Enhancing 3D Face Recognition by Combination of Voiceprint^{*}

Yueming Wang, Gang Pan, Yingchun Yang, Dongdong Li, and Zhaohui Wu

Department of Computer Science and Engineering
Zhejiang University, Hangzhou, 310027, P.R. China
{ymingwang, gpan}@zju.edu.cn

Abstract. This paper investigates the enhancement of identification performance when using voice classifier to help 3D face recognition. 3D face recognition is well known for its being superior to 2D due to the invariance in illumination, make-ups and pose. However, it is still challenged by expression variance. The partial ICP method we used for 3D face recognition could implicitly and dynamically extract the rigid parts of facial surface and be able to get much better performance than other methods in 3D face recognition under expression changes. This work serves to further improve the performance of recognition by combining a voiceprint classifier into partial ICP method. We implement 9 combination schemes, and experiments on database of 360 models with 40 subjects, 9 3D face scans with four different kinds of expression and 9 sessions of utterance for each subject, shows improvement of performance is very promising.

1 Introduction

Identity recognition has been received much attention during past two decades and biometric technologies, such as fingerprint, iris, face and voice, has been studied extensively. Though biometric technologies based on fingerprint and iris can offer greater accuracy, they require much greater explicit cooperation from the user. For example, fingerprint requires that the subject cooperate in making physical contact with the sensor surface. Recognition through face and voice are nature methods for their unlimited feature. Thus it appears that there is significant potential application-driven demand for improved performance in face and voice recognition system.

Most efforts have been made for face recognition from 2D images[1], but only a few approaches exploited 3D information[5]. Although the 2D face recognition system has good performance under constrained conditions, it is still challenged by changes in illumination, pose and expression [1, 6]. Because the 2D image is only a projection of the 3D human face essentially, the system performance can be improved by utilizing 3D information which is the explicit representation of facial surface[6]. However, facial expression is still a big challenge even

^{*} The authors are grateful for the grants from NSF of China (60503019, 60525202) and Program for New Century Excellent Talents in University (NCET-04-0545).

using 3D data in face recognition because in fact facial surface is a non-rigid object. "Different expressions between the gallery and probe sets degrade rank-one recognition rates in 3D face by as much as 33%", reported by [3].

Our previous work has explored facial expression effects in 3D face recognition using partial ICP [7]. The partial ICP method could partly overcome the problem of facial expression variance by dynamically extracting the rigid parts of facial surfaces to match in face identification. This paper will evaluate the enhancement of performance by partial ICP when combining a voice classifier into it. Since Gaussian Mixture Models(GMM) provide a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech [8], we use GMM-based method as a voiceprint component in combination.

The paper is organized as follows: Sec. 2 and Sec. 3 introduce the partial ICP and GMM classifiers respectively. Sec. 4 describe the combining schemes used in the experiments. The experimental results and conclusions are in Sec. 5 and Sec. 6 respectively.

2 3D Face Recognition Component

Our previous work has provided detailed description about partial ICP in 3D face recognition [7]. We briefly review it as follows.

The famous ICP algorithm, developed by Besl and McKay [4], is used to register the point sets by an iterative procedure which is widely used in field of 3D rigid object registration. Let point set $P_1 = \{p^1_1, \dots, p^1_M\}$ and point set $P_2 = \{p^2_1, \dots, p^2_N\}$. The ICP algorithm can be summarized as:

1. $P_2(0) = P_2, l=0$
2. **Do**
3. **For** each point p^2_i in $P_2(l)$
4. Find the closest point y_i in P_1
5. **End For**
6. The closest points y_i form a new point set $Y(l)$ where the pairs of points
7. $\{(p^2_1, y_1), \dots, (p^2_N, y_N)\}$ describe the correspondences between P_1 and
8. $P_2(l)$.
9. **If** registration error E between P_1 and $P_2(l)$ is too large
10. Compute transformation $T(l)$ between $(P_2(l), Y(l))$,
11. Apply transformation $P_2(l+1) = T(l) \bullet P_2(l), l=l+1$
12. **Else**
13. Stop
14. **End If**
15. **While** $\|P_2(l+1) - P_2(l)\| > \text{threshold}$

where point y_k in set $Y(l)$ denotes the closest point in P_1 to the point $p^2_k(l)$ in $P_2(l)$ and the registration error between P_1 and $P_2(l)$ is

$$E = \frac{1}{N} \sum_k^N \|y_k - p^2_k(l)\|^2 \quad (1)$$

For convergence of ICP, the coarse registration usually is carried out before the iterative process.

In general, when utilization of ICP in 3D face recognition, two facial surfaces are registered by the above method. Then the value of E computed in the last time of iterative steps is treated as dissimilarity measure of two faces.

When matching two facial surfaces with different expressions, the difference between the pairs of nearest points may become large due to shape deformation which may have a large effect when performing least-squares minimization and E is no longer accurate as a dissimilarity metric. If only those pairs of points with relatively less deformation are selected as input of calculation of E , the registration error E may be still able to distinguish different subjects while remain small when matching models of same subjects with different expression.

Fig. 1. Discarded area in facial surface with different p -rate(in red)

While the traditional ICP-based method in 3D face recognition uses all point pairs in computing transformation $T(l)$ and E [3], we do it by selecting parts of the point pairs. After sorting the distances of pairs of points in increasing order, we reject the worst $n\%$ of pairs based on distance in each pair. That is, only first $(1-n\%)$ part of distances and corresponding point pairs from sorted distances are chosen to compute transformation E and $T(l)$. Considering the last E that is used as dissimilarity measure of matching, discarding $n\%$ of pairs means removing those points in non-rigid region of facial surface. Thus, it is an implicit method to extract points in rigid parts of facial surface to register and match and the rigid parts extracted are varied according to deformation of facial surface among different matching models. We denote it *partial ICP* for 3D face recognition approach and call $(1-n\%)$ p -rate.

Fig. 1 shows some deformation images in which the darker indicates more deformation and the lighter means less deformation and areas in red indicate regions discarded by setting certain p -rate.

3 Speaker Recognition Component

3.1 GMM Description

A Gaussian mixture density is a weighted sum of M component densities which provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker [8]. It can be formalized as:

$$p(\vec{x}|\lambda) = \sum_{i=1}^M (p_i * b_i(\vec{x})) \quad (2)$$

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{u}_i)' \Sigma_i^{-1} (\vec{x} - \vec{u}_i)\right\} \quad (3)$$

where $\vec{x}$ is a D -dimensional random vector, $b_i(\vec{x})$, $i = 1, \dots, M$, are the component densities and p_i , $i = 1, \dots, M$, are the mixture weights, Σ_i , $i = 1, \dots, M$, are the covariance matrices, $\vec{u}_i$, $i = 1, \dots, M$, are the mean vectors and $\sum_{i=1}^M p_i = 1$. Given p_i , $\vec{u}_i$ and Σ_i , a Gaussian mixture density can be completely parameterized as:

$$\lambda = \{p_i, \vec{u}_i, \Sigma_i\}, i = 1, \dots, M. \quad (4)$$

In speaker recognition based on GMM, each speaker has a corresponding GMM represented by λ .

3.2 GMM Parameter Estimation and Speaker Recognition

Given some utterances of several speakers, the purpose of GMM training is to estimate the parameter of the GMM for each speakers. Suppose vector set $X = \{\vec{x}_1, \dots, \vec{x}_T\}$ is the feature vectors from a speaker, we use maximum likelihood estimation to calculate the parameters of the speaker's GMM. The GMM likelihood function is as

$$p(X|\lambda) = \prod_{t=1}^T p(\vec{x}_t|\lambda) \quad (5)$$

For $p(X|\lambda)$ is a nonlinear function of λ , we use an expectation-maximization algorithm(EM) to compute the parameters [10].

Suppose S speakers labelled as $1, \dots, S$ are trained, each has corresponding GMM with parameters $\lambda_1, \dots, \lambda_S$ and X are feature vector from a test speaker, identification can be performed by finding the speaker model with label $\hat{S}$ which has the maximum posteriori probability. It can be formalized as:

$$\hat{S} = \arg(\max_{1 \leq k \leq S} P_r(\lambda_k|X)) = \arg(\max_{1 \leq k \leq S} \frac{p(X|\lambda_k)P_r(\lambda_k)}{p(X)}) \quad (6)$$

4 Combining Schemes

Following fusion schemes are used in our experiments:

1. Minimum, Maximum, Average and Product scheme
2. Sum, Weighted Sum and Scores Difference-based Weighted Sum
3. "Naive"-Bayes Combination(NB) and Behavior-Knowledge Space

Let $x \in R^n$ be a feature vector, $\{1, 2, \dots, c\}$ be the label set of c classes, $D_i(x) = [d_{i,1}(x), \dots, d_{i,c}(x)]^T$, $i = 1, 2$, be the output of the classifier D_i and $d_{i,j}(x)$ be the degree of "support" given by classifier D_i to the hypothesis that x comes from class j . The fusion result can be presented as

$$\hat{D}(x) = \zeta(D_1(x), D_2(x)) \quad (7)$$

where ζ denotes fusion scheme. Thus, the class label $\hat{L}_D$ of feature vector x can be decided by:

$$\hat{L}_D = \arg(\max_{1 \leq i \leq c} (d_i)), d_i \in \hat{D} \quad (8)$$

The decision profile(DP) of two classifier's outputs can be organized as:

$$DP(x) = \begin{bmatrix} d_{1,1}(x), \dots, d_{1,j}(x), \dots, d_{1,c}(x) \\ d_{2,1}(x), \dots, d_{2,j}(x), \dots, d_{2,c}(x) \end{bmatrix} \quad (9)$$

• Minimum, Maximum, Average and Product Scheme

$$\text{Minimum} : \hat{D}(x) = \{d_i(x) | d_i(x) = \min(d_{1,i}(x), d_{2,i}(x))\}, \quad (10)$$

$$\text{Maximum} : \hat{D}(x) = \{d_i(x) | d_i(x) = \max(d_{1,i}(x), d_{2,i}(x))\}, \quad (11)$$

$$\text{Average} : \hat{D}(x) = \{d_i(x) | d_i(x) = (d_{1,i}(x) + d_{2,i}(x))/2\}, \quad (12)$$

$$\text{Product} : \hat{D}(x) = \{d_i(x) | d_i(x) = d_{1,i}(x) * d_{2,i}(x)\}, \quad (13)$$

where, $d_{1,i}(x) \in D_1(x)$, $d_{2,i}(x) \in D_2(x)$, $i = 1, \dots, c$, $\min(\bullet, \bullet)$, $\max(\bullet, \bullet)$ denotes the smaller and larger element of two given elements respectively.

• Sum, Weighted Sum and Scores Difference-Based Weighted Sum

$$\text{Sum} : \hat{D}(x) = \{d_i(x) | d_i(x) = d_{1,i}(x) + d_{2,i}(x)\}, \quad (14)$$

$$\text{WeightedSum} : \hat{D}(x) = \{d_i(x) | d_i(x) = w_1 * d_{1,i}(x) + w_2 * d_{2,i}(x)\}, \quad (15)$$

where,

$$w_i = \frac{1 - 2E_i}{2 - 2 \sum_{j=1}^2 (E_j)}, i = 1, 2, j = 1, 2, i \neq j \quad (16)$$

and E_j is the error rate of classifier j .

Scores Difference-based Weighted Sum (SDWS) is a trained fusion method based on Weighted Sum rule which using training data and the error rates of classifiers to compute the weights. The detail of SDWS can be seen in [11].

• "Naive"-Bayes Combination(NB) and Behavior-Knowledge Space.

These two schemes tested in our experiments are both trained fusion methods. Using the label results of different classifiers on training data, both "Naive"-Bayes Combination(NB) and Behavior-Knowledge Space schemes compute the probability of each kind of labels combination. When given a testing model, NB and BKS decide the class of the model by indexing into the probability using current labels combination. The detailed description of these techniques can be found in [9].

5 Experimental Results

5.1 Data Acquisition

The data used in our experiments include two parts for partial ICP classifier and GMM classifier respectively. One is facial range data set and the other is the utterance set. The database consists of total 40 different persons.

The facial range data set includes 360 scans in which there are 9 scans for each subject, 2 scans with smile expression, 2 scans with surprise expression, 2 scans with sad expression, 3 scans with neutral expression. All face models are acquired by InSpeck 3D MEGA Capturor DF. Fig. 2 shows some examples of face models. We put one neutral expression face model for each subject into gallery and the other 320 scans are put into probe.

Fig. 2. Face models acquired by InSpeck 3D MEGA Capturor DF

The speech data of each person is divided into 9 sessions with different text. One session is used to train GMM to get the subject's model parameter λ . In the other 8 sessions, 10 prompts are asked to read per session for each subject which are used for testing.

5.2 Results

In the paper, two experiments are conducted to evaluate the enhancement of combining GMM classifier into partial ICP. For those fusion schemes that must be trained, we used leave-one-out method to train them before combination. Before combination, the output of each classifier are normalized so that each matching process issues a support score in $[0,1]$.

Firstly, 5 different values of p -rate $\{0.6, 0.7, 0.8, 0.9, 1\}$ and 9 fusion methods are tested in our experiments. Rank-1 recognition rates are shown in table 1.

From the table, it can be seen that:

- (1) Rank-1 recognition rates can be enhanced remarkably by combining GMM classifier into partial ICP classifier with following fusion method: Minimum, Average, Product, Sum, Weighted Sum and SDWS. The largest improvement of performance is done by SDWS. An average improvement of SDWS, about 5.283%, is reached.
- (2) While the schemes Average, Product, Sum and Weighted Sum achieve similar improvement of performance, the methods, Maximum, NB and BKS, have some decline in rank-1 recognition rate. With each p -rate, we do not see any evidence, in this study, of increasing recognition rate with these three fusion schemes under leave-one-out training method when necessary.

Table 1. Recognition performance varied by different p -rate and fusion schemes

Fusion Scheme	Rank-1 Recognition Rate				
	p -rate=100%	p -rate=90%	p -rate=80%	p -rate=70%	p -rate=60%
partial ICP	90.63%	94.06%	96.56%	95%	94.69%
GMM	91.25%				
Minimum	91.87%	96.88%	95.31%	96.88%	96.88%
Maximum	91.25%	92.19%	94.69%	93.13%	93.13%
Average	96.25%	96.88%	97.5%	98.12%	98.12%
Product	96.88%	96.88%	97.5%	98.12%	98.12%
Sum	96.25%	96.88%	97.5%	98.12%	98.12%
Weighted Sum	96.25%	96.88%	97.5%	97.81%	98.12%
SDWS	97.19%	97.5%	98.44%	98.44%	98.44%
NB	90.94%	91.87%	94.37%	92.81%	92.5%
BKS	90.63%	91.87%	94.37%	92.81%	92.5%

Fig. 3. Performance results by different fusion methods

- (3) When p -rate=60% or 70%, the rank-1 recognition rates are maintained better than other p -rate. With SDWS fusion method, the experiment yields 98.44% rank-1 recognition rate.

Secondly, setting p -rate = 60%, we also plot several Cumulative Match Characteristics (CMC) Curves with different fusion methods, shown in Fig. 3.

From the figure, when considering first 5 best matching models instead of only rank-one model in recognition, identification rate are above 99% using almost all fusion methods except BKS and NB(Experimental results of fusion scheme Average and Sum are not shown in Fig. 3 because their curve will be overlaid by others).

6 Conclusion

The paper explores the improvement of identity recognition rate when combining a voice classifier, GMM-based method, into a 3D face classifier, partial ICP-based method. Several fusion methods are tested and the experimental results are compared with each other. 6 fusion methods can improve the performance of identification. The best average improvement of rank-1 recognition rate, about 5.283%, demonstrates that voice classifier can enhance the performance of 3D face recognition remarkably.

References

1. W.Zhao, R.Chellappa, P.J.Phillips, A.Rosenfeld. Face recognition: a literature survey. *ACM Computing Surveys*, 35(4):399-458, 2003
2. W. Zhao, R. Chellappa. Illumination-insensitive face recognition using symmetric shape-form-shading. *Proc. IEEE ICCV*, 1:286-293, 2000.
3. K.Chang, K.Bowyer, P.Flynn. Effects on Facial Expression in 3D Face Recognition, *SPIE Conference on Biometric Technology for Human Identification*, Apr. 2005.
4. P.J.Besl, N.D.McKay, A method for registration of 3-D shapes, *IEEE Trans.Pattern Anal.Mach.Intell.* 14:239-256, 1992.
5. K.Bowyer, K.Change, and P.Flynn, A short survey fo 3D and multi-modal 3D+2D face recognition, *IEEE ICPR*, 2004.
6. Face Recognition Vendor Test 2002, <http://www.frvt.org/>.
7. Y.M.Wang, G.Pan and Z.H.Wu, Exploring Facial Expression Effects in 3D Face Recognition using Partial ICP, *IEEE ACCV* , 2006. to Appear. Oral.
8. D.A.Reynolds, R.C.Rose, Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE Transactions on Speech and Audio Processing*,3-1:1995.
9. L.I.Kuncheva, J.C.Bezdek and R.P.W.Duin, Decision templates for multiple classifier fusion: An Experimental Comparison. *Pattern Recognition*, 34(2), 2001, 299-314.
10. A.Dempster, N.Laird and D.Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J.Royal Stat.Soc.*, vol.39, 1977, 1-38.
11. D.D. Li, Y.C. Yang and Z.H. Wu, Combining Voiceprint and Face Biometrics for Speaker Identification Using SDWS. To appeared in *Interspeech* 2005.