Template Aging in 3D and 2D Face Recognition

Ishan Manjani[*] ishan12041@iiitd.ac.in

Hakki Sumerkan[†] hsumerka@nd.edu

Patrick J. Flynn[†] flynn@nd.edu

Kevin W. Bowyer[†] kwb@nd.edu

Abstract

This is the first work to explore template aging in 3D face recognition. We use a dataset of images representing 16 subjects with 3D and 2D face images, and compare short-term and long-term time-lapse matching accuracy. We find that an ensemble-of-regions approach to 3D face matching has much greater accuracy than whole-face 3D matching, or than a commercial 2D matcher. We observe a drop in accuracies with increased time lapse, most with whole-face 3D matching followed by 2D matching and the 3D ensemble of regions approach. Finally, we determine whether the difference in match quality arising with an increased time lapse is statistically significant.

1. Introduction

A major motivation for studying 3D face recognition is its potential to achieve higher recognition accuracies and advantages over 2D intensity images [7, 10, 15]. 3D face recognition is robust to variations in illumination and pose which are the major factors impeding 2D face recognition. However the presence of facial expressions and occlusions in 3D face scans is known to diminish recognition performance [1, 3, 4].

Template aging is defined as "the increase in error rates caused by time related changes in the biometric pattern, its presentation and the sensor" [8]. The face is known to undergo various changes with age which have been well researched and modeled. The phenomenon is well summarized as - "Facial aging reflects the dynamic, cumulative effects of time on the skin, soft tissues, and deep structural components of the face, and is a complex synergy of skin textural changes, loss of facial volume, progressive bone resorption, decreased tissue elasticity, and redistribution" [5]. The effect of template aging in face biometrics is inevitable with the degree of changes the human face undergoes [9, 12, 13]. Practical applications of face recognition systems such as surveillance or use in passports - which require identification after long periods of time - may not be effective if the problem arising due to an increased time lapse is not considered.

This paper reports on the first work to look at template aging in 3D face recognition, which we also contrast with 2D recognition. We use a database of sixteen subjects with over 200 3D face scans in all. The scans have been captured over a decade, making the database appropriate to study template aging. 2D face recognition has been performed using the Verilook SDK 5.0 [11]. For 3D face matching we use an ensemble of face regions matched individually using ICP, as proposed in [6], and also compare it with whole-face ICP matching. Using these algorithms we match scans over short and long time lapse to determine if template aging is present. We also study how aging effects matching concerning each 3D region. We assess if the variation in face recognition performance over short and long time lapse is statistically significant. Finally, we compare the magnitude of 3D and 2D template aging.

The paper is organized as follows. In section 2 and 3 we give information regarding the dataset and the protocol used to study the template aging phenomenon. Section 4 describes an automatic method for preprocessing the 3D scan data. Section 5 details our techniques to match scans of subjects in the 2D and 3D domain. We then present the results of the matching algorithms and discuss whether template aging is present in 3D and 2D face recognition. Finally, we conclude our findings.

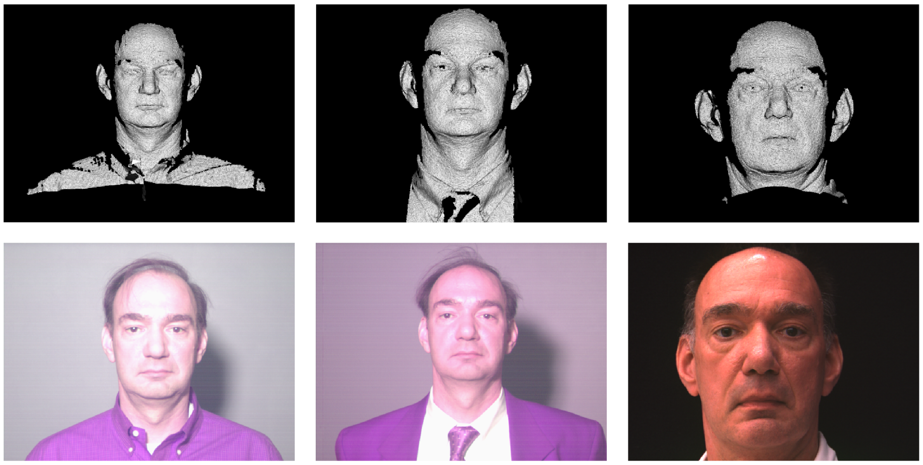
2. Dataset

The dataset was created by selecting all of the subjects out of the Notre Dame database who had a five-year time lapse between earliest and latest 3D face scan. 214 scans of sixteen subjects captured over a period from 2003 to 2012 are utilized, and made available from the University of Notre Dame[1]. The number of subjects is on the small side, but five years of time lapse with the same 3D face sensor is very rare and difficult to obtain. Table 1 summarizes the demographic information of the subjects. The minimum

[*] I. Manjani is with the Department of Computer Science and Engineering, IIIT-Delhi, Delhi 110020, India

[†] H. Sumerkan, P. J. Flynn, and K. W. Bowyer are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

[1] For info on obtaining the dataset: http://www.nd.edu/∼cvrl



01-21-2003

04-15-2003

02-23-2012

Figure 1: Sample 3D meshes and intensity images (underlying text is date of acquisition).

Table 1: Demographic information of the subjects.

Subjects	Gender	Race	Age at enrollment
16	Male 50% Female 50%	White 87.5% Hispanic 6.25% Asian 6.25%	19 - 56 yrs

Table 2: Number of Short/Long, Authentic/Impostor pairs.

	Short Term	Long Term	
Authentic	44	154	198
Impostor	287	2310	2597
	331	2464	2795

number of scans that a subject has is 5, while 24 is the maximum. Each scan consists of 3D and 2D data, with a spatial resolution of 640 × 480.

Each scan is a frontal, shoulder-level-up view of the subject. Sample scans in 3D and their corresponding intensity images for a subject are shown in Figure 1. A 3D scan has approximately 300,000 points, of which typically about 100,000 are valid 3D points. The number of points varies due to a number of factors, including subject-to-sensor distance, the lens used, and hair covering the face.

Each subject chosen for the study has some scans captured within a short time lapse, say three months, and some captured after a long time lapse, say five years, from the first capture. Variations in pose leading to lack of 3D points through self occlusion, or expression instigate a drop in recognition performance. Hence, for an authoritative study of the template aging effect, all included scans are frontal head pose and bear neutral facial expressions. We have assumed that sensor aging does not play a significant role. The same physical sensor was used at Notre Dame the whole time, kept under a maintenance contract.

3. Protocol To Study Template Aging

For each subject, the earliest scan is taken as the enrollment scan. The gallery has the 16 enrollment scans, one per subject. The remaining scans for each subject form the probe set. If the time lapse between the verification scan and the enrollment scan is within three months it is termed as Short Term, while a time lapse greater than five years is termed Long Term. The Short Term period is too small for major modifications of the human face, while over a minimum five year period changes are generally visible. A comparison of face recognition performance over the Short and Long Term period would give an insight into the template aging effect.

A pair of enrollment and verification scans belonging to the same subject are called authentic, while those belonging to different subjects are called impostor. For the experiment, all verification scans in the probe set are matched to all the enrollment scans in the gallery. An enrollment verification scan pair may be Short Term or Long Term depending on the time lapse between their capture, and either authentic or impostor according to the subject they belong to. An enrollment verification scan pair is discarded if it does not belong to either the Short Term or Long Term category. Table 2 presents the number of subject pairs belonging to each of the four categories as defined above.

We look at the verification scenario in which the user claims an identity and the captured verification scan is matched to the enrollment scan of the claimed identity. The receiver operating characteristic (ROC) curve for Short Term, Long Term, and both types of pairs (Complete curve) are plotted as the performance measure.

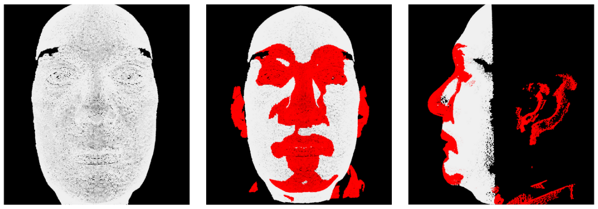


Figure 2: A scan represented as a set of points in the Eucledean space. The XY and YZ plane view of the high curvature regions marked red has been potrayed.

4. 3D Scan Preprocessing

3D scans may have noise spikes and holes due to absorption of the triangulating laser at certain regions such as the eyebrows. These may lead to irregular correspondence of points while registration. We use the moving least squares (MLS) surface reconstruction method to smooth and resample noisy data. The resampling algorithm recreates the missing parts of the surface by polynomial interpolation between the surrounding data points. The holes caused by missing points are filled and noise points are removed. Variations in subject-to-scanner distance or scanner lens may lead to dissimilar number of points across scans. Resampling also ensures the facial scans have similar number of points. The 3D scans do not have any major missing portions due to self-occlusion, since all scans are chosen to be frontal pose.

4.1. Automatic Nose Detection

An automatic technique for nose detection is performed to obtain the nose tip which would be used to segment the facial region. For a given scan we define the high curvature region. For each point of the given surface the principal curvatures p_1 and p_2 are calculated. The mean curvature χ is defined as $\chi = (p_1 + p_2)/2$. A point is said to belong to the high curvature region if the value of χ for the point is higher than the average value of χ across all points. Figure 2 illustrates a scan and the high curvature regions on it. The nose tip is a peak on the surface and is expected to belong to the high curvature region.

We use an iterative approach to detect the nose tip. Each iteration chooses the max z-value point in the high-curvature region as the candidate nose tip point, c_k. The scan is cropped using a sphere with center c_k and registered to a well-segmented template scan. c_k is declared the true nose tip if the registration error is below a threshold, else the process is repeated eliminating points from the scan lying within a horizontal strip of 2mm of c_k.

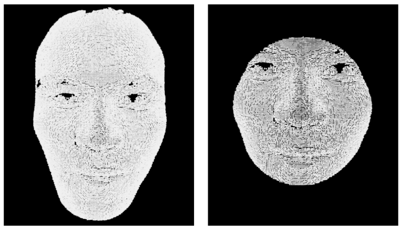


Figure 3: The cropping for an enrollment and probe scan done through a sphere of radius 100mm and 60mm respectively for the whole-face 3D matching. The regions have been represented as a collection of points in 3D.

5. Experimental Methods

The 2D scans are matched using the commercially available Verilook SDK 5.0 [11]. For a pair of images the SDK provides a similarity score. The scale starts from 0, which indicates a poor match, while higher scores indicate it is more likely are the two scans to belong to the same subject.

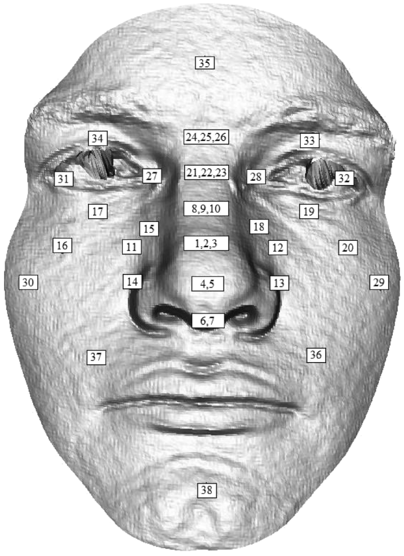
For 3D face matching the 3D point clouds are preprocessed as described in section 4. The point clouds are translated such that the nose point becomes the origin, so that nose tips for a pair of scans coincide. This important to prevent ICP from settling to a local minimum. The gallery set scans are cropped through a spherical region with the nose point as center and a 100mm radius. Probe scans are matched to the enrollment scans using either the whole-face, or an ensemble of regions as formulated in [6]. The Point Cloud Library [14] is used to process and match 3D scans.

5.1. Whole-face 3D Matching

The cropping radius for the probe scans is set at 60mm to be able to utilize as much of the face as possible for matching. Figure 3 illustrates cropped enrollment and verification scans of a subject. The crop radius for the probe scans is lower than the enrollment scans to ensure correspondence for each probe scan point to the enrollment scan points. It also excludes irregular forehead hair points which may cause poor matches. Pairs of scans are registered through the ICP algorithm [2], which returns the root mean squared distance between the closest points as the match score.

5.2. 3D Ensemble of Regions

In [6], the authors define 38 independent spherical regions of the face by a sphere center (as an x and y offset from the nose point) and a sphere radius. We employ independent region matching of the probe scans with a set of face regions as defined in [6]. See Table 3. Positions of the sphere centers on the face are presented in Figure 4. The described regions are extracted from the probe scan and registered to the enrollment through the ICP algorithm producing



(a)

Figure 4: Location of the center points on the face for the independent facial regions as proposed in [6].

Table 4: GAR at 0.1% FAR for different algorithms and time lapse periods.

Matching Technique	Short Term	Long Term	Complete
2D - Verilook	98.36	96.80	93.84
3D Whole-Face	86.54	62.98	64.56
3D Region Ensemble: Sum	99.04	98.15	98.23
3D Region Ensemble: Product	99.04	98.79	98.71

a set of scores. Different regions of the face age differently and this method allows studying how each region performs independently. Further, the region scores are fused with the sum and product rule to seek whether a region ensemble performs better than individual face regions. The probe region match scores are considered in the decreasing order of their verification performance and added one after the other to an initially empty region ensemble. Region scores are normalized by the number of probe region points prior to fusion to neutralize difference in probe region sizes.

6. Results

Table 4 summarizes the GAR obtained at 0.1% FAR for the different matching algorithms. The complete matching performance is 93.84% GAR for 2D face recognition obtained through the Verilook SDK. Whole-face 3D performance is 64.56% which is almost a 30% drop from the 2D performance mark. It indicates that registering the whole-face with ICP algorithm substantially degrades performance, even though 3D face recognition has the potential to perform better than algorithms utilizing 2D information.

Table 3 presents verification rates obtained with each region. Individual regions have a GAR as high as 97.58% (region 2) and as low as 0%. Certain areas of the face age visibly more than others, and matching with those regions may not be as good as others. We observe small radial regions with center near the nose perform better than larger, or regions with center farther away. The best independent region matching is better than the 2D performance by 4%, supporting the claim of better 3D face recognition performance over 2D (for this sensor, dataset, etc).

Fusion of scores from the independent region matching further improves accuracy. The GAR with the sum and product rule is 98.23% and 98.71%, improving over the most-accurate single region by 1%. Peak accuracy for the sum or product is obtained using only three regions: 2,9,7. These regions have radii either 25 or 35 mm and center displaced in the y direction within 20mm of the nose tip. Interestingly, all regions fused by the product rule has a 80.48% GAR, 16% better than whole-face 3D matching.

We look at the verification rates as the time lapse is switched from Short to Long. There is a 24% drop from 86.54% to 62.98% for the whole-face 3D matching. The drop is restricted to an approximate 2% for 2D matching, 0.9% for the best performing independent facial region (region 2), and 0.25% for the region ensembles. Whole-face matching is affected most by template aging. There is a smaller degradation in 2D matching. The effect becomes almost negligible for the region ensemble matching.

Figure 5 shows the authentic and impostor score histograms for each of the four recognition algorithms. Looking at the authentic distributions of each of the four algorithms, the Long Term scores are shifted towards numbers higher than the Short Term scores. The shift of the Long Term scores indicates a decrease in match quality. If the increased scores due to an increased time lapse lead to further overlap of the authentic and impostor distributions, the performance would degrade. The impostor scores show no direct impact of increased time lapse.

To understand the relationship between the Short and Long Term scores of both the authentic and impostor distributions we perform two statistical tests on the set of scores obtained from the experiments of the four algorithms. We consider the t-test with the null hypothesis that the Short and Long term score vectors come from independent random samples of normal distributions with equal means and equal but unknown variances, with the alternate that the vectors come from populations with unequal means. We then look at the Kolmogorov-Smirnov (ks-)test. The null hypothesis is that the the Short and Long term score vectors come from the same continuous distribution while the alternate is that they are from different continuous distributions.

The p-values of the results of the two tests are shown in Table 5. All four algorithms reject the null hypothesis of both the tests for the authentic distributions. For the impos-

Table 3: ICP matching results (GAR at FAR = 0.1%) for the independent face regions for different time lapse periods. Sphere represents the x and y offset of the sphere center from the nose tip, and sphere radius.

Region	Sphere	Short Term	Long Term	Complete	Region	Sphere	Short Term	Long Term	Complete
1	0, 10, 25	94.71	83.35	81.96	20	40, 10, 45	13.30	8.35	7.73
2	0, 10, 35	98.56	97.70	97.58	21	0, 30, 40	14.74	21.61	17.46
3	0, 10, 45	87.98	82.46	82.02	22	0, 30, 35	41.67	14.88	13.92
4	0, 0, 25	93.11	84.76	84.05	23	0, 30, 45	9.94	24.07	11.37
5	0, 0, 45	94.07	78.79	79.03	24	0, 40, 40	4.97	9.88	3.19
6	0, -10, 25	95.67	82.06	81.93	25	0, 40, 35	6.09	10.85	4.86
7	0, -10, 35	95.03	85.48	85.21	26	0, 40, 45	5.29	8.87	2.51
8	0, 20, 35	93.75	48.51	48.07	27	-15, 30, 35	42.63	8.51	7.89
9	0, 20, 25	98.72	95.73	95.14	28	15, 30, 35	16.67	8.75	8.31
10	0, 20, 45	69.07	30.85	30.09	29	50, 0, 45	13.62	7.26	7.22
11	-20, 10, 25	85.42	74.44	72.55	30	-50, 0, 45	4.33	2.02	2.00
12	20, 10, 25	78.69	89.19	82.67	31	-40, 30, 45	0.00	7.02	0.00
13	20, 0, 25	64.90	59.80	57.96	32	40, 30, 45	4.17	0.81	0.84
14	-20, 0, 25	75.16	38.63	37.69	33	30, 40, 45	5.29	0.04	0.06
15	-15, 15, 45	80.29	44.92	45.14	34	-30, 40, 45	0.16	3.75	0.03
16	-40, 10, 45	1.28	4.15	0.32	35	0, 60, 35	2.24	1.13	1.16
17	-30, 20, 45	0.96	21.49	2.45	36	30, -20, 35	62.50	9.35	8.31
18	15, 15, 45	88.62	19.35	18.75	37	-30, -20, 35	68.91	4.64	4.03
19	30, 20, 45	20.35	6.41	5.38	38	0, -55, 35	12.34	0.73	0.68

Table 5: The p-values for the t-test and the Kolmogorov-Smirnov test on the authentic and impostor distributions.

	Authentic		Impostor	
Matching Technique	**t-test**	**ks-test**	**t-test**	**ks-test**
2D - Verilook	9.01E-240	0	0	9.97E-296
3D Whole-Face	3.71E-14	1.47E-20	**0.59**	2.33E-4
3D Region Ensemble: Sum	2.53E-22	3.06E-19	0.005	0.002
3D Region Ensemble: Product	9.42E-10	1.43E-19	**0.27**	**0.013**

tor distributions, the 2D and region ensemble matching with sum rule algorithms reject the null hypothesis of the t-test, while the whole-face 3D matching in addition to these two reject the null hypothesis of the ks-test. The 3D whole-face, and region ensemble matching with the product rule fail to reject the null hypothesis of the t-test. The region ensemble matching with the product rule fails to reject the null hypothesis of the ks-test at 1% significance level.

7. Summary

We study four face recognition algorithms and observe that 3D region ensembles outperform 2D performance, and the whole-face 3D matching. Small facial regions centered near the nose tip lead to high 3D recognition rates, while larger regions and regions further from the nose have lower performance. We look at the Short and Long time lapse performance to investigate template aging for 3D face recognition and to compare it with template aging for 2D face recognition. We observe a 24% drop in GAR at 0.1% FAR for the whole-face 3D matching, while it is 2% for the 2D matching. 3D region ensemble matching substantially improves over the 2D performance, and whole-face 3D matching by restricting the drop to less than a mere 0.25%.

We look at the match scores produced by each of the algorithms to study their relationship with an increased time lapse. We observe that the impostor distributions have poorer scores than the authentic distributions. The more distinct the two distributions are, the better are the verification rates. We observe that, relative to the short-term distributions, the long-term authentic distribution is shifted toward the impostor distribution. We perform two statistical tests with null hypothesis that the Short and Long Term score vectors are from the same distributions. Results from all four algorithms reject the null hypothesis for both the tests for the authentic distributions. For the impostor distributions, the whole-face 3D algorithm fails to reject the null hypothesis of the t-test, and the region ensemble matching with product rule score fusion fails to reject the null hypothesis of both the t-test and the ks-test.

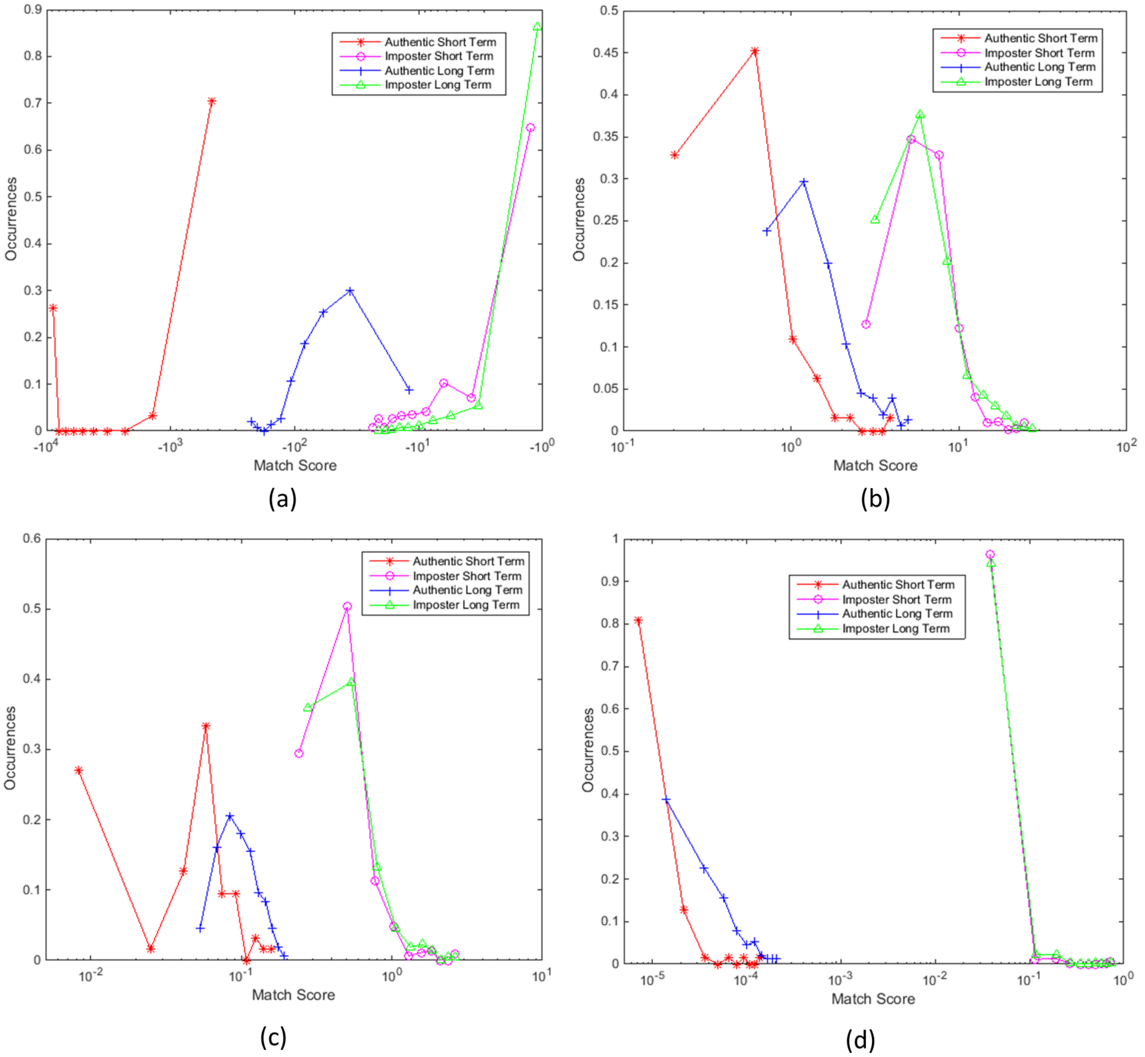
Decreasing accuracy with increased time lapse will be a problem in almost every application of face recognition. Re-enrollment may solve the problem, but increases cost of running a system and may not be practical in surveillance applications. Designing algorithms specifically seeking to counter the effect of aging, similar to pose, illumination, or expressions, for both 3D and 2D systems, is the work for future research.

References

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

[2] P. J. Besl and H. D. McKay. A method for registration of 3-D shapes. *IEEE Trans. PAMI*, 14(2):239–256, 1992.

[3] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.



(a)

(b)

(c)

(d)

Figure 5: The histogram curves obtained for the authentic and impostor distrubution score across the Short and Long Term time lapse periods. The curves for the four different algorithms have been shown: (a) 2D - Verilook, (b) whole-face 3D matching, (c) region ensemble matching - sum rule, and (d) region ensemble matching - product rule.

[4] K. I. Chang, W. Bowyer, and P. J. Flynn. Multiple nose region matching for 3D face recognition under varying facial expression. *IEEE Trans. PAMI*, 28(10):1695–1700, 2006.

[5] S. R. Coleman and R. Grover. The anatomy of the aging face: volume loss and changes in 3-dimensional topography. *Aesthetic Surgery Journal*, 26(1 Supplement):S4–S9, 2006.

[6] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. A region ensemble for 3-D face recognition. *IEEE Trans. Information Forensics and Security*, 3(1):62–73, 2008.

[7] C. Hesher, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range imaging. In *IEEE Symposium on Signal Processing and its Applications*, volume 2, pages 201–204, 2003.

[8] I. ISO. Iec 19795-1: Information technology-biometric performance testing and reporting-part 1: Principles and framework. *ISO/IEC, Editor*, 2006.

[9] A. Lanitis. A survey of the effects of aging on biometric identity verification. *International Journal of Biometrics*, 2(1):34–52, 2009.

[10] G. Medioni and R. Waupotitsch. Face recognition and modeling in 3D. In *IEEE Workshop on AMFG*, page 232233, 2003.

[11] Neurotechnology. Verilook SDK. Available from: http://www.neurotechnology.com/verilook.html.

[12] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, and M. King. Aspects of age variation in facial morphology affecting biometrics. In *IEEE BTAS*, pages 1–6, 2007.

[13] N. Ramanathan, R. Chellappa, and S. Biswas. Age progression in human faces: A survey. *Journal of Visual Languages and Computing*, 15:3349–3361, 2009.

[14] R. B. Rusu and S. Cousins. 3D is here: Point cloud library (PCL). In *IEEE ICRA*, pages 1–4, 2011.

[15] C. Xu, Y. Wang, T. Tan, and L. Quan. Depth vs. intensity: Which is more important for face recognition? In *IEEE ICPR*, volume 1, pages 342–345, 2004.