

Using Multi-Instance Enrollment to Improve Performance of 3D Face Recognition

Timothy C. Faltemier^{*}, Kevin W. Bowyer, and Patrick J. Flynn

*Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana 46556*

Abstract

This paper explores the use of multi-instance enrollment as a means to improve the performance of 3D face recognition. Experiments are performed using the ND-2006 3D face data set which contains 13,450 scans of 888 subjects. This is the largest 3D face data set currently available and contains a substantial amount of varied facial expression. Results indicate that the multi-instance enrollment approach outperforms a state-of-the-art component-based recognition approach, in which the face to be recognized is considered as an independent set of regions.

Key words: 3D face recognition, multi-instance, component-based recognition, biometric, expression variation, range image

^{*} Corresponding author. Telephone: (574) 631-9978. Fax: (574) 631-9260
Email addresses: tfaltemi@nd.edu (Timothy C. Faltemier), kwb@nd.edu (Kevin W. Bowyer), flynn@nd.edu (Patrick J. Flynn).

1 Introduction

In recognition experiments, the gallery contains labeled images of known subjects to be identified, and a probe image is matched to the gallery for purposes of identification. Many 3D face recognition experiments have explicit or implicit assumptions about the type of facial expression expected for the gallery image. A neutral expression is often used because it is generally accepted that the subject will be cooperative in the enrollment phase. However, a neutral expression may not match a non-neutral expression from the same subject as well as it matches a neutral expression images of other subjects [1].

The Iterative Closest Point (ICP) registration algorithm [2] is often used in 3D face recognition [1, 3, 4, 5, 6, 7]. ICP iteratively attempts to align the probe to the gallery by estimating a rigid transformation. Non-neutral expressions or occlusion in the probe or gallery images can result in a poor match. One of our hypotheses is that by using multiple 3D images to enroll a subject, and varying the expressions among the gallery images, we will be able to achieve a better overall recognition rate.

Component-based recognition works by splitting a probe image into many small pieces which are independently matched to complete images present in the gallery. The scores from the component matching results are fused, and a final decision is made. Component-based approaches have been explored in both 2D [8, 9] and 3D [4, 5, 10] face recognition and found to improve performance. In particular, it has been proposed as a method of dealing with expression variation between the enrolled image and the image to be recognized [10]. We compare the multi-instance enrollment scheme explored here against state-of-the-art component based approaches [10, 5, 4].

Multi-instance enrollment has been found to improve performance, both in 2D face recognition and 3D face recognition [11, 12], but has not previously been evaluated as a means to deal with expression variation. This paper uses the ND-2006 3D face data set [13], containing 13,450 total 3D images and their corresponding 2D images, to explore multi-instance enrollment as a means to deal with expression variation in 3D face recognition. The performance of a multi-instance enrollment approach is evaluated against a state-of-the-art component-based recognition algorithm [10, 5] and found to demonstrate superior performance.

The paper is organized as follows. Section 2 gives an overview of related prior work in 3D face recognition and the use of multiple images in the gallery. Section 3 discusses experimental data sets and preprocessing methods. Next, section 4 presents the experimental design in addition to the results found by adding and varying the images contained in the gallery. Finally, section 5 provides conclusions regarding the work presented in this paper.

2 Related Work

A recent broad survey of face recognition research is given in [14] and a survey focusing specifically on face recognition using 3D data is given in [15]. This section focuses on face recognition using multiple enrollment images, multiple expressions, and on component-based approaches.

Hesher et al. [11] used multiple range images per subject with a principal component analysis (PCA) based matcher to allow a greater possibility of matching. Once the sensor acquires the range images and they are normalized, PCA is used to reduce the dimensions of the image representation and facilitate matching. Noise and background information were documented as factors that degraded performance.

The authors perform experiments on a 3D face database containing 222 scans of 37 unique subjects containing a total of 6 facial expressions. They report a range of identification percentages based on the size of the training and testing sets. For a training set containing 185 scans and a testing set containing 37 scans, a maximum identification rate of 94% is reported.

Lu *et al.* [3] created a method for face recognition that uses a combination of 2.5D scans to create a single 3D image for gallery enrollment. For their experiments, the authors employed 598 2.5D probe models of various poses and expressions that were matched to 200 3D gallery models collected at the authors' institution. The authors found that by using the full 3D image of a subject in the gallery and implementations of the ICP and Linear Discriminant Analysis (LDA) algorithms, they were able to achieve a 90% rank one recognition rate using a data set consisting of arbitrary poses. Of the errors reported, nearly all were caused by a change in expression between the probe and the gallery images. In [16], Lu *et al.* present an algorithm for matching 2.5D scans in the presence of expressions and pose variation using deformable face models. A small control set is used to synthesize a unique deformation template for a desired expression class (smile, surprise, etc.). A thin-plate-spline (TPS) mapping technique drives the deformation process. The deformation template is then applied to a neutral gallery image to generate a subject specific 3D deformation model. The model is then matched to a given test scan using the ICP algorithm. The authors report results on three different types of experiments. The first data set contains 10 subjects with 3 different poses and seven different expressions. Rank one results of 92.1% are reported when deformation modeling is used compared to 87.6% when it is not. The second data set consists of 90 subjects in the gallery and 533 2.5D test scans and similar results are reported. Data for the first two experiments was gathered at the authors' institution. The data for the final experiment was taken from the FRGC v2 data set and consisted of 50

randomly chosen subjects in the gallery and 150 2.5D test scans. When deformation modeling is employed a rank one recognition rate of 97% is reported compared to the 81% when it is not.

Yacoob et al. [17] use data sets containing 2D images of 20 subjects and 60 subjects to demonstrate that using a single non-neutral expression (such as a smile) for both the probe and the gallery images in a biometric experiment has more discriminatory power than a neutral expression. The experiments use the PCA algorithm for comparison and the authors define a metric called “discrimination power” to capture the relationship between the within-class and between-class scatters of the images in subspace. This work used the AR database of 2D images.

Martinez [8] uses multiple local region patches to perform 2D face recognition in the presence of expressions and occlusion. He believes that different facial expressions influence different parts of the face more than others. His algorithm addresses this belief by weighting areas that are less affected by the current displayed emotion more heavily. The weighting scheme uses training data labeled with subject and expression information and can be used on either the same subjects or subjects that display similar levels of expressions. The author states that when large cultural differences exist in the testing or training expressions, the weights should be recalculated each time. Reported results show that up to 1/6th of the face can be occluded without a loss in recognition, and 1/3rd of the face can be occluded with a minimal loss. This work also used the AR face database of 2D images.

Heisele *et al.* [9] demonstrates a novel approach for face recognition that combines the techniques of 3D morphable models and component-based recognition. The authors use three 2D face images to generate a 3D head model of each subject. That subject is then rendered under varying illumination and pose conditions to build a large gallery of synthetic images. Recognition is performed on a single global

image as well as 35 individual facial components. Results show that fusing results from the individual components performs better than the single global image. For this experiment, results were reported on 2000 images (10 subjects, 200 images per subject) collected at the authors' institution.

Chang *et al.* [4] investigate the effects of using three overlapping nose regions to improve the performance of 3D face recognition. These regions include a circle centered at the nose, an ellipse centered at the nose, and a region composed of just the nose itself. This method uses the ICP algorithm to perform image matching. 2D skin detection is performed for automated removal of hair and other non-skin based artifacts on the 3D scan. They perform four experiments on a superset of the FRGC v2 experiment 3 data containing 4,485 total scans, 2,798 neutral image sets (449 subjects), and 1,590 non-neutral expression sets (355 subjects). They report results of 97.1% rank one recognition on a neutral gallery matched to neutral probes and 87.1% rank one recognition on a neutral gallery matched to non-neutral probes. The product fusion metric was used to process the results from multiple regions. In the neutral gallery matched to the neutral probe set, maximum performance was reported when only two of the three regions were combined. The author mentions that increased performance may be gained by using additional regions.

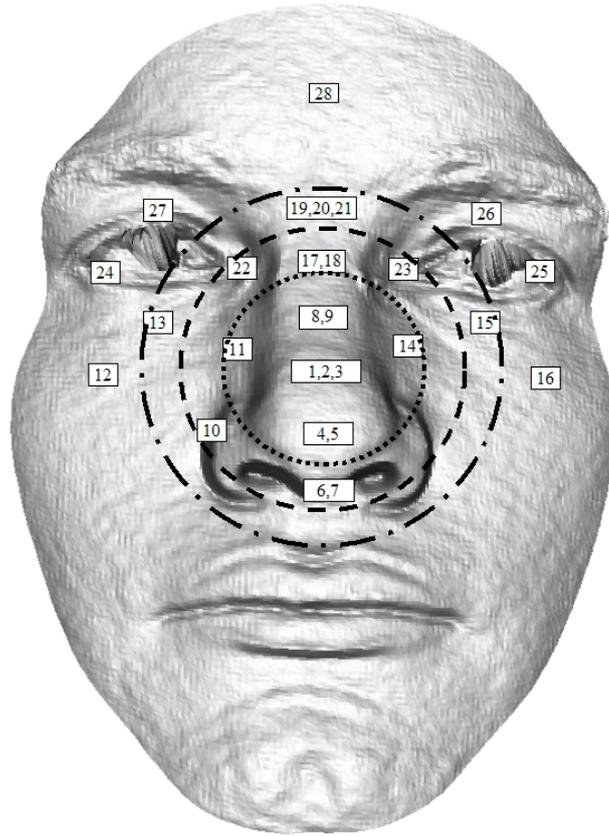
Gökberk *et al.* [18] perform a comparative evaluation of five face shape representations, (point clouds, surface normals, facial profiles, PCA, and LDA) using the well known 3D-RMA data set [19] of 571 images from 106 subjects. They find that the ICP and LDA approaches offer the best average performance. They also perform various fusion techniques for combining the results from different shape representations to achieve a rank-one recognition rate of 99.0%.

Mian *et al.* [7] propose an expression invariant approach to 3D face recognition that reports results on the FRGC v2 data set [20]. The authors automatically detect the

nose, perform pose correction and normalization in both 2D and 3D, create a rejection classifier to reduce the overall processing time, and finally segment the 3D images into two regions (nose and eye/forehead) and match them independently to increase the overall recognition rate. They report verification rates of 98.5% at 0.1% FAR and rank one identification rates of 96.2% based on a neutral gallery and a probe set comprising the remaining images using their R3D algorithm. In addition, the authors report that the eye/forehead and nose regions of the face contain maximum discriminating features necessary for expression invariant face recognition.

Cook *et al.* [21] present a novel method for accomplishing expression insensitivity in 3D face recognition based on Log-Gabor Templates. The authors apply 18 Log-Gabor filters on 49 windows to generate 147 feature vectors comprising 100 dimensions. After matching is performed, they report results using the FRGC v2 data set [20]. When the 4007 x 4007 similarity matrix is calculated, the authors report a 92.31% verification rate at 0.1% FAR. In the identification scenario, the authors employ the first image of a subject in the gallery set (466 images) and the remainder in the probe set (3581) for a rank one recognition rate of 92.93%. The authors also discuss how the best performance is achieved when using windows surrounding the upper nose area while the inclusion of outer areas adversely affects the accuracy of the system.

In this paper, Experiment 3 compares the performance of the Multi-Instance Enrollment approach to the previously developed Region Ensemble for Face Recognition (REFER) algorithm [10]. The REFER algorithm exploits the presence of subregions on the face that are relatively expression-invariant and uses a committee of classifiers based on these regions to improve performance. Each region matches independently to a gallery surface using the ICP algorithm [2], resulting in 28 unique



(a) Image number 04514d324

Fig. 1. Image number 324 of subject 04514 displaying the location of the 28 centroids used by the REFER algorithm during the region of interest (ROI) selection. Each number corresponds to the region location information in Table 1. The circles show the relative size and coverage of the three spheres for the selected location.

error distances for a single probe to gallery comparison (which is increased from the 14 regions used in [5]). The component-based analysis of a probe image matched to an enrolled image has been used successfully in 2D face recognition [8, 9], and now also in 3D face recognition [4, 7, 22].

For recognition experiments, the REFER algorithm uses the Borda Count (BC) method [23, 24, 25, 18, 26] for score based decision fusion. We use a modified version of the BC method. Unlike the standard BC method that provides a rank score (1st place, 2nd place, 3rd place, ..., n th place) to each component-to-gallery similarity score entry in the set from $1..n$ where n is the number of gallery images,

Table 1

ICP Probe Region Information - The X and Y parameters determine the offset values (in mm) for the new sphere center in relation to the origin

Region	X	Y	SphereRadius	Region	X	Y	SphereRadius
1	0	10	25	15	30	20	45
2	0	10	35	16	40	10	45
3	0	10	45	17	0	30	40
4	0	0	25	18	0	30	45
5	0	0	45	19	0	40	40
6	0	-10	25	20	0	40	35
7	0	-10	35	21	0	40	45
8	0	20	35	22	-15	30	35
9	0	20	45	23	15	30	35
10	-20	0	25	24	-40	30	45
11	-15	15	45	25	40	30	45
12	-40	10	45	26	30	40	45
13	-30	20	45	27	-30	40	45
14	15	15	45	28	0	60	35

we only give a rank score to the first α (in our experiments $\alpha = 4$) entries in the gallery. The traditional BC method takes the sum of the ranks to determine a final score. Our version distributes points such that the first place gets $(\alpha)^2$ points, second gets $(\alpha - 1)^2$ points, until the number of points equals 0. This modification yielded experimentally higher results than the traditional BC method. We believe this result is due to the larger weight given to rank-one matches.

Figure 1 shows the centroid location of the 28 overlapping sphere regions used by the REFER method [5, 10]. Table 1 shows the cropping parameters used to generate each region. X and Y determine the offset values for the new sphere center in relation to the origin and SphereRadius determines the new sphere radius. By selecting multiple regions, any small inconsistencies found in a single model can be overcome when combining the different views. Further details of this algorithm and implementation are discussed in our previous work [5, 10].

3 Data Sets and Preprocessing

3.1 The ND-2006 Data Set

Images for the experiments performed in this paper are a superset of the FRGC v2 data set [20] named ND-2006. The FRGC v2 data set only contains 66 subjects that display all of the 5 expression types (N, H, Sp, D, and Sd) and at least one additional probe image. Of these 66 subjects, all of the additional probe images are only comprised of neutral expressions. In the context of testing a multi-gallery enrollment approach, this does not provide enough diversity of expressions in the probe set. For this reason, we chose to use the larger ND2006 data set for comparisons. The ND-2006 data set contains a total of 13,450 images containing 6 different types of expressions (Neutral, Happiness, Sadness, Surprise, Disgust, and Other) as seen in Figure 2. A total of 888 distinct persons, with as many as 63 images per subject, are available in the ND-2006 data set. A table containing the number of images based on expression type is shown in Table 2. While the “Neutral” category contains the largest number of images, expressions account for more than a quarter of the data set. The “Other” expression listed in Table 2 consists of a variety of expressions that did not fit into a predetermined category (such as a face displaying a whistling expression, cheeks puffed out, and others). These expressions are only included in experiments as probe images due to their non-specific categorization. Information about obtaining a copy of this data set will be available soon [13].

The images were acquired with a Minolta Vivid 910 range scanner [27]. The Minolta 910 scanner uses a laser stripe projector and triangulation to build a 3D model of the face. The image is initially in a range image format of 640x480 with a flag field indicating 0 if a 3D point was not sensed and 1 if it was. Color texture (r, g, b) and 3D location (x, y, z) are produced, but not perfectly simultaneously, as the laser

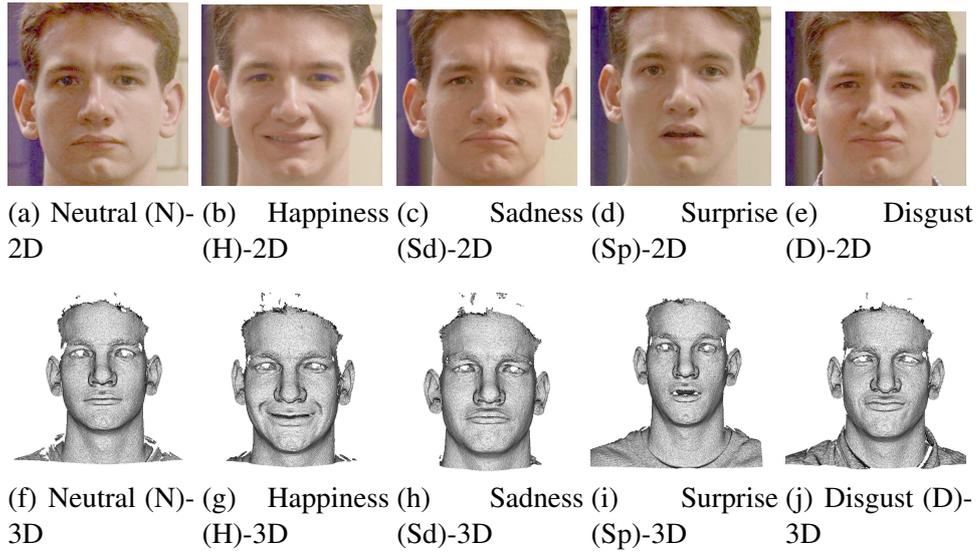


Fig. 2. Different types of expressions gathered for subject 04514 and their associated texture and 3D images.

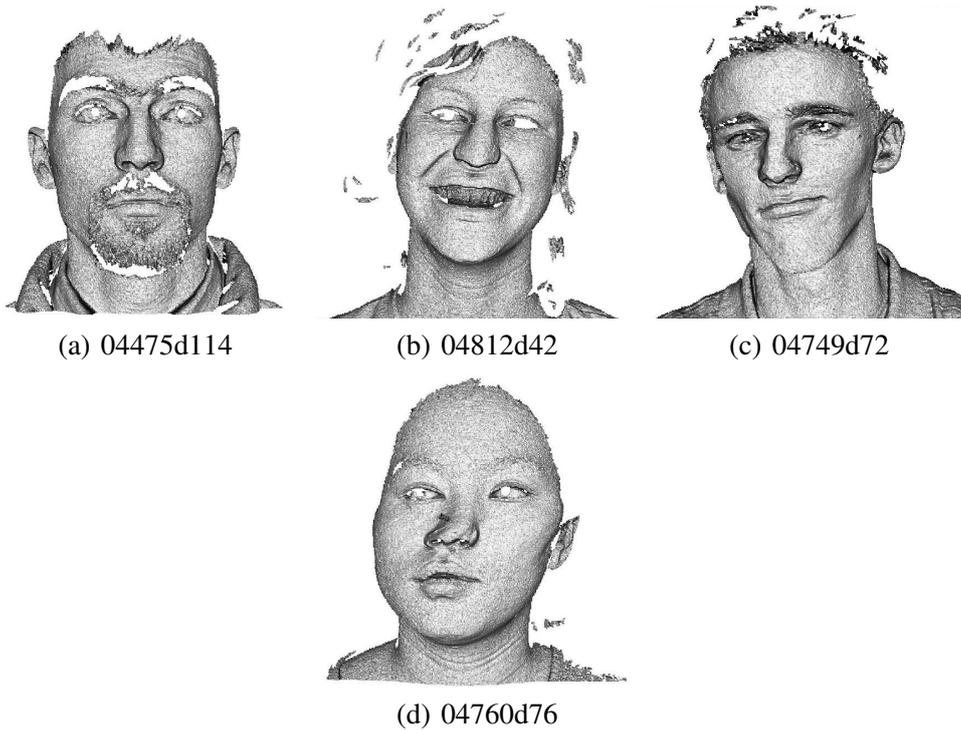


Fig. 3. Examples of images that contain artifacts in the ND-2006 data set. Images (a) and (b) show large holes in the mesh, and images (c) and (d) show distortion due to movement during acquisition.

Table 2
Types of expressions in the ND-2006 data set

Expression	Number of Images
Neutral (N)	9889
Happiness (H)	2230
Disgust (D)	205
Sadness (Sd)	180
Surprise (Sp)	389
Other (O)	557

stripe requires a few seconds to cross the face and the color image is taken after the shape is acquired. Subject movement can result in texture misregistration or mesh distortion, examples of which can be seen in Figure 3. None of the images in the ND-2006 data set were excluded from the experimental data sets in this paper due to artifacts or holes in the images. The number of 3D points on a frontal face image taken by the Minolta camera is typically around 112,000, and depends on the lens used as well as standoff. Example images from this sensor can be seen in Figure 2.

3.2 Data Preprocessing

Our algorithm operates automatically using only the 3D shape from a frontal view of the face. First, small holes in the range image are filled by locating “missing” points that are surrounded by 4 or more “good” points. The x, y, and z coordinates of the missing point are interpolated from its valid neighbors. Boundary points are initially determined by a row-wise sweep through the range image row by row to find the first and last valid 3D point. This process is repeated until no additional points are created. Once hole filling is performed, a final pass over the range image with a 3x3 median filter smooths the data and removes spikes in the z-coordinate. The process of hole filling and filtering is illustrated in Figure 4.

Finally, the nose tip point is detected using a consensus of three methods [5]. The

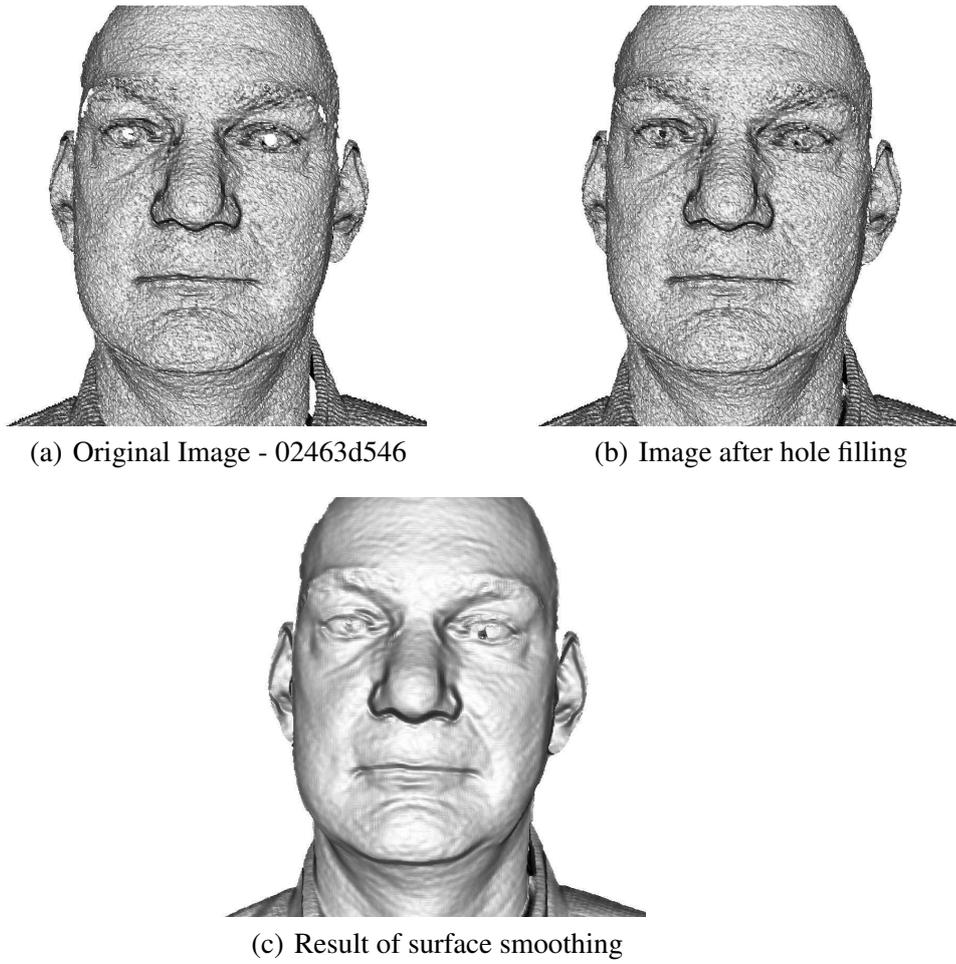


Fig. 4. Preprocessing steps used on the ND-2006 data set.

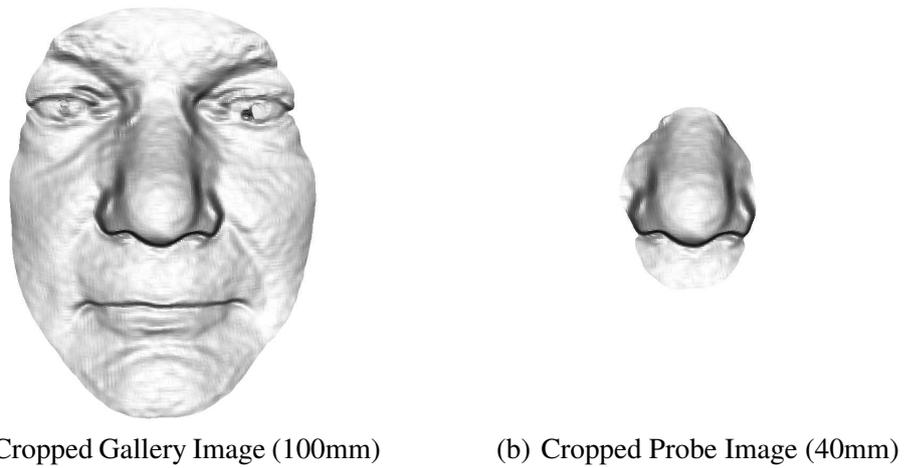


Fig. 5. Gallery and Probe images used in the Multiple Instance Experiment.

first method finds the curvature and shape index [28] of each point on the face to find nose tip candidates labeled c_n . The next method aligns the input image to a template, using the ICP algorithm, where the position of the nose tip is the highest Z value in the image (assuming that spikes and holes are removed) and is labeled p_n . If the candidate nose tip points found in these methods are within 20mm, then the average of p_n and c_n is reported as the final nose tip location. If this is not the case, a tiebreaker step is performed. The position of the nose tip is known on the template image and when properly aligned, this point should be the nose tip of the input image as well and is labeled m_n . Distances between m_n , p_n , and c_n are calculated and the pair that contains the smallest distance is averaged and reported as b_n , the best nose tip.

If the image is being used to enroll a person in the gallery, the points located within a 100mm radius sphere centered at the nose tip are extracted for use as the gallery image. If the image is a probe for our multiple gallery experiment, a sphere containing a 40mm radius centered at the nose tip is extracted. A significantly smaller surface is used for the probe than the gallery to ensure that the the probe image will always be a subset of the gallery image. Another benefit of using a smaller probe region is that the area subject to variation across expressions (i.e. the cheek and nose bridge) is limited. Examples of gallery and probe images can be seen in Figure 5. The number of vertices in a gallery image is typically 60,000 while the number of vertices in a probe image is only 10,000.

4 Experimental Design and Results

This section reports the results of three main groups of experiments. One group of experiments looks at multi-instance galleries containing up to five different facial

expressions per subject. A second group of experiments looks at multi-instance galleries of up to five happy or neutral expression images per subject. A third set of experiments compares the results of our multi-instance gallery approach to the result of using the algorithm by Chang *et al.* [4], to the results of the component-based REFER algorithm [5, 10], and to using a single frontal image for enrollment.

The core matching engine is an implementation of the iterative closest point (ICP) algorithm [2]. The ICP algorithm iterates until the difference in successive RMS scores is less than 0.0001 or until reaching a limit of 100 iterations. Provided that the initial alignment is good, convergence typically occurs in less than 30 iterations. The RMS of the final alignment is reported as the matching value between a probe image and a given gallery image. For a multi-instance gallery representation of a person, the minimum RMS of the probe to any one of the multiple instances is reported as the matching value for that person. Based on experimental results, we determined that the ICP algorithm is sensitive to initial alignment. The experimental results suggest that degraded performance can result if the starting position is misaligned by more than 10mm in the X or Y directions.

4.1 Multi-Instance Gallery Using Varied Expressions

In this first set of experiments, the size of the gallery is varied from one to five images, and the different possible combinations of one of each of up to five different expressions are used. The probe set is the same for each of the 31 different sub-experiments in this overall experiment. It contains 2500 neutral-expression images and 791 images with varying facial expression, with from 1 to 50 probe images for each of the 112 subjects in the gallery.

Table 3
Experiment 1 - Multiple Expression Experiments (3921 Probe Images) - Overall Results

Sub-Exp #	Gallery Type	# of Gallery Images	Rank One Recognition	TAR at 0.1% FAR
1	N	112	82.8	63.5
2	H	112	85.1	61.3
3	Sp	112	89.8	70.2
4	D	112	66.5	44.8
5	Sd	112	91.0	73.1
6	N+H	224	90.6	75.4
7	N+Sp	224	94.8	80.5
8	N+D	224	87.9	61.7
9	N+Sd	224	94.5	81.9
10	H+Sp	224	94.9	78.8
11	H+D	224	87.6	68.4
12	H+Sd	224	95.5	83.1
13	Sp+D	224	91.2	74.7
14	Sp+Sd	224	97.4	85.4
15	D+Sd	224	91.9	76.6
16	N+H+Sp	336	95.8	84.8
17	N+H+D	336	92.8	77.2
18	N+H+Sd	336	96.9	87.6
19	N+Sp+D	336	94.9	81.3
20	N+Sp+Sd	336	97.9 (ns)	88.4
21	N+D+Sd	336	95.2	83.7
22	H+Sp+D	336	94.7	79.8
23	H+Sp+Sd	336	97.7 (ns)	87.9
24	H+D+Sd	336	95.7	83.1
25	Sp+D+Sd	336	97.7 (ns)	86.2
26	N+H+Sp+D	448	95.9	84.5
27	N+H+Sp+Sd	448	98.0	89.2
28	N+H+D+Sd	448	96.8	86.3
29	N+Sp+D+Sd	448	97.9 (ns)	88.5
30	H+Sp+D+Sd	448	97.8 (ns)	88.4
31	N+H+Sp+D+Sd	560	98.0	91.0

The accuracy of the nose-tip feature detection gives an approximate practical limit to the recognition accuracy. This is because if the nose tip detection is inaccurate, the probe region may be inaccurately cropped from the probe image and. The accuracy of the automatic nose tip was checked by comparing to manually marked nose tip location, and 97.9% of the automatically determined nose tip points were within 20mm of the manually-marked point.

Results of this set of experiments are summarized in Table 3. Results are reported in terms of both a rank-one recognition rate appropriate to an identification scenario, and a true accept rate (TAR) at fixed false alarm rate (FAR) value appropriate to a verification scenario. The particular FAR point chosen is 0.1%, which is the value used in the Face Recognition Grand Challenge program [20].

One interesting point from the results in Table 3 is the performance obtained using the “sad” and “surprise” expression images in the gallery. Of the five different expressions used as single-instance galleries - neutral, happy, surprise, disgust, and sad - the highest performance was obtained using the sad expression images and the second-highest performance was obtained using the surprise expression images. This is true even though the probe set contains very few images representing either of these two expressions. None of the 791 non-neutral expression images in the probe set represent the sad expression, and only 65 represent surprise. The results across the two-instance gallery representations reinforce this result. The highest performance of the ten different two-instance gallery representations is with the gallery containing sad + surprise expressions. Similarly, the three highest-performing three-instance galleries are the ones that contain sad and surprise, along with each of the other three expressions. And the highest-performing four-instance galleries are ones that contain both the sad and surprise expressions. These two facial expressions have not previously been recognized as being particularly good expressions for enrollment when the probe set contains other varying expressions.

To investigate why these two expressions caused such high performance, we took 10 people who have at least one image containing each one of each of the five expressions (N, H, Sp, D, and Sd). For each subject we selected the first image of each expression type to include in this experiment (for a total of 50 images).

After performing an “All vs. All” experiment, where each image is matched to the remaining images, the following RMS averages were extracted from the scenarios listed in Table 4. Based on the experimental results in this table, we see that the RMS distance is not the only factor that determines the quality of the match.

Table 4
Neutral, Sadness, and Surprise RMS average decomposition.

Probe Expression	Gallery Expression	RMS Average
Neutral	Neutral	0.488
Happiness	Neutral	0.426
Surprise	Neutral	0.376
Disgust	Neutral	0.551
Sadness	Neutral	0.480
Neutral	Surprise	0.490
Happiness	Surprise	0.407
Surprise	Surprise	0.317
Disgust	Surprise	0.562
Sadness	Surprise	0.491
Neutral	Sadness	0.587
Happiness	Sadness	0.514
Surprise	Sadness	0.424
Disgust	Sadness	0.463
Sadness	Sadness	0.395

One of the possible reasons the Surprise and Sadness galleries perform better than the other gallery expressions may be that the surprise and sadness expressions are much more reproducible in terms of prompted expressions. When a subject is asked to display a neutral expression, often times they may simply display the expression that is on their face at that time, which may vary from session to session. This is in contrast to the images containing surprise and sadness expressions. After manually examining images containing these expressions, we saw a much more consistent resulting image. In addition, after preprocessing, we did not see a source of significant noise that would cause the regions to be mismatched for any other reason than expression. Another reason why the Surprise and Sadness gallery expressions perform better than the Neutral expression gallery may be due to a uniqueness factor.

For our multiple expression experiments in this paper, we used a small region of interest containing only the nose area. Nose regions across neutral expressions of non-matching images had a lower RMS score than nose regions across surprise and sadness expressions of non-matching images. These results suggest that the nose regions of neutral images are more similar to each other than those of the surprise and sadness nose regions.

Another interesting point from the results in Table 3 is that the “disgust” expression leads to particularly poor performance. This suggests that facial movement caused by the disgust expression is very dissimilar to the rest of the expressions. It could also be explained by an increase in difficulty in obtaining the expression itself. During acquisition, many subjects questioned exactly what an expression of disgust looked like. This confusion could cause the subject to attempt to do something significantly different with their face than is normal. Still, even adding this expression of disgust (which performed poorly by itself) to any other image caused an increase in recognition rate. This implies simply that the disgust image resulted in a different permutation of the facial features and yielded another opportunity for correct matching.

The rank-one recognition rates for sub-experiments 20, 23, 25, 27, 29, and 30 are not statistically significantly different from each other, using a z test at the 0.05 level. However, the rank-one recognition rates for these sub-experiments are statistically significantly greater than for the other sub-experiments in this group. The result variation for one-image galleries is $70.2 - 44.8 = 25.4$, for two-image galleries 23.8, three-image galleries = 10.7, and four image galleries is 4.7. This suggests that the variability is less for a larger and more heterogenous gallery, or that the types of expressions are less important as the size of the gallery increases.

Table 5
Experiment 1 - Multiple Expression Experiments (3921 Probe Images)

Sub-Exp #	Gallery Type	N (2500)	H (507)	Sp (65)	Other (219)
1	N	2043	444	58	180
2	H	2081	479	56	187
3	Sp	2237	474	63	184
4	D	1608	366	42	173
5	Sd	2291	450	58	196
6	N+H	2231	495	62	196
7	N+Sp	2373	476	65	196
8	N+D	2171	464	59	200
9	N+Sd	2362	484	63	202
10	H+Sp	2345	490	63	195
11	H+D	2137	490	58	199
12	H+Sd	2381	497	62	205
13	Sp+D	2254	484	64	201
14	Sp+Sd	2444	493	64	205
15	D+Sd	2294	473	58	201
16	N+H+Sp	2394	495	65	201
17	N+H+D	2292	497	62	205
18	N+H+Sd	2420	501	64	206
19	N+Sp+D	2366	490	65	204
20	N+Sp+Sd	2455	497	65	205
21	N+D+Sd	2375	489	64	206
22	H+Sp+D	2350	500	64	203
23	H+Sp+Sd	2448	500	64	205
24	H+D+Sd	2380	499	63	208
25	Sp+D+Sd	2447	497	64	209
26	N+H+Sp+D	2392	497	65	205
27	N+H+Sp+Sd	2455	501	65	207
28	N+H+D+Sd	2414	501	65	208
29	N+Sp+D+Sd	2451	498	65	209
30	H+Sp+D+Sd	2447	501	64	209
31	N+H+Sp+D+Sd	2453	501	65	209

For completeness, a breakdown of the number of probes correctly recognized at rank one is given in Table 5. Here it is clear, for example, that the sad expression is a better enrollment choice than the happy or neutral expression, even for purposes of recognizing happy or neutral probes. The sadness (Sd) and disgust (D) are not included in this table because no probes were available with either of those expressions.

Table 6
Experiment 2 - Neutral and Happiness Experiment (5968 Probe Images) - Overall Results

Sub-Exp #	Gallery Type	Number of Gallery Images	Rank One Recognition	TAR at 0.1% FAR
1	N	259	80.2	69.1
2	H	259	79.9	68.8
3	N+N	518	89.2	80.4
4	H+H	518	89.0	77.3
5	N+H	518	89.4	81.6
6	N+N+N	777	93.1	83.8
7	N+N+H	777	93.5	85.1
8	N+H+H	777	93.9	85.4
9	H+H+H	777	92.4	83.4
10	N+N+N+N	1036	94.5	85.9
11	N+N+N+H	1036	95.0	88.2
12	N+N+H+H	1036	94.7	88.0
13	N+H+H+H	1036	96.9 (ns)	90.1
14	H+H+H+H	1036	94.0	84.7
15	N+N+N+N+N	1295	94.9	87.8
16	N+N+N+N+H	1295	96.0	89.5
17	N+N+N+H+H	1295	96.1	90.0
18	N+N+H+H+H	1295	97.2	91.5
19	N+H+H+H+H	1295	97.1 (ns)	90.1
20	H+H+H+H+H	1295	94.8	85.7

4.2 Comparison of Happy and Sad for Multi-Instance Gallery

Most face recognition studies to date have dealt only with neutral-expression images. The most common non-neutral expression has been a smile or happy expression. For this reason, we explore the use of these two expressions in more detail in a second set of experiments. This set of experiments contains images from all subjects in the ND 2006 data set that have at least five neutral-expression images, five happy expression images, and additional images for use in the probe set. This leads to a gallery containing 259 subjects, and a fixed probe set of 5,968 images, with from one to 50 probes per person. The feature detection module successfully located the nose tip (within 20mm of the manually annotated location) on 5872/5968 = 98.4% of the images for this experiment. We attribute the slightly higher feature

location performance to the increased number of neutral expressions present in the data set.

Results for this set of experiments are summarized in Table 6. The rank-one recognition rates are only slightly higher for a pure neutral-expression gallery than for a pure happy-expression gallery, and a mixed-expression gallery always gives the highest performance for a given gallery size.

Experiment 2 contains a probe set that contains more than four times as many neutral expressions than non-neutral expressions. This difference favors the neutral gallery over the happy gallery for rigid matching, resulting in higher performance for neutral expression matches. Table 7 provides a more detailed description of the performance. Sub-experiments 1 and 2 show that a neutral gallery image performs best when matched to a neutral probe image. Conversely, a gallery composed of a happy expression performs better than a neutral gallery for every other expression. This conclusion agrees with the findings of Yacoob *et al.* [17]. Based on experimental results, we found that when an equal number of probe images containing a happy expression and probe images containing a neutral expression were compared, the probe set containing happy expressions resulted in a higher overall recognition rate. This suggests that it may be easier to reproduce a prompted happy expression than it is to reproduce a prompted neutral expression. In addition, the individual probe results in Table 7 for neutral and happy expressions show that for each gallery that contains at least one happy expression, probes containing a happy expression will demonstrate superior performance.

Table 7 shows a plateau in performance for non-neutral probe images with only three gallery images (H+H+H). As additional gallery expressions are added, no significant increase in rank one matches is produced. This conclusion would greatly

Table 7
Experiment 2 - Neutral and Happiness Experiment (5968 Probe Images)

Sub-Exp #	Gallery Type	N (4834)	H (507)	Sp (182)	D (95)	Sd (88)	Other (262)
1	N	3975	400	129	59	53	170
2	H	3804	468	147	76	67	206
3	N+N	4348	443	164	75	74	223
4	H+H	4281	492	165	79	72	223
5	N+H	4313	485	164	79	72	224
6	N+N+N	4549	453	169	78	78	229
7	N+N+H	4532	486	169	82	75	236
8	N+H+H	4552	495	171	82	76	232
9	H+H+H	4455	500	171	81	79	228
10	N+N+N+N	4619	454	174	82	80	231
11	N+N+N+H	4613	487	173	82	79	237
12	N+N+H+H	4592	495	172	82	77	238
13	N+H+H+H	4710	500	176	82	80	236
14	H+H+H+H	4550	500	170	80	78	231
15	N+N+N+N+N	4648	456	170	80	79	232
16	N+N+N+N+H	4672	486	174	82	80	236
17	N+N+N+H+H	4666	495	174	82	79	239
18	N+N+H+H+H	4723	500	177	82	81	240
19	N+H+H+H+H	4720	501	177	82	80	237
20	H+H+H+H+H	4586	501	174	82	79	229

benefit from a larger more diverse data set containing additional images with non-neutral expressions. Another observation from the results contained in Table 6 is that expression variation in the gallery combined with an increase in the number of images (or chances to match) provides the best opportunity for accurate recognition. Peak recognition performance (97.2%) is achieved with the gallery combination of N+N+H+H+H, however this is not a statistically significant difference from using the 4 gallery combination N+H+H+H (96.9%). In addition, the results show that for a given size gallery (*e.g.* 5), the heterogeneous galleries (4N+H, 3N+2H, 2N+3H, N+4H) all outperform the homogenous galleries (5N or 5H). These rank one recognition results are effectively mirrored in the verification experiment (TAR at 0.1% FAR) across the various gallery combinations. For verification, the optimal gallery is number 18 (N+N+H+H+H) and has a 91.5% TAR with a FAR of 0.1%.

Table 8
Experiment 3 - Multi-Instance vs. Single Gallery

Sub-Experiment Description	Probe Subset	Rank One Recognition Rate
Multi-Instance	Full - 5968 images	97.2%
REFER	Full - 5968 images	96.2%
Chang <i>et al.</i> [4]	Full - 5968 images	86.9%
Single Frontal Image	Full - 5968 images	83.8%
Multi-Instance	Neutral only - 4834 images	98.3%
REFER	Neutral only - 4834 images	97.0%
Chang <i>et al.</i>	Neutral only - 4834 images	91.9%
Single Frontal Image	Neutral only - 4834 images	90.3%
Multi-Instance	Non-Neutral only - 1134 images	95.1%
REFER	Non-Neutral only - 1134 images	93.1%
Chang <i>et al.</i>	Non-Neutral only - 1134 images	69.1%
Single Frontal Image	Non-Neutral only - 1134 images	57.1%
Multi-Instance	Happiness only - 507 images	98.2%
REFER	Happiness only - 507 images	92.1%
Chang <i>et al.</i>	Happiness only - 507 images	60.8%
Single Frontal Image	Happiness only - 507 images	47.5%

4.3 Multi-Instance Gallery Versus Component-Based Approaches

This set of sub-experiments compares our multi-instance gallery approach against two earlier component-based approaches to 3D face recognition [4, 5, 10], as well as to a simple single-frontal-face-region approach. The probe set for this set of experiments is the same as the set for the experiments in the previous section. We look at performance for the overall probe set of 5,968 images, a neutral-only probe set of 4,834 images, a non-neutral probe set of 1,134 images, and a happy-only probe set of 507 images. The gallery for the multi-instance approach is the N+N+H+H+H gallery from sub experiment 18 in the previous section. The gallery for the other approaches is the first neutral-expression image for a given subject, which is the typical convention for such approaches

The “baseline” performance result (labeled Single Frontal Image) is presented by

performing ICP on a single whole frontal region of the face image used for gallery enrollment and a smaller whole frontal region of another image is used as the probe. This comparison technique has been used in several papers [20, 4, 1], and represents the potential performance gain by using the alternative methods.

Table 8 shows a performance comparison of the best multiple instance results achieved in Experiment 2 (sub-experiment 18 seen in Table 6 containing five gallery images: two neutral and three happiness expressions), the results attained by the REFER algorithm (using only a single neutral image for each subject), and a baseline performance of a single neutral image in the gallery matched to a single frontal image (a sphere cropped at 100mm from the nose tip) in the probe. The first two rows show the overall similarity in terms of rank one performance between the two approaches. Both techniques perform well except for the final test which employs a probe set consisting of only images that display a happiness expression. In this test, the disparity in the rank one recognition rate is due to the number of happiness expressions present in the gallery. In the REFER setup, exact rigid matching is not possible due to the forced expression variation between gallery (neutral) and probe (happiness). In contrast, the Multi-Instance setup contains three images displaying a happiness expression, which equals three times the chance of encountering a similar happiness variation in the gallery. The results in Table 8 also demonstrate the poor performance of the baseline experiment (labeled Single Frontal Probe Image). For a neutral gallery matched to a neutral probe, the rank one recognition rate is comparable to the other two methods. However, when alternative expressions are introduced to the probe set, the limitations of the single full frontal matching method become apparent. Finally, each REFER and multi-instance pair of results in Table 8 shows a statistically significant different rank one recognition rate. The z-scores for these experiments can be seen in Tables 9 10 11 12. This allows us to conclude that if multiple expressions are available for gallery participation, higher

recognition rates can be achieved, especially if the probe images contain the same expressions. However, if only a single neutral image is available, the performance decrease is not substantial when using the REFER method.

Table 9
z-Scores: Multi-Instance vs. Single Gallery - Full - 5968 images

	Multi Instance	REFER	Chang et al.	Single Frontal Image
Multi-Instance	N/A	3.05	20.79	24.96
REFER		N/A	18.26	22.57
Chang et al.			N/A	4.78
Single Frontal Image				N/A

Table 10
z-Scores: Multi-Instance vs. Single Gallery - Neutral only - 4834 images

	Multi Instance	REFER	Chang et al.	Single Frontal Image
Multi-Instance	N/A	4.21	14.57	16.96
REFER		N/A	10.95	13.50
Chang et al.			N/A	2.76
Single Frontal Image				N/A

Table 11
z-Scores: Multi-Instance vs. Single Gallery - Non-Neutral only - 1134 images

	Multi Instance	REFER	Chang et al.	Single Frontal Image
Multi-Instance	N/A	2.02	16.14	21.21
REFER		N/A	14.59	19.82
Chang et al.			N/A	5.92
Single Frontal Image				N/A

Table 12

z-Scores: Multi-Instance vs. Single Gallery - Happiness only - 507 images

	Multi Instance	REFER	Chang et al.	Single Frontal Image
Multi-Instance	N/A	4.52	14.75	18.15
REFER		N/A	11.74	15.46
Chang et al.			N/A	4.24
Single Frontal Image				N/A

4.4 Statistical Significance

In order to determine statistical significance for experiments performed in this paper, we use a standard Z-test, which has been used in similar situations by Yan *et al.* and Yambor *et al.* [29, 30] who based their use of the statistic after Devore and Peck [31]. However, this test is valid only on assumption that the binomial distribution converges to a normal distribution. Likewise, the rank one recognition and verification rate comparisons can be viewed as a binomial distribution problem under the assumption that the probability (p) of a correct match (either correctly identifying or verifying a subject) is constant across all subjects. If p varied significantly the binomial assumption may be weak; however, empirical estimation of p from multiple subsets showed p to be reasonably consistent across the subjects. Assuming this, the probability of an incorrect match is $(1 - p)$. It is assumed that if a sufficiently large sample size (N) is used, the binomial distribution will converge to a normal distribution and therefore the population standard deviation can be approximated. For a large enough sample size N , a binomial variable X is equal to $N(N_p, N_{pq})$. Yan *et al.* [29] use the assumption that good comparative results are achieved when $N_{pq} \geq 3$. In experiment 1, $N = 3921$, and in experiment 2, $N = 5968$. For each set of results listed in Tables 3 and 6, $N_{pq} \geq 3$.

For identification result comparisons, \hat{p} is the empirical (sample) estimate of p , the probability that a subject is correctly identified from a gallery set. For verification result comparisons, \hat{p} is the probability that a probe to gallery match results in a true accept at a given false accept rate (*i.e.* 94.5% at a 0.1% FAR would result in $\hat{p} = 0.945$ for that comparison). Verification experiments tend to have a higher N as each comparison is included in the total number of samples. Identification experiments only contribute a single sample for each probe to gallery set match. This discrepancy causes experimental verification results to require a much smaller difference than identification results to be statistically significant. Given two results, with sample sizes N_1 and N_2 , and the percentage of observed correct matches as \hat{p}_1 and \hat{p}_2 the test statistic for $H_0 : p_1 = p_2$ using a 0.05 level of significance is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{N_1+N_2}{N_1N_2}\right)\left(\frac{X_1+X_2}{N_1N_2}\right)\left(1 - \frac{X_1+X_2}{N_1N_2}\right)}}$$

$$X_1 = \hat{p}_1 \times N_1, X_2 = \hat{p}_2 \times N_2$$

If $z \leq 1.64$ then it is assumed that there is no statistically significant difference between the pair of results and their associated sample sizes.

5 Conclusions and Future Work

In this paper, we examined the value of using multiple gallery images per subject to increase the performance of 3D face recognition. This study employs the largest 3D face data set available to date with over 13,450 total images. Our experimental results (seen in Tables 6, 3, and 8) support many conclusions. First, using multiple images to enroll a person in a gallery can improve the overall performance of a biometric system, at least up to the level of four or five images tested here. We have not yet found a clear “plateau” in performance for the number of images used to enroll

a person, however we saw a statistically significant difference between using any single image compared to using multiple images in the gallery. Next, Experiments 1 and 2 demonstrated that when using multiple images to enroll a person, sampling from different expressions (heterogeneous) improves performance over sampling only the same expression (homogeneous). Finally, the results have shown that the types of expressions in the gallery become less important as the gallery increases in size.

The results in this paper show that a homogeneous gallery composed of images containing neutral expressions outperforms a gallery containing the same number of happy images in a verification experiment. In an identification experiment, no statistically significant difference was found between either of the two gallery compositions. Finally, we show that a statistically significant difference in recognition rate exists when using a single region from multiple galleries and using multiple regions from a single neutral gallery for 3D face recognition. While the difference is not substantial (97.2% vs. 96.2%), using a single region from multiple galleries results in superior recognition performance.

In this paper we have shown how the performance of a recognition system can benefit by varying the number and types of expressions in the gallery of real subjects. A logical extension of this work would be to combine the REFER and multiple gallery methods into a single technique. We have experimented with this fusion and have found that we require a larger and more challenging data set in order to see statistically significant gains in recognition performance. With the ND-2006 data set, our multi-instance approach is able to achieve a 97.2% rank one recognition rate and REFER is able to achieve a 96.2% rank one recognition rate. When these methods were combined, we saw gains in these results; however, none were of a statistically significant margin. In addition, combining these algorithms results in a method that

is 5 times slower than the original REFER algorithm, performing a total of 140 ICP matches.

Another extension of this work is to determine how well images containing synthetic expressions perform compared to real expressions. Lu *et al.* [16] propose a method of using subject specific facial surface deformation models to synthetically create 3D face images containing a variety of pre-trained expressions. This process would reduce the need for subject enrollment cooperation and possibly boost results over those attainable by any single image.

Although this paper reports results using the largest 3D face data set available, these conclusions would benefit from further experimentation. The first problem is the actual acquisition of images containing expressions. In a controlled setting, getting prompted expressions is challenging. Variation between the same prompted expression can be caused by numerous things outside of the operators control (mood, weather, or personal preference). One day the subject may interpret a prompted expression in one manner, and the next day it may be different. One of the most exciting avenues to further this line of research is through video recognition. By capturing the changes in expressions, we would be able to determine all possible ways that a region of interest can change. The experiments in this paper used permutations of up to 5 different expression, however these were the final expressions and did not represent the full range of motion in the face. For example, if a subject starts with a small smile and slowly shifts into the largest possible grin, any future image containing a smile should be recognized as part of the smile subset. This would be done for other expressions to create what we determine as a total range of facial motion. With this subject specific knowledge, recognition challenges stemming from expression variation may be addressed.

6 Acknowledgements

Biometrics research at the University of Notre Dame is supported by the National Science Foundation under grant CNS-0130839, by the Central Intelligence Agency, by the National Geo-Spatial Intelligence Agency, by UNISYS Corp., and by the US Department of Justice under grants 2005-DD-BX-1224 and 2005-DD-CX-K078.

References

- [1] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, and G. Medioni. Performance of Geometrix ActiveID 3D Face Recognition Engine on the FRGC Data. *Face Recognition Grand Challenge Workshop*, pages III: 154–154, 2005.
- [2] P. Besl and N. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:239–256, 1992.
- [3] Xiaoguang Lu, Anil K. Jain, and Dirk Colbry. Matching 2.5D Face Scans to 3D Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):31–43, 2006.
- [4] K. Chang, K. W. Bowyer, and P. Flynn. Multiple nose region matching for 3D face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–6, 2006.
- [5] T. Faltemier, K. Bowyer, and P. Flynn. 3D face recognition with region committee voting. *Proc. the Third International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pages 318–325, 2006.
- [6] X. Chen, T. Faltemier, P. Flynn, and K. Bowyer. Human face modeling and recognition through multi-view high resolution stereopsis. *Proc. IEEE Workshop on Biometrics (affiliated with CVPR2006)*, pages 17–22, 2006.

- [7] A. S. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1584–1601, 2007.
- [8] A. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:748–763, 2002.
- [9] B. Heisele and T. Koshizen. Components for face recognition. *Proc. IEEE International Conference on Face and Gesture Recognition*, pages 153–158, 2004.
- [10] T. Faltemier, K. Bowyer, and P. Flynn. A region ensemble for 3D face recognition. *Under Review*, 2007.
- [11] C. Heshner, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range images. *Seventh Intl Symposium on Signal Processing and Its Applications*, 2:201–204, 2003.
- [12] James L. Wayman. A path forward for multi-biometrics. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1069–1072, 2006.
- [13] ND-2006 Face Data Set. <http://www.nd.edu/~vrl/>. 2007.
- [14] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.
- [15] K. 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.
- [16] X. Lu and A. Jain. Deformation modeling for robust 3D face matching. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2006)*, pages 1377 – 1383, 2006.

- [17] Y. Yacoob and L. Davis. Smiling faces are better for face recognition. *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 59–64, 2002.
- [18] B. Gokberk and L. Akarun. Comparative analysis of decision-level fusion algorithms for 3D face recognition. *Proc. 18th International Conference on Pattern Recognition*, pages 20–24, 2006.
- [19] C. Beumier and M. Acheroy. Face verification from 3D and grey level cues. *Pattern Recognition Letters*, 22(12):1321–1329, 2001.
- [20] P. J. Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 947–954, 2005.
- [21] J. Cook, V. Chandran, and C. Fookes. 3D face recognition using log-gabor templates. *Proc. British Machine Vision Conference*, pages 83–83, 2006.
- [22] Kin-Chung Wong, Wei-Yang Lin, Yu Hen Hu, Nigel Boston, and Xueqin Zhang. Optimal linear combination of facial regions for improving identification performance. *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, 2007.
- [23] O. Melnik, Y. Vardi, and C. Zhang. Mixed group ranks: preference and confidence in classifier combination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 973–981, 2004.
- [24] S. Jeong, K. Kim, B. Chun, J. Lee, and Y.J. Bae. An effective method for combining multiple features of image retrieval. *Proc. IEEE Region 10 Conference: TENCON99*, pages 982–985, 1999.
- [25] T.K. Ho, J.J. Hull, and S.N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 66–75, 1994.

- [26] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, 2005.
- [27] Konica-Minolta USA. Minolta Vivid 910 non-contact 3D laser scanner. <http://www.minoltausa.com/vivid/>. accessed 2004.
- [28] A. Jain, X. Lu, and D. Colbry. Three-dimensional model based face recognition. *Proc. 17th International Conference on Pattern Recognition*, I:362–366, 2004.
- [29] P. Yan and K.W. Bowyer. Empirical evaluation of advanced ear biometrics. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pages 41–48, 2005.
- [30] W. Yambor, B. Draper, and R. Beveridge. Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures. *Empirical Evaluation Methods in Computer Vision*, pages 1–14, 2002.
- [31] J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data, Third Edition*. Brooks Cole, 1997.