Multimodal Facial Gender and Ethnicity Identification

Xiaoguang Lu, Hong Chen, and Anil K. Jain

Michigan State University, East Lansing, MI 48824.
{Lvxiaogu, chenhon2, jain}@cse.msu.edu

Abstract

Human faces provide demographic information, such as gender and ethnicity. Different modalities of human faces, e.g., range and intensity, provide different cues for gender and ethnicity identifications. In this paper we exploit the range information of human faces for ethnicity identification using a support vector machine. An integration scheme is also proposed for ethnicity and gender identifications by combining the registered range and intensity images. The experiments are conducted on a database containing 1240 facial scans of 376 subjects. It is demonstrated that the range modality provides competitive discriminative power on ethnicity and gender identifications to the intensity modality. For both gender and ethnicity identifications, the proposed integration scheme outperforms each individual modality.

1 Introduction

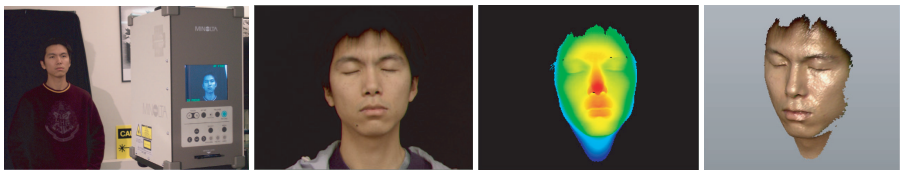
Human face contains a variety of information for adaptive social interactions with people. Humans are able to process a face in a variety of ways to categorize it by its identity, along with a number of other demographic characteristics, such as gender, ethnicity, and age. Gender and ethnicity are involved in human face perception and recognition [1–4, 7].

Unlike gender, ethnic categories are loosely defined due to the intermingling of races and the natural variations within races. We reduce the ethnicity classification into a two-category classification problem, *Asian* and *non-Asian*, which was also used in [14]. Anthropometrical statistics showed the ethnic craniofacial morphometric differences [8] and a close relationship between the 3D shape of the human face and ethnicity [9].

A lot of effort has been spent on the gender and ethnicity classification from different modalities. Most of them are focused on a single modality [11, 10, 12, 15, 13, 14]. Only a few studies have investigated multiple modalities [16].

We address the problem of gender and ethnicity identification using two different facial modalities, range and intensity. With the advances of 3D imaging technology, commercial 3D sensors provide not only the range data, but also registered intensity information [17, 18] (see Fig. 1 for an example of a facial scan). We explore the surface shape (range) of the human face, which captures the craniofacial structure, for determining the ethnicity. Furthermore, since the identification from each individual modality can provide confidence of the assigned class membership for each test sample, the decision accuracy can be enhanced by integrating the confidence from different modalities.

Since the precise facial landmark localization is difficult due to the variations of facial structures, we do not use the anthropometrical measurements. Instead, we explore the appearance-based scheme [**?**], which has demonstrated its power in image-based face recognition.



(a) (b) (c) (d)

Fig. 1. An example of facial scan captured by Minolta Vivid 910. (a) Data-capture scenario; (b) texture image; (c) range image, with points closer to the sensor displayed in red; (d) 3D visualization.

2 Methodology

The system architecture is illustrated in Fig. 2. Range images are normalized in 3D space, and intensity images are normalized consequently. Data within a certain region are cropped from the normalized range and intensity images. Two SVMs classify the cropped range data and the intensity data. The classification results are integrated to achieve the final decision.

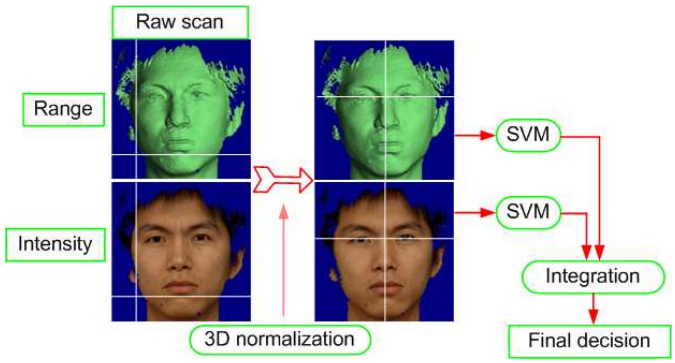


Fig. 2. System Diagram for gender and ethnicity identification.

2.1 Normalization

To apply the appearance-based scheme, the raw scans are required to be aligned [24]: the raw scans are translated, scaled, and rotated so that the coordinates of the reference points are aligned.

The scans obtained from the 3D sensor are a set of points $S = \{(x, y, z)\}$. For the purpose of normalization, we manually specify 6 points in the scan: the inside and the outside corners of the left eye, $E_{l,i}$ and $E_{l,o}$, the inside and the outside corners of the right eye, $E_{r,i}$ and $E_{r,o}$, the nose tip N, and the chin point C. We use $E_{l,i,x}$ and $E_{l,i,y}$ to represent the x and y value of $E_{l,i}$, and $E_{r,i,x}$ and $E_{r,i,y}$ to represent the x and y value of $E_{r,i}$. After rotation, translation and scaling, the points are normalized so that the centers of the left and the right eyes (midpoints of the inside and outside eye corners) are located respectively at $(100, 0, 0)$ and $(-100, 0, 0)$, and the plane that passes the centers of eyes and the chin point, is perpendicular to the z-axis. This transformation is defined as:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = s \cdot R \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}, \tag{1}$$

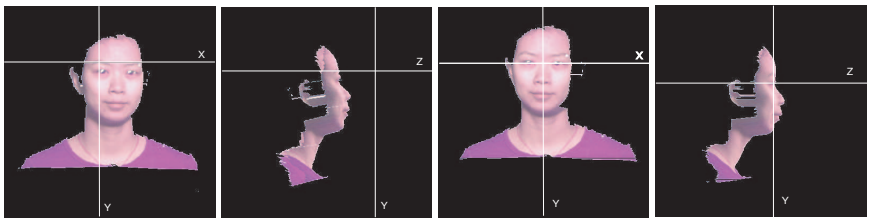
where

$$\begin{aligned} &\begin{pmatrix} t_1 & t_2 & t_3 \end{pmatrix} = -(\overrightarrow{E}_{l,i} + \overrightarrow{E}_{l,o} + \overrightarrow{E}_{r,i} + \overrightarrow{E}_{r,o})/4, \quad s = 400/\|\overrightarrow{E}_{l,i} + \overrightarrow{E}_{l,o} - \overrightarrow{E}_{r,i} - \overrightarrow{E}_{r,o}\|, \\ &\begin{pmatrix} x_0 & y_0 & z_0 \end{pmatrix} = (\overrightarrow{E_{l,i}} - \overrightarrow{C}) \times (\overrightarrow{E_{r,i}} - \overrightarrow{C}), \qquad R = M_z \cdot M_x \cdot M_y, \end{aligned}$$

$$M_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}, \quad M_y = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}, \quad M_z = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\alpha = -\arctan(y_0/\sqrt{x_0^2 + z_0^2}), \quad \beta = \arctan(x_0/z_0), \qquad \gamma = \arctan(\frac{E_{l,i,y} - E_{r,i,y}}{E_{l,i,x} - E_{r,i,x}}),$$

Figure 3 shows the frontal and profile views of a face scan before and after normalization.



(a) (b) (c) (d)

Fig. 3. (a) Frontal view before normalization. (b) Profile view before normalization. (c) Frontal view after normalization. (d) Profile view after normalization.

2.2 Feature Vector Construction

To avoid the effect of hairstyle and other facial accessories, a close facial scan cropping scheme is applied. Given a normalized 3D face data set C, x and y coordinates of a

rectangular area R to be cropped, and the numbers of rows and columns of the grid in the rectangle R, m and n, we crop the face areas and construct feature vectors as follows:

(1) Build a grid G. The grid G is in a plane parallel to the x-y plane. It has m rows and n columns. The borders of G are set to be the rectangle R. A grid G is shown in Fig. 4.

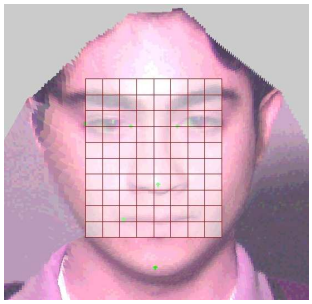


Fig. 4. Cropping face areas for construction of feature vectors. A 10×8 grid is overlaid on the facial scan for demonstration.

(2) Build the $m \times n$ projection matrices XM, YM, ZM. The elements $XM(i,j)$, $YM(i,j)$ and $ZM(i,j)$, $i = 1, \cdots, m$, $j = 1, \cdots, n$, correspond to the grid node $G(i,j)$. Denote the set of points inside $G(i,j)$ as C', where $C' = \{(x,y,z) | (x,y,z) \in C,$ and x, y are inside $G(i,j)\}$. If C' is empty, the corresponding element is labeled as a hole (see Fig. 5). Otherwise, the value of each grid is computed as follows:

$$XM(i,j) = \frac{1}{|C'|} \sum_{\text{for all } (x,y,z) \in C'} x,$$

$$YM(i,j) = \frac{1}{|C'|} \sum_{\text{for all } (x,y,z) \in C'} y,$$

$$ZM(i,j) = \frac{1}{|C'|} \sum_{\text{for all } (x,y,z) \in C'} z,$$

where $|C'|$ is the number of elements in C'.

(3) Interpolation. After the 3D rotation, the occluded points in the original scan cause holes in the normalized scan. The holes in XM, YM, and ZM are filled by interpolating the nearest neighbors as shown in Fig. 5.

(4) Vector formation. The columns in matrices ZM are concatenated to generate the vector V of length $m \times n$, which is used by the classifiers for identification.

2.3 Identification and Fusion of Modalities

The gender and ethnicity identification using individual modalities are formulated as a two-class classification problem. In the appearance-based scheme, Support Vector

Fig. 5. (a,b) Examples of the holes (shown as white patches) after 3D normalization. (c,d) The holes are filled by interpolation.

Machines are a type of classifiers that provide high gender classification accuracy [13]. We use SVMs in our experiments for both ethnicity and gender classifications. Instead of matching scores, the posterior probabilities are extracted from the SVMs [29].

The combination strategy we used in our experiments is the sum rule [19] conducted at the decision level, which has more generality, when classifiers have physically different types of features.

For gender classification, the fusion process is formulated as:

$$p(male|s) = (p(male|s_{range}) + p(male|s_{intensity}))/2, \quad (2)$$

$$p(female|s) = (p(female|s_{range}) + p(female|s_{intensity}))/2, \quad (3)$$

where s is the subject to be classified, s_{range} and $s_{intensity}$ are respectively the range and the intensity maps of the subject, $p(male|s_{range})$ and $p(female|s_{range})$ are the posterior probabilities provided by the SVM that uses range data for gender classification, and $p(male|s_{intensity})$ and $p(female|s_{intensity})$ are the posterior probabilities provided by the SVM that uses intensity data for gender classification. The final decision is made by comparing $p(male|s)$ and $p(female|s)$. The same fusion scheme is applied to the ethnicity identification.

3 Experiments and Discussion

A mixture of two frontal 3D face databases is used for evaluating the proposed schemes. Representative facial scans are given in Fig.6. One database is from University of Notre Dame (UND) [27], composed of 944 scans from 276 subjects. The other is collected at Michigan State University (MSU), containing 296 scans of 100 subjects. The demographic information of the entire mixed database is summarized in Table 1. The samples of both databases were collected using the Minolta Vivid series 3D scanner [17].

For ethnicity identification, a 10-fold cross-validation is conducted. Each time we use 9 folds as the training set and the remaining fold as the test set. Scans from the same subject are grouped into the same set to ensure that the ethnicity classification results are not biased by the similarity between the testing and the training data in terms of the identity. The mean and the standard deviation of the matching error rates from these 10 experiments are reported. The same scheme is applied for gender identification.

Table 1. Number of subjects and scans (given in parenthesis) of the combination of UND and MSU databases in each category.

	Non-Asian	Asian	Subtotal
Female	106 (255)	33 (110)	139 (465)
Male	176 (563)	61 (212)	237 (775)
Subtotal	282 (918)	94 (322)	376 (1240)

Fig. 6. Scan examples in the database. Intensity images (top) and the corresponding range images (bottom). From left to right, they are non-Asian female, non-Asian male, Asian female, Asian male.

The ethnicity and gender identification performance is provided in Tables 2 and 3. For both ethnicity and gender identifications, the experimental results show that 3D (range) information provides competitive results to the 2D (intensity) modality. It is demonstrated that the integration of range and intensity outperforms each individual modality.

Table 2. Ethnicity identification performance. The average and standard deviation of the error rates using 10-fold cross-validation are reported.

	Non-Asian	Asian	Overall
Range	$2.7\% \pm 0.028$	$6.7\% \pm 0.052$	$3.8\% \pm 0.024$
Intensity	$2.1\% \pm 0.027$	$5.9\% \pm 0.051$	$3.2\% \pm 0.029$
Range + Intensity	$0.7\% \pm 0.010$	$5.5\% \pm 0.039$	$2.0\% \pm 0.016$

3D sensors in the current market are not as mature as 2D sensors. Typical problems with range images include missing data near dark regions (e.g., eye regions) and spikes at the region with high reflectivity. The interpolation and smoothing results are the approximations. These problems would deteriorate the gender and ethnicity identification performance using range images even though there exist methods for recovering some data in such areas [27].

Table 3. Gender identification performance. The average and standard deviation of the error rates using 10-fold cross-validation are reported.

	Female	Male	Overall
Range	$24.5\% \pm 0.101$	$9.0\% \pm 0.030$	$14.6\% \pm 0.044$
Intensity	$19.2\% \pm 0.123$	$11.3\% \pm 0.066$	$14.0\% \pm 0.047$
Range + Intensity	$17.0\% \pm 0.093$	$4.4\% \pm 0.032$	$9.0\% \pm 0.030$

4 Conclusions

Gender and ethnicity identifications are important topics in face recognition. The extracted demographic information are useful in many applications. Two different modalities of human faces, range and intensity, are explored. The range information, containing 3D shape of the face object, is utilized for ethnicity identification. A fusion scheme is developed by integrating the range and intensity to identify the gender and ethnicity from facial scans. The proposed scheme can be extended to combine other facial modalities, such as thermal images. Experimental results demonstrate that the range modality provides effective capability for gender and ethnicity identifications. It also shows that the proposed combination strategy obtain better classification accuracy than the classifiers based on each individual modality.

References

1. R. Malpass and J. Kravitz, "Recognition for faces of own and other race," *J. Perc. Soc. Psychol.*, vol. 13, pp. 330–334, 1969.
2. J. Brigham and P. Barkowitz, "Do 'they all look alike?' the effect of race, sex, experience and attitudes on the ability to recognize faces," *J. Appl. Soc. Psychol*, vol. 8, pp. 306–318, 1978.
3. A. O'Toole, A. Peterson, and K. Deffenbacher, "An other-race effect for classifying faces by sex," *Perception*, vol. 25, pp. 669–676, 1996.
4. A. Golby, J. Gabrieli, J. Chiao, and J. Eberhardt, "Differential responses in the fusiform region to same-race and other-race faces," *Nature Neuroscience*, vol. 4, no. 8, pp. 845–850, 2001.
5. A. Puce, T. Allison, J. Gore, and G. McCarthy, "Face-sensitive regions in human extrastriate cortex studied by functional MRI," *J. Neurophysiol.*, vol. 74, pp. 1192–1199, 1995.
6. A. O'Toole, K. Deffenbacher, D. Valentin, and H. Abdi, "Structural aspects of face recognition and the other race effect," *Memory & Cognition*, vol. 22, pp. 208–224, 1994.
7. A. K. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," in *LNCS 3087*, 2004, pp. 259–269.
8. L.G. Farkas, *Anthropometry of the Head and Face*, Raven Press, 2nd edition, 1994.
9. D. Enlow, *Facial Growth*, W.H. Saunders, 3rd edition, 1990.
10. A. O'Toole, T. Vetter, N. F. Troje, and H. H. Bulthoff, "Sex classification is better with three-dimensional structure than with image intensity information," *Perception*, vol. 26, pp. 75–84, 1997.
11. B. Glolomb, D. Lawrence, and T. Sejnowski, "Sexnet: A neural network identifies sex from human faces," in *NIPS*, 1990, vol. 3, pp. 572–577.
12. S. Gutta, J. Huang, P. Phillips, and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," *IEEE Trans. Neural Networks*, vol. 11, no. 4, pp. 948–960, Jul. 2000.

13. B. Moghaddam and M. Yang, "Learning gender with support faces," *IEEE Trans. PAMI*, vol. 24, no. 5, pp. 707–711, May. 2002.
14. G. Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for real time face detection and classification," in *Proc. IEEE FG*, 2002.
15. J. Davis and H. Gao, "Gender recognition from walking movements using adaptive three-mode PCA," in *Proc. IEEE Workshop on Articulated and Nonrigid Motion*, Washington DC, 2001.
16. L. Walavalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma, "Support vector learning for gender classification using audio and visual cues," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 3, pp. 417–439, 2003.
17. *Minolta Vivid 910 non-contact 3D laser scanner*, <http://www.minoltausa.com/>.
18. *Cyberware Inc.*, <http://www.cyberware.com/>.
19. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. PAMI*, vol. 20, no. 3, pp. 226–239, 1998.
20. R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. PAMI*, vol. 17, no. 10, pp. 955–966, Oct. 1995.
21. L. Hong and A.K Jain, "Integrating faces and fingerprint for personal identification," *IEEE Trans. PAMI*, vol. 20, no. 12, pp. 1295–1307, 1998.
22. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Mar. 1991.
23. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
24. S. Shan, Y. Chang, W. Gao, and B. Cao, "Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution," in *Proc. IEEE FG*, Korea, 2004, pp. 314–320.
25. L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. SMC*, vol. 22, no. 3, pp. 418–435, 1992.
26. Y. Wang, C. Chua, and Y. Ho, "Facial feature detection and face recognition from 2D and 3D images," *Pattern Recognition Letters*, vol. 23, pp. 1191–1202, 2002.
27. K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multi-modal 2D and 3D biometrics for face recognition," in *Proc. AMFG*, France, Oct. 2003.
28. C. Wilkinson, *Forensic Facial Reconstruction*, Cambridge University Press, Cambridge, UK, 2004.
29. John C. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in large Margin Classifiers*, Alexander J. Smola, Peter Bartlett, Bernhard Schlkopf, and Dale Schuurmans, Eds. MIT Press, Cambridge, MA, 2000.