Section 2.

2. ENHANCED FACE DETECTION AND FRONTAL TRANSFORMATION

2.1 RGB-D Camera Based Face Detection and Tracking

We use Haar Feature [14], [15] to detect face in a color image, this is a simple start of the whole process for one frame. Single image detection naturally ignores the consistence of video sequence, and this is just what tracking strategy addresses. Most of the tracking strategies only depend on color images. We use Intel RGB-D camera RealSense r200, which can obtain available depth information that ranges from 0.5m to 5m. So another dimension's information could be used for tracking a face. In this section, a tracking algorithm based on both color and depth information is proposed.

Considering a human face in a depth frame, there are small outstanding regions whose depth values are extremely different from the background's. Fig. 1b shows a depth frame sample and Fig. 1a. is the corresponding color frame. This outstanding feature is what the tracking algorithm mainly depends on.

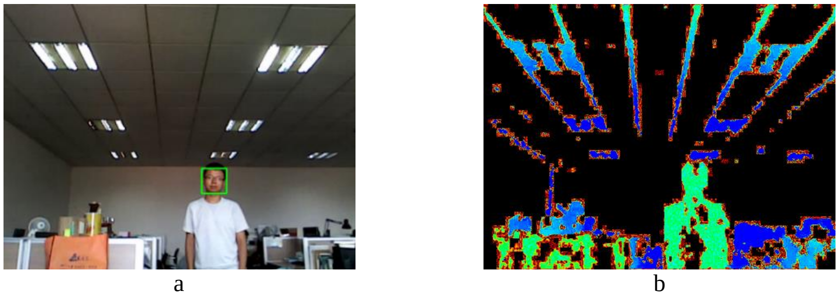


Figure 1. a Shows indoor color information while b is a corresponding depth image (pixel's hue value varies with depth)

If a face is detected in the current frame, noticing that this detection is based on only color information, face region (ROI) in the frame is considered as the start of tracking. Color histogram (hue value counted) and depth histogram of ROI are calculated. In the next frame, two back-projections are produced, including depth back-projection which are generated from depth frame where each pixel's value are substituted by its corresponding vertical coordinate in depth

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term

histogram and color back-projection which are generated from color frame where each pixel's hue value are substituted by its corresponding vertical coordinate in color histogram. These two back-projection then are fused according to specific weights. In near range color back-projection holds a large proportion while in far range depth back-projection holds a large proportion. Fusion is detailed with formula (1), where c and d are pixel's values in color back-projection and depth back-projection respectively, k is a coefficient for normalization.

$$fusion\ result = \ k \times [e^{-depth} \times c + (1 - e^{-depth}) \times d] \tag{1}$$

This fused image, we call it the probability distribution image, displays the probability of where a human face exists. Mean Shift algorithm is utilized on the fused image, and result of tracking, the most reliable region where a face exists, is obtained. A tracked Face is shown in Fig. 1a, surrounded by a green rectangle. Fig. 2 displays two back-projection images and the fused image.

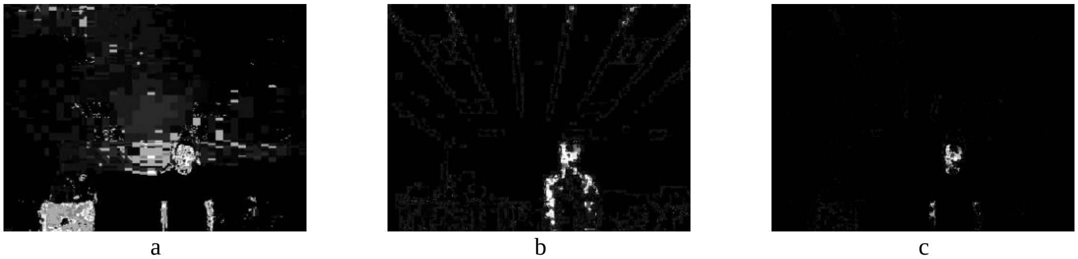


Figure 2. a is color back-projection, b is depth back-projection, and c is fused image from a and b

2.2 Frontal Face Transformation

A real-time frontal transformation upon a non-frontal face is proposed, which differs from [3], we use 78 face key-points obtained by RealSense-SDK [16], and achieve a face transformation in video sequences rather than in intensity image. A 3D face model is applied in this proceeding. Firstly, camera calibration is calculated according to 78 face key-points in 2D image and corresponding 3D points in 3D face model, this calibration naturally establishes the camera coordinate system. Then, all points in 3D face model are projected to camera coordinate system, and this projection result indicates the corresponding coordinates of 3D points in the 2D image. Finally, 3D points are substituted with RGB values in 2D image, and a frontal face image is synthetized by bilinear interpolation. The largest pose angle of non-frontal face to be transformed is 60°. The processing has a speed of 10~15 FPS on intel i5 CPU, which meets real time requirements for VIP assisting. Fig. 3 shows the faces before and after transformation. Obviously, intra-class differences have been considerably decreased. As a result, recognition accuracy has been raised from 80% for face images without transformation up to 95% with frontal faces.

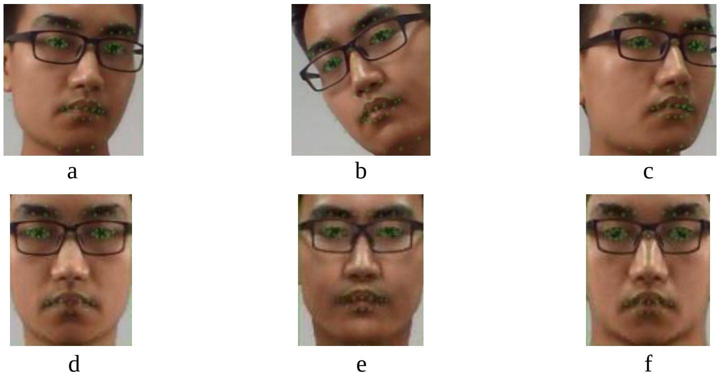


Figure 3. a, b, and c are faces detected in color images, while d, e, and f are the corresponding frontal transformed faces

3. FACE RECOGNITION AND INTERACTION

Face detection and transformation play the role of preparation of data set for the subsequent recognition. With automatic labeling, face recognition is regarded as a classification problem. Each data is a face image, uniformed as 100×100 pixels, and is transformed into a 1×10000 vector when calculated; each corresponding label is automatically set as a vector, length of which is the same as the number of subjects. For example, if there are 3 people to

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term

be recognized, label of first one will be set as [1, 0, 0], label of second one will be set as [0, 1, 0], similarly label of last one will be set as [0, 0, 1].

Neural network is applied in our system to train the face classifier, hence we have an input layer which size is 10000 and an output layer which size is the same as number of subjects. Such input layer is too large to be computed so that neural network is not available. Thus, dimensional reduction for input data becomes necessary. PCA [17] is a popular strategy to handle this problem. All face data is synthesized as a matrix, each row of which is a 1×10000 vector, transformed from a 100×100 image. For instance, if there are n people and each of whom has m face images, we index these n × m face images from 1 to k, named as $F_1, F_2, \dots, F_k$. Uniformed k vectors, sized as 1×10000, are combined as a matrix, that is, $[F_1, F_2, \dots, F_k]^T$. This matrix, named as M, is calculated with formula (2).

$$R = M \times M^T \tag{2}$$

R's size will be $k \times k$. Then, eigenvalues and eigenvectors of R are calculated. K eigenvectors, $E_1, E_2, \dots, E_k$, saved as column vectors, are combined in a matrix in descending order of corresponding eigenvalues. This matrix, named as E, is $[E_1, E_2, \dots, E_k]$. E is multiplied by transpose of M, as shown in formula (3).

$$M_{EF} = M^T \times E \tag{3}$$

We get matrix M_{EF}, that is 'EigenFaces' matrix. The size of M_{EF} is $10000 \times \mathrm{k}$, and each column of M_{EF} is an 'EigenFace', which represents a specific feature of face. Finally, dimensional reduction result is calculated with formula (4).

$$M_{DR} = F_i \times M_{EF} \tag{4}$$

F_i is uniformed face vectors, with a size of 1×10000, and M_{DR}, with a size of $1 \times \mathrm{k}$, is the dimensional reduction result of F_i. Each component of M_{DR} is the weight of corresponding feature in M_{EF}.

Dimension of face data is reduced from original 10000 to final k. Since k is the product of n and m, while n is usually no more than 10 and we set m as 20, k will be pretty smaller than 10000. On the one hand, input layer size of neural network is decreased; on the other hand, redundant information of face is diminished. After dimensional reduction with PCA, face data and label serve as input layer and output layer in neural network respectively, and parameters of neural network are trained with backpropagation algorithm (BP). When a new face image needs to be recognized, similarly it undergoes dimensional reduction, then neural network is calculated from input layer to output layer. Experiments are conducted on the Yale face database and we achieved recognition accuracy up to 95%. Faces under real circumstances had also been tested. Fig. 4 and Fig. 5 show how a neural network implements training and recognition respectively.

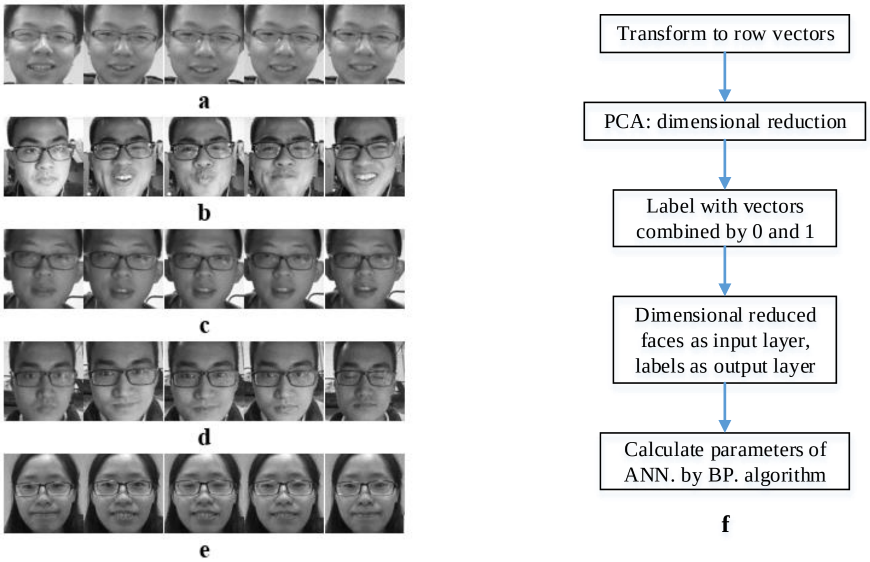
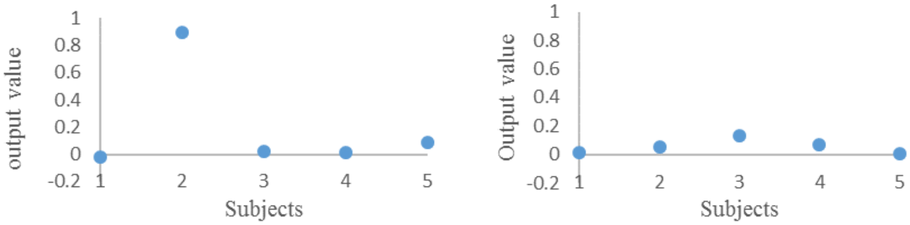

a
b
c
d
e
Transform to row vectors
PCA: dimensional reduction
Label with vectors combined by 0 and 1
Dimensional reduced faces as input layer, labels as output layer
Calculate parameters of ANN. by BP. algorithm
f

Figure 4. a, b, c, d, e are 5 subjects and each has 5 face samples. f shows the flow of data pre-refine, labeling and training

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term

In Fig. 5a, ANN is inputted with a test face image, and among the 5 output values, maximum is found at index 2, which indicates the test face image belongs to subjects 2. A threshold of 0.5 is set to estimate whether a test face image belongs to stranger. In Fig. 5b, output values are calculated from another test face image. Since none of 5 values is larger than 0.5, the test face image is decided to be stranger's.



a b

Figure 5. a is the output of ANN for a face image belonging to one of the trained 5 subjects, b is the output of ANN for a face image belonging to none of the trained 5 subjects, that is, stranger

An effective way to inform the visually impaired of detection and recognition result is adopted. We use stereo sound to implement this. Stereo sound is synthesized according to face's direction in the field of view and face's distance away from the user. Blind user who wears a headset can not only be aware of the existence and identities of people around himself (herself) but also perceive those people's directions and distances. Meanwhile, for the situation when two or more people's faces are detected, the nearest one will be informed to the blind, then the others.

4. CONCLUSION

Automatically detecting and recognizing facilitate the practical application of this apply face recognition technology for assisting the visually impaired. With RGB-D camera, face tracking using color and depth information have been implemented in order to achieve automatic and fast labeling. Meanwhile, it is proved that, compared with raw face images which may be non-frontal, frontal face images after transformation significantly increase the recognition accuracy. The system satisfies the need of real time detection and recognition. Friendly interaction approach by using stereo sound is proposed to provide convenience for users.

REFERENCES

[1] Intel Corporation, http://click.intel.com/intel-realsense-developer-kit-r200.html

[2] D. Comaniciu, V. Ramesh and P. Meer, "Real-Time tracking of non-rigid objects using mean shift," Proc. IEEE Conf. CVPR, (2015)

[3] T. Hassner, S. Harel, E. Paz and R. Enbar, "Effective face frontalization in unconstrained images," Proc. IEEE Conf. CVPR, 4295–4304, (2015)

[4] Stanford University, http://deeplearning.stanford.edu/wiki/index.php/Neural_Networks

[5] Y. LeCun, L. Bottou, G.B. Orr and K.R. Muller, "Efficient backprop," Neural Networks Tricks of the Trade, **1524**(1), 9-50, (1998)

[6] M. Turk and A. Pentland, "Face recognition using eigenfaces," Proc. IEEE Conf. CVPR, 586–591, (1991)

[7] P.N. Belhumeur, J. Hespanha and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE Trans. PAMI, **19**(7), 711–720, (1997)

[8] T. Ahonen, A. Hadid and M. Pietikainen, "Face recognition with local binary patterns," Computer Vision - ECCV, 469–481, (2004)

[9] R. Jafri and H.R. Arabnia, "A survey of face recognition techniques," Journal of Information Processing Systems, **5**(2), 41-68, (2009)

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term

[10] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham, “Active shape models-their training and application,” Computer Vision & Image Understanding, **61**(1), 38–59, (1995)

[11] T.F. Cootes, G.J. Edwards and C.J. Taylor, “Active appearance models,” ECCV, 2:484–498, (1998)

[12] M. Haghighat, M. Abdel-Mottaleb and W. Alhalabi, “Fully automatic face normalization and single sample face recognition in unconstrained environments,” Expert Systems with Applications **47**(5), 23-34, (2016)

[13] L. Ding, X. Ding and C. Fang, “Continuous pose normalization for pose-robust face recognition,” IEEE Signal Processing Letters **19**(11), 721-724, (2012)

[14] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” Proc. IEEE Conf. CVPR, (2001)

[15] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” IEEE ICIP, **1**, 900-903, (2002)

[16] Intel Corporation, https://software.intel.com/en-us/intel-realsense-sdk/

[17] M. Turk and A. Pendland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience **3**(1), 71-86 (1991)

Downloaded From: http://proceedings.spiedigitallibrary.org/pdfaccess.ashx?url=/data/conferences/spiep/91011/ on 02/14/2017 Terms of Use: http://spiedigitallibrary.org/ss/term