Database Construction & Recognition for Multi-view face

Won-Sook LEE
SITE, University of Ottawa, Canada
wslee@uottawa.ca

Kyung-Ah SOHN
HCI Lab, Samsung AIT, Korea
kasohn@sait.samsung.co.kr

Abstract

We present data collection and recognition experiment focused on multi-view face recognition/descriptor. Many face databases and face recognition systems have been constructed and experimented in terms of various illumination, time, poses, or expressions. However none of databases yet satisfies a large variation of poses to study systematic 3D human face information, which results unsatisfactory success rate for the posed face recognition while many quite satisfactory frontal view reconstructions have been shown. It is due to the difficulty of data collection of facial images to satisfy the large variation of poses to fully represent the 3D characteristic of human faces. We show two possible multi-view face data collection either using rendering of 3D models or using a video camera. We also illustrate our approach to build a face descriptor containing 3D information of human face using multi-view concepts. This multi-view face recognition descriptor is a 3D face descriptor which takes systematic extension of 2D face descriptor using the concept how much powerful a view influences over nearby views, so called as "quasi-view" size.

1. Introduction

One of the most significant factors making people's faces look different is the pose, or the viewing direction at which the face is seen. This is the basic reason that the typical face description algorithms based on a single view, especially frontal view, should have a fundamental limitation in its performance in a real situation. Some approaches such as novel view generation with given limited view [12] were proposed but they reveal limitation to obtain high success rate. This is because the face is a 3D-object and it cannot be reconstructed from 2D information. There are other approaches to describe the 3D characteristics of human face for recognition purpose such as building 3D model itself of the given person. Some approaches using 3D morphable models [9] show quite good result but due to its heavy computational costs, it is still unavailable in real markets. A 3D model reconstruction with stereoscopic camera is also used to retrieve the face identity [5], but the successful result has not come out yet.

Here, we propose to describe a face as a mosaic of many one-views [11] and show which criteria are used to extend the one-view to multi-view. By choosing certain appropriate views, we can contain 3D information of a face in a compact form. Here, we aim the multi-view face descriptor to contain the information of any view between horizontal rotation ranged in from $-90°$ to $90°$, say [$-90°...90°$], and vertical rotation [$-30°...30°$] for video images and [$-60°...60°$] for computer generated images. For this multi-view face descriptor, we need databases with wide pose variation and dense pose distribution. Many kinds of face databases already exist but none of them satisfies enough view range. Therefore we constructed our own face databases. Instead of building a complicated and high-cost studio for data collection, we used two simple methods, one with 3D mesh models and the other with a video camera.

In next sections, we briefly analyze the existing face databases and then describe multi-view face data collection procedure. Then we show experimental results with a 3D face descriptor which takes systematic extension of a 2D face descriptor to multi-view one using "quasi-view" concepts. Quasi-frontal is widely used term to show how faces in off-frontal views can be recognized by front view recognizers and we generalize this to general views using "Quasi-view".

2. Face database

2.1. Existing face databases

There have been various databases in several research groups in the world, but no database has been constructed covering a big range of angles. Good summarization can be found at the Peter Kruizinga's webpage [4] or in a book [6] by S. Gong et al.

Many 2D face images have been obtained in a studio built entirely for face images' sake. Some selected examples of face databases intended to cover poses and illumination are shown in Table 1. They set up cameras and lights in different positions to simulate pose and light variances, which produces satisfactory images in short time. Nevertheless, certain constraints exist due to the studio setup. The total number of poses equals to the actual number of camera, and the limited number of pose variance such as 7 to 13 pose images in these examples are far too small to perform experiments for multi-view face recognition.

KISA[10]	YALE[2]	CMU[8]
43 illumination combinations	64 light cones	21 flashes
total 7 poses (0°,0°) (±15°,0°) (±30°,0°) (±45°, 0°) 15° horizontal gap no vertical variance	total 9 poses (0°, ±12°) (9°, ±9°) (12°, 0°) (17°, ±17°) (24°, 0°) five poses at 12° and three poses at 24° from camera's axis	total 13 poses (0°, 0°) (±22.5°, 0°) (±45°, 0°) (±67.5°, 0°) (±90°, 0°) (0°, ±22.5°) 22.5° horizontal gap in on row, 2 cameras vertically up and down, 2 cameras to simulate surveillance camera
1000 people (52,000 images)	10 people (4500 images used)	68 people (41,368 images)

Table ***1****: Three examples of face DB are shown in terms of illumination and pose variances. Angles inside parenthesis are (horizontal rotation, vertical rotation).*

There have been two approaches to cover a large range of poses using artificial markers or using sensors and a calibrated cameras [6]. The approach using artificial markers is suffered by noisy between eye positions and head rotation. The other approach using a magnetic sensor and a camera still needs manual intervention, but seems good result to produce big range of views. Anyhow it still needs special purpose hardware equipment such as magnetic sensor and it has limitation to be accurate as magnetic sensor contains its own error.

2.2. Ideal face recognition

Ideal posed face recognition means to cover the wide cases in terms of people and environment, in other words intrinsic and extrinsic parameters. We focus on extrinsic parameters in this paper. For the pose, the training set must be complete in every view covering any view between horizontal rotation [-90^{o}… 90^{o}] and vertical rotation [-30^{o}… 30^{o}] ~ [-60^{o}… 60^{o}] as a view-mosaic in *Figure* **1** while registration can be done with only selected views. Background is better to be homogeneous to allow automatic segmentation by chroma-keying, so that the background can be replaced by any background later.

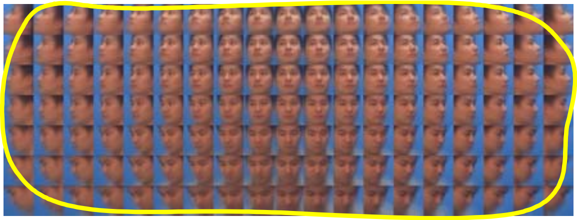


Figure ***1****: Aimed recognizable region with multi-view face descriptor by registering some selected views*

3. Two Multi-view face DBs construction

3.1. Video streams

Face acquisition using a video camera is a method to construct a face database with a large pose variance. We position a video camera at the front of a human subject and video streams are taken by rotating the human subject's chair, so we were able to obtain images of horizontal at interval 1^{o}. The details processes are as follows:

Capture process using a video camera - Two lamps are used to counterbalance shadow. A camera is positioned at eye level. Blue Screen is used for background. A chair is used to rotate the human subject horizontally. A USB video camera connected to PC is used. A conventional graduated protractor is used to measure the vertical degree of the human subject's head. The human subject holds the neck with given vertical degree while whose chair is being rotated in constant speed in horizontal rotation range about [-120^{o}… 120^{o}]. This redundant rotation was due to keep rotation speed constant within horizontal rotation range [-90^{o}… 90^{o}] and to check whether frames of horizontal rotation -90^{o} and 90^{o} are well indicated. The vertical degree of the neck are taken as [-30^{o}, -20^{o}, .. , 30^{o}]. So for each human subject, total 7 video frames are taken to get a full set of horizontal rotation and vertical rotation.

Interpolation-based eye positioning process - After getting the frame numbers corresponding to the views (angle) of horizontal rotation -90°, 0° and 90°, the correspondence between frame numbers and view information is obtained by interpolation. Then an interactive system is used to position eye locations in a few images, say about 12 key frames, and then the all eye locations in 181 views are calculated using Hermit Spline and smoothing methods.

Definition of face localization –

Figure *2* shows one example of face localization (face alignment) definition. Positions of two eyes in the front view are on (0.3, 0.32) and (0.7, 0.32) when width and height are considered as 1.0. For example for an image with size 100 by 200, (0.3, 0.32) means coordinates (30, 62), which is 30th pixel in horizontal coordinate from left and 62nd pixel in vertical coordinate from top. Left eye position of the positive horizontal rotation keeps (0.3, 0.32) while right eye position of the negative rotation does (0.7, 0.32). Vertical rotation has the same eye positions as the ones on zero vertical rotation images.

Cropping images – Cropping requires two parameters, translation and scaling of a cropping window. Eye potions are given in previous process and the scaling factor is approximated by constant horizontal rotation of the human subject. The distance between two eyes and the horizontal rotation degree are used to calculate scaling factor. The distance between the eyes in the front view is considered as *0.4* (scaling factor), that is roughly the length for cropping.

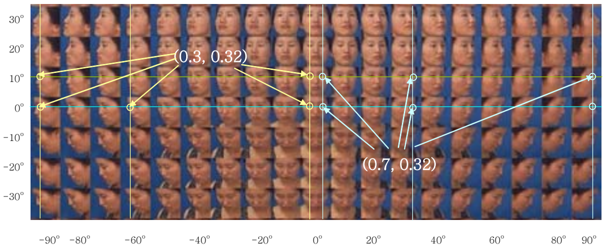


Figure 2: eye-positions on view-mosaic of 7 video streams of a person. Both the horizontal and vertical rotations are considered together for cropping.

However, this scaling factor has to vary according to the rotation of the head. Image is merely a projection of 3D object to the view plane, and this fact can be applied to the distance between the eyes as well. When an imaginary line between two eyes is projected on the view plane, the projection degree would equal to the rotation degree of the head. Therefore, the scaling factor is approximated by an equation *1: 0.4* $\times$ *cos*θ *= Image width: Distance between two eyes*. The cropping parameters for vertical rotation are also obtained in the same way. The results are shown in *Figure 2*.

This video acquirement method seems promising as one of the best systems when it is combined with a built-in studio with several video cameras and also various illumination variances to cover both illuminations and poses.

3.2. 3D facial models

Another way to collect large range of poses and also illumination is using rendering of 3D facial mesh models. We apply rotation of wanted degree to the mesh and then render it using 2D projection. The advantage of this method is that it ensures exact rotation degree of the face, various illumination conditions can be simulated virtually and background is free to design. The used database contains 108 laser scans with texture images of size 480 by 400. The model is aligned to the front view using the information of a nose and two eyes found as facial features in the texture image and directly calculated their 3D locations in a mesh model, and then rendered with the wanted view direction.

Detection of features - Texture images of our mesh models are well aligned in the process of scanning, that is, the face center is located in the middle of the image and its orientation is upright as in the first image of Figure 3. So, heuristic approach using gray values is used to find the features. First, the minimal rectangular search region for nose location is set and then summates the gray pixel values of each row in the region. The sum around the nose-end row will be the smallest, or the darkest, due to nostrils and shadows. The darkest pixel block along that row corresponds to the real nose-end position. With similar methods, the left and right eye positions are captured in separate search regions. For eye cases, search regions are defined automatically from the acquired nose position. And column-wise summation is done first to reduce false detection caused by the eyebrow. In post-processing step for correction, if the height of two detected eyes differs too much, the higher detection is from eyebrow regions, hence, move the search space to lower part. If the detected eyes are too close, they are from nostrils, not real eyes, so move the search region to upper part. By re-examining the updated search region, we were able to locate feature positions successfully for each image in the database.

Alignment of the 3D mesh model - Because we know the 3D coordinates of two eyes and nose, say, $p_{left\_eye}, p_{right\_eye}, p_{nose}$, from their locations in the image, the alignment process is quite straightforward: if we define $v_1 = p_{left\_eye} - p_{right\_eye}$, v_2 = the vector from p_{nose} and orthogonal to v_1 , which is similar to $(p_{right\_eye} + p_{left\_eye})/2 - p_{nose}$, and $v_3 = v_1 \times v_2$, the rotation given by $\begin{bmatrix} \frac{v_1}{\|v_1\|} & \frac{v_2}{\|v_2\|} & \frac{v_3}{\|v_3\|} \end{bmatrix}$ will transform the model to near-frontal view. The resulting model, however, is slightly down-headed for the up-vector approximation, is not parallel to the real up-direction in the face model. So, we rotate the model to upper view with fixed amount of degree at the last stage. Now, the face is aligned and we can render it with arbitrary view direction. We normalized the final image by fixing two eye locations for each view as in the last column of Figure 3.

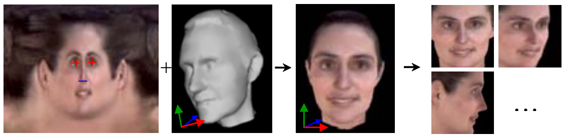


Figure 3: Detected feature points in the texture image are mapped to the 3D mesh model, which is aligned to the frontal view and then rendered with various view directions.

4. Multi-view face descriptor

With large-pose-variation database, we create a multi-view face descriptor for any-view face recognition. If we register every 10° apart, we have to register 19 x 7 views to cover horizontal rotation [-90° … 90°] and vertical rotation [-30° … 30°] and a very naive descriptor does integration of N single view descriptor results size $133\times$ single view descriptor size. Then the descriptor becomes too big. In this paper, we focus on feature extraction and descriptor optimization among many issues in multi-view face recognition.

As we learn from frontal-view face descriptors, a registered view is used to retrieve nearby views (quasi-frontal) with high retrieval rate and we extend the concept of quasi-frontal to quasi-view, from frontal view to general view.

Definition: Quasi-view with error rate *K* - Quasi-view V^q of a given (registered) view V with error K means faces on V^q are retrieved with V with error retrieval rate less than or equal to K.

4.1. Feature extraction

We use a modified version of the best algorithm in the MPEG-7 advanced face descriptor XM which is selected in competition among various algorithms to retrieve faces. More details are found in MPEG document [1]. The choice of descriptor is not the main point in this paper as we focus on how to extend one-view based descriptor to the multi-view descriptor.

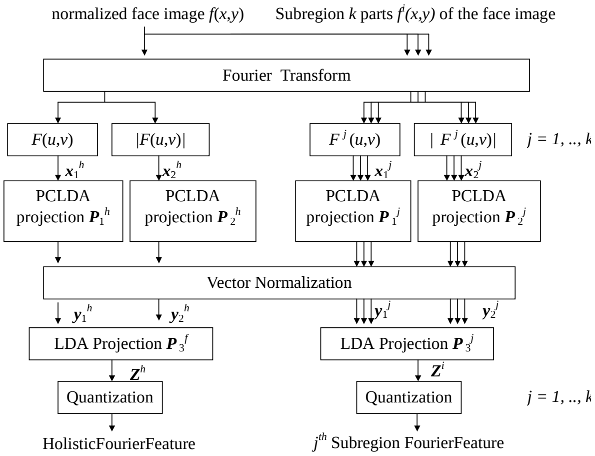


Figure 4: Feature extraction used for Multi-view 3D Face Descriptor

Our Subregion-based LDA on Fourier space as shown in Figure 4 is designed for multi-view purpose. The biggest differences between the MPEG-7 XM [1] and our model are (i) feature extraction in luminance space is removed (ii) the subregion decomposition, which was in luminance space, is now in Fourier space (iii) the number and positions of subregions are various depending on a given view. The first two modifications give more efficient feature extraction with smaller descriptor size. The third modification is caused by the non-frontal-view extension. The new view-dependent subregion is shown in Figure 5.

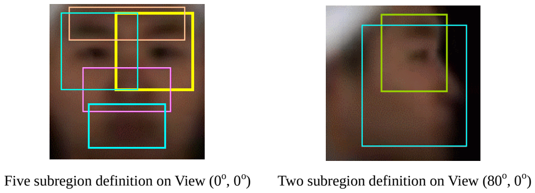
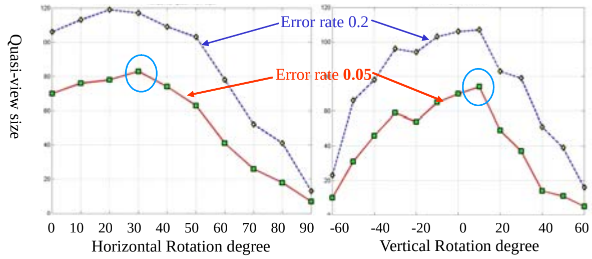

Five subregion definition on View (0°, 0°)
Two subregion definition on View (80°, 0°)

Figure 5: Subregion definition depending on views is shown as superimposed on the center face of our rendering face database

4.2. Quasi-view

Graham and Allinson [3] have calculated the distance between faces over pose to predict the pose dependency of a recognition system. The faces are further apart they will be easier to recognize using distance measures in the Eigenspace and consequently, the best pose samples to use for an analysis should be concentrated around this range. Note that they have checked only horizontal rotation of human heads.



(a) Horizontal rotation (b) Vertical rotation

*Figure **6**: Quasi-view sizes depending on a registered view. (a) It shows 20 to 30 degree horizontally rotated view has the biggest the quasi-view size. (b) It shows about 10 degree vertically rotated view has the biggest the quasi-view size.*

Quasi-view size also depends on the view. We experiment quasi-view inspection with error rate 0.05. Here Figure **6** (a), which has very similar pattern to the ones in [3], which measured distance in the Eigenspace, shows how the quasi-view size varies with horizontal and vertical rotations of a head. To make fair comparison between different views, we extracted 24 holistic features (without using subregion features) for each view. The views (20°, 0 °) ~ (30 °, 0 °) have both the biggest quasi-view size and the biggest Euclidean distance between the people in Eigenspace among views (0 °, 0 °), (10 °, 0 °), ... and (90 °, 0 °). Figure **6** (b) shows the views (0 °, 0 °) ~ (0o, 10 °) have the biggest quasi-view size among views (0 °, -60 °), (0 °, -50 °), … and (0 °, 60 °). Interesting point is that the heading downward views have bigger quasi-view size than ones of heading upward and it shows it is easier to recognize people when they look downward more than they look upward.

4.3. Descriptor optimization

Optimization criteria depends on where the emphasize goes such as the smallest number of registration poses, the smallest descriptor size, or the biggest coverage of poses by selecting views efficiently describing 3D of the face features.

First experiment is done with 50/50 ratios with rendered images from 3D mesh models, which means half of the images are used for training and the other half for test. We concentrate on selection by quasi-view size, number of subregions and number of features on subregions. So for some views, 5 holistic features and 5 features for 5 subregions are extracted and for other views 5 holistic features and 2 features for 5 subregions are extracted. If a view is close to profile, we use 5 holistic features and 5 features for 2. For one view, one image is selected.

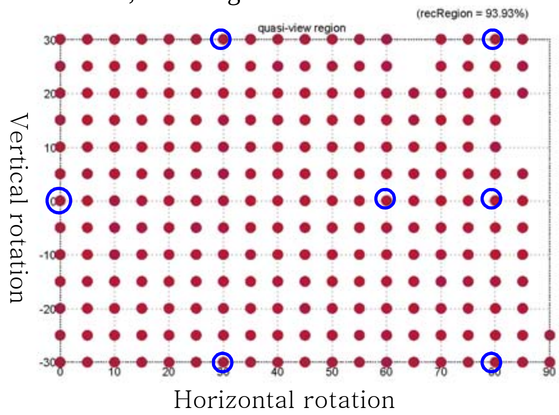


*Figure **7**: Rendered image experiment: The region (93.93%) covered by 7 quasi-views in the view-mosaic of positive horizontal rotation with error rate 0.05. Registration with 13 views are enough to retrieve over 90% of views of horizontal rotation [-90 ° … 90 °] and vertical rotation [-30 ° … 30 °] with error rate 0.05.*

Through experiments with various combinations of quasi-views, a set of views is selected to create multi-view 3D descriptor. An example of several possible descriptors shown in *Figure* **7** has 240 dimensions with rendered images. This descriptor is able to retrieve the rendered images in the test database with error rate 0.05 covering 93.93 % views of total region of view-mosaic of horizontal rotation [-90°… 90 °] and vertical rotation [-30 ° … 30 °]. For a reference, it covers 95.36% for error rate 0.1, 97.57% for error rate 0.15 and 97.98% for error rate 0.2.

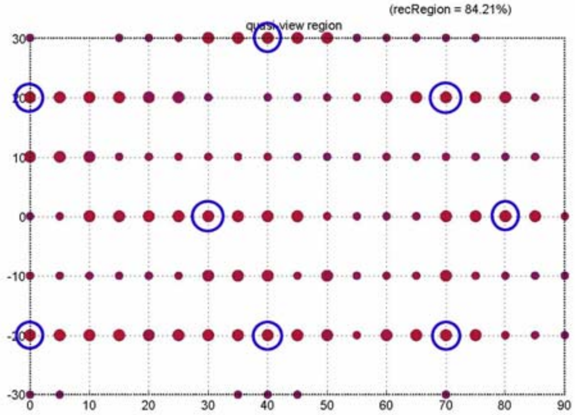


Figure 8: Video image experiment: The region (84.21%) covered by 14 quasi-views in the view-mosaic of horizontal rotation [-90°…90°] and vertical rotation [-30° …30°] with error rate 0.2.

With images extracted from video images have lower success rate. Our experiment is done with 114 different identities of people. The training data does not have pose information as accurate as the rendered images from the 3D-facial mesh models, especially vertical axis of views. In addition the human subject has been sat for a while and rotated to take video streams, so facial, hair and body movement and deformation exist. So the quasi-view size of video images is smaller than the one of rendered images. An example of many possible descriptors is shown in Figure 8. Other examples are 300 dimensions with 87.22% coverage and 270 dimensions with 85.71% coverage with the same error rate.

As a reference for dimension, the current Advanced Face Descriptor [1][7] has 48 dimensions with error rate, called as ANMRR, 0.3013 and 128 dimensions with ANMRR 0.2491 for 50/50 ratio for photograph images. Here ANMRR is a MPEG definition for error rate and our error rate is same as ANMRR. AFD is verified only with limited pose variation such as about (±30°, 0°) and (0°, ±30°), which means the coverage of view is very small compared to our 180° horizontal rotation and 60° vertical rotation.

5. Conclusion

To build a face recognition descriptor containing 3D information of human face features, two different kinds of face database are built. Each of them has number of human identity over 100. The database construction for very dense poses either using 7 video streams from a single video camera or using rendering of a 3D facial mesh obtained from laser scanner is explained. The database is used to construct the multi-view face recognition descriptor. Our multi-view approach with quasi-view dedicated to human face structure shows acceptable size of descriptor and satisfactory recognition rate for recognizing identity of human face in any view.

6. References

[1] A. Yamada and L. Cieplinski, “MPEG-7 Visual part of eXperimentation Model Version 17.1”, ISO/IEC JTC1/SC29/WG11 M9502, Pattaya, Thailand, March 2003

[2] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, “From few to many: Generative models for recognition under variable pose and illumination”, In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[3] D. B. Graham and N. M. Allinson, “Characterising Virtual Eigensignatures for General Purpose Face Recognition” in Face Recognition: From Theory to Applications (H. Wechsler, PJ Phillips, V. Bruce, FF Soulie and TS Huang, eds.), Berlin: Springer-Verlag, pp. 446-456, 1998

[4] http://www.cs.rug.nl/~peterkr/FACE/face.html)

[5] http://www.geometrix.com

[6] S. Gong, S. J. McKenna, A. Psarrou, Dynamic Vision, Imperial College Press, ISBN 1-86094-181-8, 2000

[7] T. Kamei, A. Yamada, H. Kim, W. Hwang, T.-K. Kim, S. C. Kee, “CE report on Advanced Face Recognition Descriptor”, ISO/IEC JTC1/SC29/WG11 M9178, Awaji, JP, December 2002

[8] T. Sim, S. Baker, and M. Bsat, “The CMU Pose, Illumination, and Expression Database”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003.

[9] V. Blanz, T. Vetter, “Face Recognition Based on Fitting a 3D Morphable Model”, IEEE Transactions on Pattern Analysis and Machine Intelligence archive, Volume 25, Issue 9, September 2003

[10] Virtual Media, “Face database constriction for research”, KISA, 2002. 12

[11] W.-S. Lee and S. C. Kee, “3D-Face Descriptor: proposal for CE,” ISO/IEC JTC1/SC29/WG11 M9422, Pattaya, Thailand, March 2003

[12] Whoi-Yul Kim , Jae-Ho Lee, Hyun-Sun Park, Heun-Jin Lee, “PCA/LDA Face Recognition Descriptor with Pose”, MPEG-7 video, Oct. 2002.