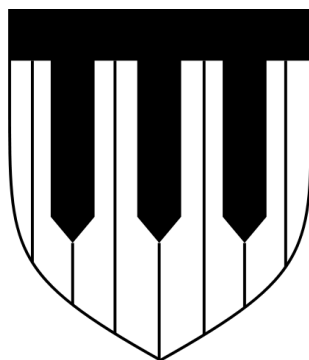


SURFACE MATCHING WITH STATISTICS AND GEOMETRY: TECHNICAL REPORT FOR 2011

Roy Schestowitz*

January 8, 2012



*Dr. Roy S. Schestowitz holds a Ph.D. in Medical Biophysics, which he received from the Victoria University of Manchester where he specialised in statistical analysis of shape and intensity characterising soft tissue. He also worked on a novel approach for assessing dissimilarity using combined models of 2-D faces and 3-D brain data (notably MRI) before working on cardiac MRI for real-time tracking purposes.

Abstract

Statistics of disparate and absolute parts of the human face are a complex area of exploration due to high variation which is caused by facial expressions. There have been studies – despite scarcity in numbers – into how this variation can be modeled, but there is not sufficient consideration of different paradigms for studying this variation. Generalised Multi-Dimensional Scaling (GMDS) can overcome this by considering image surface rather than handle the complexity introduced by applying directional decomposition in a high-dimensional hyperspace. In both 2- and 3-D, information about depth can be used, although in the latter case this information is accurate, whereas in the former case there is reliance on estimation based on shadows or stereo vision, i.e. multiple angles. Three-dimensional methodologies usually rely on accurate measures that are not just relative but also absolute, meaning that the location of objects in the image should be capable of alignment wrt other images too. The application of these ideas in areas such as face analysis – including recognition, modeling, synthesis, and interpretation – is seen as promising with the advent of new acquisition equipment and modalities. A lot of data is made available and exploitation of its full potential is made possible by accounting for large sets of data. The more data becomes available, the more viable it becomes to study the statistics of faces and make inference based on the learnt information. Our attempt to reproduce some of the results of F. Al-Osaimi *et al.* and furthermore improve

them using other methods and different datasets (with a 3-D scanner at our disposal), are described in this informal document, which in essence contains research notes for 3-D facial expression analysis through statistics (project starting 2011). It is work in progress¹, so this text is eternally an informal draft that deals with comparing a principal component analysis (PCA) approach to a GMDS approach. Shall the goal be met by reasoning about the advantage of the latter, portions of this document may prove handy.

¹A lot of material in textual form was being assembled throughout development, including technical explanations and explanations in the form of visual elements that demonstrate textual descriptions (e.g. tables, images, screenshots).

<i>CONTENTS</i>	4
-----------------	---

Contents

1 Overview	44
1.1 A PCA Approach	45
1.2 Rationale	45
1.3 Goals	46
1.4 Document Structure	47
2 Existing Work/Literature Survey	48
2.1 Generalised Multi-dimensional Scaling	49
2.2 Expression Deformation	50
2.2.1 Viola-Jones	53
2.2.2 Similar Work	54
2.2.3 Neutral Images as Reference	55
2.2.4 Similar Work on Image Sequences	55
2.2.5 PCA-based Models	56
3 Methods	62
3.1 Model-building	62
3.2 Data Alignment	63
3.3 PCA	63

<i>CONTENTS</i>	5
3.3.1 Robust Generalised PCA	64
3.4 PCA for Animation	66
3.5 GMDS vs. PCA	67
3.6 GPCA	74
3.7 Algorithm	75
3.7.1 Related Work	77
3.7.2 Wavelets (Texas University)	80
3.8 Outline/Thoughts About Operation	87
3.9 Systematic Experiments	89
4 Data	93
4.1 FRGC	93
4.1.1 Imperfections in Signal	93
4.1.2 Segmenting Parts of the Face	96
4.1.3 Isolated Faces	97
4.2 GIP	97
4.2.1 Experimental Framework for GIP Dataset	101
4.2.2 Additional Data on Demand	102
4.3 Synthetic Data	102
4.4 GIP Data Localisation	103

<i>CONTENTS</i>	6
-----------------	---

5 Implementation	104
-------------------------	------------

5.1 Preparation and Preprocessing	107
5.2 Normalisation	109
5.3 Expression Models	111
5.3.1 Reproducibility Concerns	111
5.3.2 Apparent Limitations	112
5.4 Registration	113
5.5 Modelling	114
5.5.1 PCA	115
5.5.2 GMDS	115
5.6 Control Files	116
5.7 Graphical User Interface	116
5.8 Remote Access to the Program	117

6 Experimental Framework	119
---------------------------------	------------

6.1 Validation	120
6.2 Residual	120
6.2.1 Residue Filtering	122
6.2.2 Further Data Preparation	123

6.2.3 Binary Masks	123
6.3 PCA and Projection	124
6.4 ROC Curves	126
6.5 Benchmarks	129
6.6 Extensive Work	129
6.7 FRGC Experiments	130
7 Ongoing Progress and Results	131
7.1 Visualisation	133
7.2 Statistical Analysis	134
7.3 Detection	134
7.4 Automation	134
7.5 Similarity Measures	135
7.5.1 EDM Versus Pure Residual Approach: Correlation Be- tween PCA-based Approaches	135
7.6 Performance	136
7.6.1 Performance with GPU Boost	137
7.6.2 Dataset	137
7.6.3 Data Deficiency	138
7.7 Benchmarks	141

7.7.1	FRGC 2.0 Experiment 3	143
7.8	Full Model (EDM) for FRGC Data	150
7.8.1	Newly-built EDMs	151
7.9	ICP	177
7.9.1	ICP Experiments	182
7.10	Systematic Experiments	184
7.10.1	Residue Adjustments	196
7.10.2	ROC Curves	196
7.10.3	Initial ROC-based Benchmarks	202
7.10.4	Downsampled Images for PCA	209
7.10.5	Model-based approach	210
7.10.6	ICP Revisited	226
7.10.7	New Similarity Measure	226
7.10.8	Effects of Lambda Changes	228
7.10.9	Debugging ICP	230
7.10.10	Translation Explored	252
7.10.11	Multi-feature PCA	253
7.10.12	Multidimensional Scaling - Animated Example	265
7.10.13	Exploratory GMDS Integration	269

7.10.14 Full-face PCA	295
7.10.15 GMDS on Smaller Face Parts	297
7.11 Texas Database	299
7.11.1 Geodesic Cutoff	335
7.11.2 Scale Issues	353
7.11.3 Improving GMDS	360
7.12 Planning for Final Stages	366
7.12.1 Caching Code	369
7.12.2 Backporting	382
7.12.3 C++ FMM Debugging	384
7.12.4 Workaround implemented	387
7.12.5 Resolution Increases	388
7.12.6 Smoothing	394
7.12.7 Resolution increased	395
7.12.8 Residuals	411
7.12.9 Higher Resolution	415
7.12.10 2006 Experiments and Geodesic Masks	419
7.12.11 Preparing Larger Experiments	421
7.12.12 False Positive - a Dilemma	424

7.12.13 Fallback Discriminant	426
7.12.14 Making a More Stable Classifier	427
7.12.15 Occlusion Based on FMM for Matching	428
7.12.16 More FMM Results	434
7.12.17 FMM-based Dissimilarity	435
7.12.18 Rotation	437
7.12.19 More Data Points	438
7.12.20 Geodesic Slices	438
7.12.21 Surface Signatures	439
7.12.22 Vectorised Signatures	440
7.12.23 Hybrid and Bugs	462
7.12.24 Alternations to the Algorithm	462
7.12.25 Eyes vs Nose	462
7.12.26 Exact geodesic_library	466
7.12.27 Removing Cases of Uncertainty	467
7.12.28 Recognition Results	467
7.12.29 Number of Vertices vs Recognition Rates	469
7.12.30 Trial and Error in Parallel	470
7.12.31 Geodesic lenses	471

7.12.32 Weighting for Source Points	482
7.12.33 Weighted Similarity Measure	483
7.12.34 Early Results With Weighting	483
7.12.35 Nose Tip Revisited	484
7.12.36 Similarity Measure Variants	485
7.12.37 ROC Curve - Without Smoothing	485
7.12.38 Spectral Masks	486
7.13 Diffusion Distance	486
7.13.1 Spectral Rings	487
7.13.2 Values Reset	487
7.13.3 Diffusion in Facial Features	487
7.13.4 Similar Work	490
7.13.5 Gabor Filtering In Combination With Diffusion Distance	492
7.13.6 Mask Dilation Approach	506
7.13.7 Dilation Range	506
7.13.8 Automatic 3=D Correspondence Finding	508

List of Figures

1	Expression parameterisation in action (image from Al-Osaimi <i>et al.</i>)	51
2	The effect of varying the first (top row), second, and third parameter of a brain appearance model by ± 2.5 standard deviations	60
3	The analysis performed by a Robust Generalised PCA algorithm	65
4	The proposed framework for GMDS improvement	69
5	A closer look at the GMDS approach	70
6	Crude visual example of how typical PCA and GMDS relate to one another, approach-wise	71
7	Example face from the FRGC dataset	94
8	3-D Image example from the FRGC dataset, demonstrating points on the side of the face – points which need to be removed	95
9	Example face with holes remaining in the data	98
10	Another example	99
11	Same as above, different angle	100
12	Program steps broken down into an overview-type flowchart .	105
13	A replotted block diagram of the components in the IJCV paper (top) and our proposed extension/modifications (bottom), and the already-implemented procedures	106

14	Translation of the given (cropped) face applied so as to position it with the nose tip at the front and at the centre	154
15	A before/after overview	155
16	Early prototype of the GUI	155
17	The same GUI at a later date	156
18	The shape-residual extracted from two different images of different people, where the faces are aligned so as to fit a common frame of reference.	156
19	A sample image and a corresponding residual wrt to another (unseen) image	157
20	Example of what happens when the nose is incorrectly identified	157
21	Examples of challenging residuals that have a lot of noise . . .	158
22	Process of cleaning up the residual of two images (GIP data) .	158
23	Examples of residual images before and after outliers removal .	159
24	Examples of faces that, given the default set of parameters, do not detect the nose correctly (with old-style cropping) . . .	159
25	Improved cropping of faces that takes spatial measurements into account	160
26	Example of binary masks being applied to image residue . . .	160
27	Examples of 4 raw residuals being reduced (notice the Z scale) using thresholds and masks	160

28	Cropping of GIP data shown on the top left	161
29	Model modes decomposed for a couple of GIP datasets (ab- breviated to account for top 10 modes alone)	162
30	Example 3-D representation of an arbitrary face image from FRGC 2.0	163
31	Example of alignment and cropping of the rigid parts of an- other face (mind axes scale) with the result shown at the top surface and right image inside the GUI	164
32	Example of alignment and cropping of the rigid parts of an- other face with the result shown at the top surface and right image inside the GUI	164
33	Example of alignment and cropping of the rigid parts of an- other face there there is some noise and hole that pose a challenge	165
34	Decomposition based on sample GIP data (Pareto)	165
35	Decomposition based on just 4 registered faces with different expressions (Pareto)	166
36	Same as prior figures, but with 90 images in the set	166
37	Overview of an experiment dealing with expression-to-expression variation model-building	167

38	The masks used to crop residuals in the FRGC dataset. The left-hand one is more restrictive and selective in the sense that it omits some of the data associated with the face near the edges.	167
39	The top row shows images of the same subject and the bottom one is a group of hard cases (image from Huang <i>et al.</i>)	168
40	Left to right: Texture image mapped to 2-D, 3-D representation, and cropped parts (for alignment)	168
41	Example of an image where the face does not fit the image frame, unlike the example at the top right	169
42	Image residue incorrectly cropped by a binary mask	169
43	Examples of image residues from the FRGC datasets	170
44	Image residues from the FRGC datasets shown from frontal angle	170
45	The image shows a matrix corresponding to two things; on the left there is a top-down view of what is shown on the right. The top part shows how the 15,000 sampled (cloud)points get distributed after dimensionality reduction and the bottom part relates to the magnitude of the principal components, where the red parts show higher deviation from the mean. It is fairly smooth.	171
46	Principal components as a function of datapoints (log scale) .	171

47	Point-to-point correlation along the 10th utmost principal axis	172
48	Score per sample point, mapped back from vectorised form to the image grid	173
49	The tenth principal component derived from PCA, as visualised based on the reshaped (previously vectorised) image residues	176
50	The tenth image residual from the Experiment 3-built EDM projected onto the EDM space's most significant component .	177
51	Example of an evaluation experiment for photometric ICP . .	178
52	ExamExample points cloud for ICP to register	184
53	On the left: two faces (with binary masks cropping them for rigid parts like nose and forehead) overlaid for ICP; on the right: same from another angle	185
54	Examples of the faces used tor training and recognition, with neutrals on the left and smiles on the right	189
55	Examples of the program with the new data and methods in place	192
56	Pixel-wise difference between the images from our expressions set and resultant corresponding images following ICP-found translation	193

57	The two images that automatic detection struggles with (because of the hair)	194
58	The first six neutral images taken from the set and cropped by the algorithm correctly	195
59	The first 6 images in the set with a narrow mask used to extract and attain a neutral-to-non-neutral residue	197
60	Same as the previous figure, but with only 5 images. The top row shows the effect of using a broader mask and the bottom part shows the effect of applying a fixed mask and thresholds to make the data more trivially comparable.	197
61	ROC curves plotted for just 13 tests done on the FRGC datasets with expressions isolated	198
62	Preliminary test where several images (not complete set) are used to get a rough idea of what the ROC curves will look like	199
63	A somewhat larger test on non-neutral sets, where ICP based on PCA is used for alignment, then mean of residuals get used as a similarity measure	200
64	The same type of comparison with the same type of set (as in Figure 63) but with a more cunning similarity measure and an example of the X data at the bottom right	201
65	A combined view of X, Y, Z, the image before ICP, after ICP, and the reference image	202

66	Difference images of the first 5 pairs taken from the same people	203
67	Difference images of the first 12 pairs taken from different people	204
68	ROC curve of the 17 images from figures 66 and 67	205
69	The curve showing the performance for 83 pairs from false matches and 37 from true matches	206
70	Images <code>~/NIST/FRGC-2.0-dist/nd1/Fall2003range/04557d337.abs</code> and <code>~/NIST/FRGC-2.0-dist/nd1/Fall2003range/04557d339.abs</code> , where there is some detection difficulty	207
71	Example face-to-face comparisons	208
72	The ROC curves comparison on the left as linear scale and on the right log-scaled	208
73	The results in terms of recognition rate after widening the mask and also changing from median to mean	215
74	The results from full-resolution image sets and low-resolution equivalents (as seen at the top left)	216
75	The FP analysis of the results of a model-based approach, with the breakdown of modes shown at the top right	217
76	Performance comparison between an approach where the median of squared differences gets compared to mean of model changes	218

77	Performance of recognition when the absolute differences are gathered by their median	219
78	Performance of recognition when the squared differences are gathered by their means	220
79	Degraded performance when the compared face pairs are not ones that were used to train the PCA model	221
80	Comparison assessment with a large set of false pairs. The results of random pairs versus unseen pairs with expression differences.	222
81	The results of matching random pairs from different people and from similar people, with and without expression (based on expression models)	232
82	The results of comparing correct pairs to random (and false) pairs using the model-based approach. The right hand side shows the breakdown of model modes.	233
83	Performance measured on relatively small sets, empirically showing that coarser grids yield better recognition performance	234
84	Decomposition of the different modes of variation for the three cases, namely granularity levels 6x6, 8x8, and 10x10, respectively (10 pixels/voxels apart) demonstrating that despite the changes in resolution the model modes have a similar distribution and are probably inherently similar, as expected	235

85	Comparison between the performance of the method with smoothing applied before sampling and without any smoothing at all (which gives similar performance)	236
86	The decomposition of the model as a chart corresponding to Figure 84 on the right, this time with smoothing on	237
87	The first few face residues following alignment with ICP (sample points being around the forehead, nose, and eyes)	238
88	The results of measuring the similarity by determinant of the eigenvalues of the covariance matrix and engaging in a recognition task	239
89	Results of an experiment where the determinant is again being explored, this time with a larger set	240
90	Results of an experiment where the determinant is again being explored with a comparison of the curves for 3 values of δ . . .	241
91	The result – in terms of performance – of varying n in $\lambda_{1 < i < n}$.	242
92	The effect of changing the value of δ on the overall recognition performance	243
93	An 8x8 separation between points in the image (shown from two angles), with downsampling done for debugging purposes .	244
94	A slice or subset of the data being used for ICP (on the left) and the masked face from which it is extracted (right)	244

95	Top: Two images taken from the same individual being compared when there is insufficient compensation for noise. Bottom: another set of such images but where smoothing is applied to reduce noise-imposed anomalies	245
96	On the left: the result of poor or buggy ICP (difference); on the right, an image is shown of the type of image we expect to have and also get when ICP performs well	246
97	Difference between the first 4 images before and after ICP (rotation and translation), with two of the first reference images shown at the bottom just for a sense of what the images at the top are derived from	247
98	The effect of the bug demonstrated by showing misalignment on the X axis (and to a lesser degree in Y too).	247
99	The new distribution of modes following the bugfix	248
100	The effect of perturbing the points on ICP	249
101	The effect of noise on ICP studied by aligning images 1-5 at the top to images 6-10 at the bottom	250
102	The purely median-based performance on the Spring Semester set, without ICP	251
103	The purely model-based (determinant) performance on the Spring Semester set, without ICP	256

104	Example differences between an image before and after translation (in all three dimensions)	257
105	On the left: the results from a test run (first 20 images) using the determinant-based objective function. The model was not constructed with translation, whereas matching did. On the right: the same but with a median-based similarity measure. .	258
106	A top-to-bottom view of one's face (the rigid part) with corresponding translation and rotation	259
107	A look at some of the tweaking and debugging process of ICP, where the angle shown is pointing from underneath the nose, going towards the top	259
108	An example of the first image pair, visualised separately as stripes as coarse as the image sampling rate (for the model) .	260
109	The results of a mis-constructed experiment where ICP did not work correctly and nonetheless, the model-based approach did not fail so miserably	260
110	Aligned and misaligned derivative difference (Y only)	261
111	Multi-feature experimental data (y-only derivative)	262
112	Y derivative (left) and X derivatives (right)	263
113	A couple of faces with the Y derivative on the left and the X derivative of the Y derivative (result of a bug) on the right . .	264

114	An example Y derivative image before (left) and after (right) signal enhancement	265
115	The result of a very crude experiment on Fall Semester datasets, which build a PCA model of derivative differences and then perform recognition tasks on unseen faces. ROC curves are shown on the left, the composition of the model is abstracted on the right.	266
116	The result of a buggy code creeping into experiment as in 130 (incorrect values were sampled). ROC curves are shown on the Left, the composition of the model is abstracted on the right.	267
117	The result of a correct code dealing with an experiment like in 130 but with data from the Fall Semester	270
118	The effect of stress minimisation of the shape of a cat	271
119	Randomly chosen face sampled 10 point apart along each dimension	271
120	Example of almost randomly selected distances along the shapes	272
121	Improved selection of distances (787 vertices) and the effect of MDS reducing the stress	272
122	Top: original image. Bottom (from top to bottom, left to right): stress minimisation with MDS, one iteration at a time .	273

123	A look at the cruder among ways to perform a comparison between faces	274
124	An exploratory look at how applying MDS to face images of the same subject depends on presupplied distances	275
125	Transformation from 3-D face (left) to a subset of rigid parts and then GMDS handling of the underlying surface (right) . .	275
126	Nose and eye regions from different people (FRGC 2.0) as treated by GMDS ($N = 50$)	276
127	Nose and eye regions from different people (FRGC 2.0) as treated by GMDS when $N = 100$	276
128	Nose and eye regions of the same person (FRGC 2.0) as treated by GMDS	276
129	The first pair in the set of real matches (same person in dif- ferent poses)	277
130	An example of a problematic pair with a false signal spike (left)	277
131	A view of the program's front end (framework wrapper)	278
132	A view of the handling of image pairs and their comparison using GMDS	279
133	A simple visualisation of the algorithm's processing of images, by numbers	280

134	The correspondence problem in GMDS and an abstraction of the data by consideration of a top-down representation	281
135	Performance tests on very basic GMDS algorithm applied to rigid face parts	282
136	Examples of some of the bugs encountered and overcome while working on GMDS implementation for faces	283
137	Set of results for 10x10 grid sampling (GMDS)	284
138	Early performance measures for GMDS more properly done . .	285
139	Larger scale examples of early performance measures	286
140	Results from poor PCA model, obtained using GMDS	287
141	Model modes distribution, corresponding to Figure 140	288
142	Model modes distributions (of 10 people and 76 people), built with the proper weight, albeit with very heavy and sometimes excessive smoothing	289
143	Preliminary results from GMDS-based recognition with full face surface	296
144	A look at an alternative mask which focuses on the nose and inner eye only	297
145	Recognition results based on the mask from 144 with GMDS .	298
146	A nose-only mask, which omits areas with potential of facial hair (the examples at the centre and the left are not related) .	299

147	The performance attained by applying GMDS just to the nose region	300
148	Example of the effect of ICP-induced rotation on the Voronoi cells	300
149	Return to the old mask with additional rotation, which does not yield better results than those at the region of 92%-98% recognition rate	301
150	Cheek inclusion gradually staged in for understanding of its impact on recognition performance (geodesics and PCA) . . .	301
151	The stress map corresponding to the new binary mask (with 150 points for FMM)	302
152	Stress map and the corresponding faces (looking from beneath the nose) from which it is derived	302
153	An example of a false pair (different people) and a cleaned up stress map showing some interesting patterns	303
154	Another example of a false pair and the results of GMDS . . .	303
155	Example of a bug found in the program, leading to massively false correspondence upon the same person	303
156	An example of acceptable matching between two poses of the same person	304

157	Another example of a bug found (and resolved) after it had proven problematic to recognition rates	304
158	A look at the problem associated with narrow faces that lead to incompatible sampling	305
159	General program settings used for the subsequent experiments	306
160	Results of a large-scale test after previous bugfixes	307
161	Example of a correspondence problem in a pair of images (one image on the left, another on the right). The top 4 images show the correspondence after the bugfix for one pair and the bottom 4 show the outcome of applying a fix to another pair. .	308
162	Example of a problematic pair where hair obstruction and nose position compared to the forehead caused an issue which is now properly addressed	309
163	Example of a pair where the side of the face got sampled, leading to serious issues (top) before they got resolved (bottom)	309
164	Comparison between images of the same person, where the height of the nose relative to the cropping is causing issues . .	310
165	The images corresponding to the above example (same person, different positions)	311
166	Another example of a problematic example where the score borders on being seen as “no match” even though it is	312

167	Some recognition results from the above experiments, with denser sample on the right where the cheeks were also removed to test their impact on performance (little impact)	313
168	Smaller-scale and large-scale (right) experiments that look at how applying the methods only to the training set (many identical faces clustered together) changes the above results. It does not affect them much.	313
169	Standard program settings with which to run the Texas data .	318
170	An example of GMDS applied to just a vertical slice of the data taken from different individuals	319
171	Exploratory work around GMDS applied solely to the nose region of different people (left and right), shown from different angles	319
172	Initial experiments with the Texas3DFR Database excluded the cheeks, which were later added as various parameters were studied for their impact	320
173	Texas3DFR Database pairs with the correct correspondence .	321
174	Model modes with more than 1% variation built from correct pairs	322
175	Model modes with more than 1% variation built from false pairs	323

176	With GMDS issues still in tact, the ROC curve for recognition suffers	324
177	A pair that GMDS usually fails on	325
178	Another pair that GMDS usually fails on	326
179	A closer, GMDS-style look on the very flawed correspondence-finding (example from Figure 222)	327
180	Another example of a GMDS-type comparison applied to a real pair and failing	327
181	Pairs of facial expressions from the same person, cut in half beneath the nose and tilted sideways, then shown with GMDS applied. The score (from left to right): 3.4866, 2.4497, 2.4718, 10.9726, and 173.6779	328
182	Some of the raw images (full face) after GMDS with 50 Voronoi cells displayed	331
183	A top-down view showing the matching and the corresponding score. with flipping manually corrected. The third example from the top got the topology completely upside down.	339
184	5 examples of experiments with synthetic data, where the top part shows the pair of images in their classic form, the middle shows a top-down view, and the bottom part is the range image	340

- 185 The scores in black show the pairings between different people
and in green are the scores of matches between the same person 341
- 186 A new visualisation form where the dots signify stress at the
given point 342
- 187 **Top:** A mapping of GMDS stress when cheeks are included in
match-finding. **Bottom:** same as above but cheeks excluded.
One can assume dark means low stress and white is high stress. 343
- 188 Original images, erroneous cropping effects (still in the process
of debugging) and retriangulation of the points after omission. 344
- 189 A toy example of a very small couple of surfaces cropped from
the centre of a face of the same person, where the pairs shown
correspond to top-down view and GMDS' results 344
- 190 Example of an augmented slice from a pair of faces and GMDS
applied to these 345
- 191 Results of a comparison between arbitrary bits where some
boundaries are a Euclidean cutoff and some are geodesic . . . 345
- 192 Results of a comparison between consistently chosen bits (near
the eye) where some boundaries are a Euclidean cutoff and
some are geodesic 345
- 193 Results of a comparison between surfaces that are mostly carved
out of a geodesic boundary 346

194	Results of a comparison between noses with a boundary defined by geodesic distance constraints	346
195	A preliminary look at a predominantly geodesic mask and how it separates pairs from different people (top) and pairs from the same person (bottom)	347
196	An experimentation with a mask that includes points around the forehead and around the nose	347
197	With just 600 vertices, the ROC curve shows unimpressive ability to distinguish between true pairs and false pairs	348
198	The effect of increasing the number of vertices to 2420.	349
199	Improved performance with slight changes in surface size for the gallery	356
200	The result of changing the border threshold for surface carving	357
201	The result of growing the surface too big	358
202	Performance with (left) and without averaging (right) of the arrange image for better sampling of the GMDS process	359
203	A look at slicing at geodesic boundaries around the nose tip, with coarse resolution on the left and improved resolution that isolates regions on the right	369
204	Newer Fast Marching algorithm as applies to a face from the Texas database	374

205	Newer Fast Marching algorithm as applies to TOSCO dataset	379
206	A bug with connected triangles	381
207	Two faces and the issue with triangulation	382
208	The data (top) and the inherent bug which incorrectly connects points (bottom)	383
209	A correctly connected pair of faces with the source point highlighted	384
210	GMDS using the older FMM implementation superimposed on top of new and incompatible code	385
211	Visualisation of the increasing number of vertices	388
212	Coarse resolution performance compared	390
213	A finer resolution-oriented set of results obtained from fewer runs than before	397
214	An example where the hair entering the surfaces can interfere with GMDS-based recognition (GMDS as a similarity measure)	398
215	Example where hair is at risk as being treated like skin surface, depending on the mask/s	399
216	The result of the nose tip being misplaced (original on the left, after masking on the right)	400
217	The problem of non-overlapping faces, a result of misregistration/misalignment	401

218	Registered and correctly aligned image	402
219	Example of a correctly sliced image subset (before geodesic boundaries cutoff)	403
220	High-density (vertices) surface and the images it is carved off .	404
221	Problematic image pairs	405
222	A pair that causes GMDS to fail	405
223	Problematic real pair (same person) where GMDS works but poorly so	406
224	A 3-D representation of a pair of images from the same person	406
225	GMDS failing to work as expected	407
226	A problematic pair which is seen as too different to quality as a match	407
227	ROC curve based on the smoothed surfaces variant of the al- gorithm	408
228	Example of some GMDS (mis)matches in the initial experiments	408
229	ROC curve based on the improved smoothed surfaces and somewhat better resolution	409
230	3 problematic image pairs	409
231	The area of collision in GMDS-based face detection	410

232	Examples of shape pair residuals and the corresponding ROC curve	411
233	Residual difference and the problem of localised high signal (which makes this a weak similarity measure)	412
234	ROC curve obtained by using a residuals of just a particular image region (nose and eyes)	413
235	Examples of pixel differences for pairs of the same people . . .	413
236	ROC curve corresponding to pixel differences for the whole middle section of the face	414
237	ROC curve corresponding to pixel differences for the nose area alone	415
238	ROC curve corresponding to sum of squared differences for the nose area alone	416
239	Example of 2 pairs from which the difference image is produced (shown at the top)	416
240	Top images show the sum-of-squared-differences of the first 3 true pairs, with the mere difference shown at the bottom . . .	417
241	Examples of the first 12 false pairs (sum-of-squared-differences)	418
242	ROC curve generated by a sum-of-squared-differences-based similarity measure	418

243	ROC curve generated by a sum-of-squared-differences-based similarity measure	420
244	Examples of matches between "true" pairs and other matches between "false" pairs (different people). The separation is not yet profound enough to get state-of-the-art recognition performance.	421
245	Example of similarity values after a Euclidean delimiter (above the eyes) was removed	421
246	Example of similarity values with more points	422
247	Number of points pushed higher towards 350 (near the maximal allowed value)	423
248	Pairing examples with false pairs, 1000 vertices on each	424
249	Pairing examples with false pairs, 2000 vertices on each	425
250	At the top left is just a naive implementation, the top right shows what happens when GMDS failures get detected and removed, and the large plot shows what happens when Euclidean measures are factored into this toy example.	430
251	4 problematic image pairs	431
252	Manually-measured width values for pairs of faces corresponding to different people	431

253	Manually-measured width values for pairs of faces corresponding to the same person	432
254	A geodesic ring/circle-based measurement as applied to tell apart anatomical equivalents from inequivalents	432
255	FMM is being used in a level sets-esque approach	433
256	An extension of the original (first) experiment which explored FMM (with Euclidean measures) as a classifier	434
257	Results from an extension of the range of radii/distances traversed from 20 to 50	435
258	Interim results (70 images) show 95% recognition rate with FMM-only (no GMDS) utility, but this tends to degrade as more difficult images are presented. Two good recognisers (classifiers), one of which is a Euclidean-geodesic hybrid, might give pretty good and mutually-independent results without using texture or fiducial points.	436
259	An example of two images from two different people, which nonetheless the FMM-based recogniser cannot quite detect as being different	437
260	An FMM-based recogniser results in nearly 90% recognition rate now (without GMDS)	437
261	The FMM recogniser ROC curve after increasing the number of true pairs	438

262	Example of a 10 degrees tilt	443
263	Recognition results from tilting one corresponding eye 360 degrees, then measuring distances on the geodesic boundaries . .	444
264	ROC curve based on comparing 220 images, where their Euclidean properties are measured upon geodetic slices	445
265	Brute force implementation that measures many geodesic distances	446
266	This figure visualises the idea of encoding surfaces as a vector not of surface vertices but an ordered list of Euclidean-upon-geodesic distances, which are fast to compute and sensitive to isometric/mildly detectable alterations	446
267	Separability testing in hyperspace	447
268	The image set of the first imaged individual in then test set, as an animation. The animation of the data from the 95th person is originally a GIF file.	448
269	Animation of the data from the 96th person	448
270	ROC curve obtained by measuring geodesic-Euclidean distances on the first imaged individual vs the same on different individuals	449
271	Smoothing versus <i>no</i> smoothing before measuring distances for identification purposes	449

272	The result of running the test set further (not for comparative purposes)	450
273	Animation of the data from the 103rd person. It is based on a set of images from the same person (numbered 103), without particularly challenging variation	451
274	The ROC curves based on a comparison between arbitrary (non-identical) pairs and pairs of images from the 103th person	452
275	The results following an increase in the smoothing range, demonstrating significantly degraded performance	457
276	An illustration by example of some images that prove to be challenging in the sense that their intrinsic properties are so similar that they almost get classified as being the same person (depending on how the threshold gets set)	458
277	Detection rate of my FMM-based method without GMDS as fallback (just annulling cases where fallback is invoked). X is log-scaled.	459
278	Results of GMDS applied to one single region rather than many, demonstrating the importance of having enough samples	460
279	FMM-based method without the use of ranges for fallback (and with some errors in pairs, which degrades the quality) . .	461
280	The result of applying the new method to pairs it feels confident enough comparing, based on pre-supplied thresholds . . .	473

281	The problem with GMDS not finding a path through the graph in some cases, where eye regions get altogether cropped out as a result	473
282	The performance one gets by handing difficult cases based on nose alone or eyes alone. The results from GMDS are similar to the results attained using the other method which is still undergoing development and gradual improvements, maybe with exact geodesics.	474
283	The performance attained by removing hard cases	475
284	The performance attained by changing the number of vertices and keeping all pair examples to be judged for similarity	476
285	Example of poor alignment in the original set	477
286	Two examples of easy matches from the remainder of the dataset (which was enrolled in its entirety into the experiment)	477
287	Example of geodesic differences map around the nose (to be improved)	477
288	Example of a thin FMM spiral	478
289	Four small examples of distances spiral in isolation	479
290	2 larger examples of distances derived from pairs of images of the same people	479

291	The distances spiral overlaid on the images it corresponds to (9th person in the set)	480
292	The distances spiral overlaid on the images it corresponds to (13th person in the set)	480
293	The distances spiral with larger, clearer points	481
294	Example of a true pair (same person) with a simplified repre- sentation of distances around each source point (FMM)	482
295	Another example of a true pair with simplified representation of distances around each source point (FMM)	482
296	An expanded view on the mis-correspondence between regions, where brighter shades represent greater disagreement between the pair taken from the same person	493
297	An expanded view on the mis-correspondence between regions, where the pair taken is from different people	494
298	Overview of the debugging process with examples from two true pairs (same person) and one false pair (different people), with the eye component discrepancies shown at the top and the nose at the bottom	494
299	The ROC curve obtained by using a weighted form of the similarity measure	495

300	ROC curve for the first phase of the experiment, which compares one-to-one (same person) and many-to-many (different people excluding this person, except in one case)	495
301	A broader scope curve for performance as in the previous figure	496
302	An example of misalignment in some parts of the nose in a true pair of images (same person), with the left nostril being a prime example	496
303	An example of a borderline case (leaning towards false positive)	497
304	Debugging information for the problematic true pair shown before	497
305	Debugging information (distance differences) for the aforementioned false positive	497
306	The problematic (borderline) false positive after the new alignment scheme gets applied	498
307	A contour around nose tip candidates all of which share the same (maximal) depth value, resulting in uncertainty	499
308	The Performance attained in hard cases where the tip is determined more arbitrarily than in a sophisticated fashion . . .	499
309	The Performance attained in hard cases where the tip is chosen based on the average location of tip candidates	500


310	The result of applying a faster calculation of similarity, as applied to the first person against 90 different pairs from 90 different people	501
311	Performance when smoothing gets disabled, demonstrating little difference compared to prior results	502
312	The sort of results we get by using spectral masks without proper adjustment to make the masks shrewd enough. There is some potential there.	503
313	Example raw slice of the face of one subject	503
314	A test run with just one ring around the nose as a discriminant	504
315	The ROC curve obtained by using one single spectral/diffusion ring	504
316	The ROC curve obtained by using one single spectral/diffusion ring, applied only to the true positive gallery in the set	504
317	The ROC curve obtained by using just one diffusion distance as a discriminant	505
318	The ROC curve obtained by using two diffusion distances as discriminants	505
319	ROC curve for a rather disappointing approach of mask dilation based on diffusion distance	506
320	The same GUI in late March	516

321	Face cropping for standard experiments data	517
322	Difficulties identifying faces in GIP data	517
323	Difference images of the entire face surface before and after ICP-based registration	518
324	Assumed mark (left) extracted from accompanying GIP data (right), illustrating mis-detection	519
325	The offset problem visualised	520
326	The face before (left) and after cropping (right)	521

“The surest way to corrupt a youth is to instruct him to hold in higher esteem those who think alike than those who think differently.”

– *Friedrich Nietzsche.*

1 Overview



IN order to study the accuracy of 3-D face recognition algorithms, one must differentiate between the facial expression as a contributor to variation and the physical component which hardly ever varies. The former can help one learn something about a person at one given point in time, whereas the latter helps distinguish between people. In order to remove variational impact caused primarily by facial expressions, one can assume a commonality between facial expressions and study their statistical nature automatically. If faces have expression-free (or neutral) equivalents, it then becomes abundantly clear how to tell faces apart. A good solution in such a problem-inducing situation would be a framework that can separate the contribution of expression from the contribution that hardly ever varies. Then, the framework also becomes more able and better equipped to annul the former component.

1.1 A PCA Approach

The idea adopted here is one which involves learning variation from a large dataset. Principal component analysis (PCA) gets used for this and along the lines of Cootes *et al.* [20, 22, 21], biologically-meaningful modes of variation are found which encompass a set of faces (albeit in 3-D rather than 2-D). Once the average face is known – as extracted from the data along with the common modes of variation – it is then possible to apply transformations in reverse, reparameterising the model according to the problem domain. If all images can be brought into a common frame of reference, comparison is made trivial, using known similarity measures. It is important to only model facial expressions, however, excepting nose and forehead for instance. Viewing the outcome in terms of recognition rates (with ROC curves for example) enables tweaking and fine-tuning the method.

1.2 Rationale

The merit of the approach adopted here is that it goes beyond 2-D and makes use of the full 3-D data, relying on PCA to handle an otherwise very complicated job². One of the current pitfalls when it comes to 3-D face recognition is the inability to manage a lot of data and exploit its full potential; in 2-D there is also some guesswork associated with uncertainty, caused in part by illumination ambiguities, which only ever allow rough estimation of depth

²Humans are used to recognising faces by texture and crude stereo vision, not full 3-D.

and never a consistent method either, due to changes in light sources and scale factors. 3-D face images suffer from none of these issues, but they are more intrusive in acquisition, not to mention secondary matters relating to expense, storage limitations, and availability.

This document provides some background about a research project which deals with a fully 3-D face-related application. A clear direction has neither been determined nor finalised yet for building upon this work, but suggested improvement might be the utilisation of GMDS. The group of Dr. A. Mian ([home page](#)) has done some fantastic work on 3-D face recognition and we shall attempt to reproduce some results with a NIST-supplied database, then show potential for substitution and possibly an improvement (performance- or detection-wise, where by performance we refer to speed and hardware utilisation).

1.3 Goals

Outline of the work and general thoughts about the goal suggest an aspiration towards GMDS vs. PCA comparisons, whereby shrewd parallelisation³ of the former method may deem it more suitable for many practical needs. At the time of writing we approach this goal by writing code that can handle the data by normalising it somewhat before using the ICP algorithm to process pairs

³In his IJCV paper, Bronstein stresses that “[t]he inner geodesic distances were computed using an efficient parallel version of FMM optimized for the Intel SSE2 architecture (using our implementation, a matrix of distances of size 2500×2500 can be computed in about 1.5 seconds on a PC workstation).”

and build models separately. We also work on several different visualisation methods, cleanup methods, smoothing algorithms, and cropping methods. We are classifying images and expressions as neutral and not neutral by dividing the datasets prior to runtime, rather than determine this 'on the fly'.

1.4 Document Structure

This document is structured as follows; it starts with a brief literature survey explaining some recent work that closely relates to ours. Sections 3, 4, and 5 cover the existing method, data, and implementation which replicates some others'. Experiments appear in the later sections and they provide a look at the suggested scope of research, in addition to the available data, a foreseen methodological approach (the subject of ongoing debate), and some expected results.

It cannot be stressed strongly enough that this text is work in progress and therefore should be dealt with as such.

“We’ve always been shameless about stealing great ideas.”

– *Steve Jobs.*

2 Existing Work/Literature Survey

FIACIAL recognition is a subject of great importance and a lot of literature is already dedicated to it. Recognition of non-rigid surfaces such as faces is a difficult task to tackle both at an inter-personal and intra-personal level, mostly due to variation in one’s facial structure over time. The problem is further complicated by the addition of non-rigid elements, notably the introduction of a wide range of facial expressions, which are controlled by many minuscule muscles and can vary enormously by the combination of these muscles’ state. With the growing interest in access control technology – be it for fraud detection or for something more mundane such as personalisation upon identity detection – competing methods were developed to address a need for robust, expressions-proof, and potentially uncertainty-aware (in the sense that degree of reliability can be reported) method of pairing a given, unseen 3-D scan with an entry in a database of faces (with unique preassigned IDs or equivalents). This document only deals with one family of approaches, namely those that non-rigidly transform 3-D data so as to score

dissimilarity and therefore provide figures of merit to a given match. By trying to match many pairs or assessing their appropriateness for comparison based on a searching index⁴, one can determine a best match.

The following strands of work are most suitable for the proposed new framework and their appropriation will be described herein, where fusion of disparate ideas from each will be viewed as desirable for novelty.

2.1 Generalised Multi-dimensional Scaling

The basic idea is that expressions can be treated by inspecting their effect on the surface of a flattened face. Each expression can then be treated using isometries, which are an area explored by others too [52]. The surface of the face is deformed to a Canonical Form using Multi-Dimensional Scaling (MDS) such that geodesic distances between the points are preserved. This helps remove the impact of expressions on the surface in a different way than the one adopted with PCA. There is an extension to this work, which is known as Generalised Multi-dimensional Scaling (GMDS). Bronstein *et al.* used variants of such a non-rigid method to tackle the face recognition problem, whereas many others stick to rigid methods which preserve the geometry of the faces as they approach the recognition problem. GMDS can also be used in a wide range of other problems, including deformation-

⁴The challenge of searching for matches in large databases is a complex problem in itself. It can use a coarse-to-fine (multi-scale) approach or signatures that act somewhat like hashing.

invariant comparisons, similarity of deformable shapes with partial similarity, and correspondence of deformable shapes. This will be discussed in a later section which deals with implementational considerations.

2.2 Expression Deformation

The baseline for this work is a paper from Mian’s group [3]. In their paper “An Expression Deformation Approach to Non-rigid 3D Face Recognition,” [Faisal R. Al-Osaimi](#), M. Bennamoun, and A. Mian explain some encouraging results from experiments that apply PCA to face images (the paper was also [published online for Open Access](#)). This comprehensive paper from the group in question is 22 pages long in the raw form and about 15 in IJCV. The abstract describes an idea and quantifies some results using known benchmarks and the “FRGC v2.0 dataset”. Then, the method is alluded to vaguely and not formalised until later. “Most of the approaches in the literature are rigid,” says the text in page 2, just before the overview which states: “The main contribution of this paper is a non-rigid 3D face recognition approach. This approach robustly models the expression patterns of the human face and applies the model to morph out facial expressions from a 3D scan of a probe face before matching. Robust expression modeling and subsequent morphing gives our approach a better ability in differentiating between expression deformations and interpersonal disparities. Consequently, more interpersonal disparities are preserved for the matching stage leading to better recognition

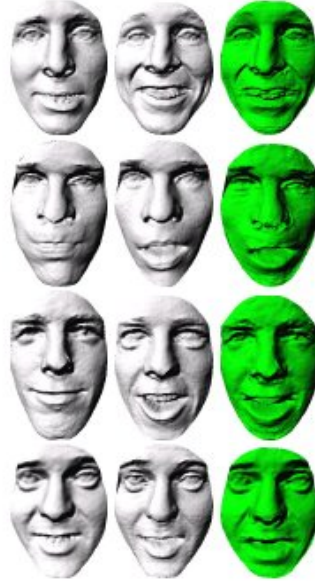


Figure 1: Expression parameterisation in action (image from Al-Osaimi *et al.*)

performance.”

The background section is followed by some classification of existing work, concluding with: “Our approach also falls into this category i.e. non-rigid 3D face recognition.” 1.1 presents a very good summary of related work and 1.2 a clear overview of the method and the ideas behind it, accompanied by a helpful diagram at the bottom of page 3 (Figure 1). The strategy is to use pairs of image of the same individuals, normalising them a bit, and then applying PCA to reduce the dimensionality that characterises expression variation.

Section 2 in page 4 starts by describing pre-processing steps that are essential

yet specific to the limitation of the FRGC v.20 dataset. Page 5 starts presenting some visual examples of the approach, with some equations relating to PCA (along with more visual examples) in pages 6 and 7. The next stage will deal with registering images from particular individuals and then building a model from them. It ought to be stressed that registration gets done with just a partial face (whose most dynamic structures are mostly omitted using a binary mask), whereas modelling brings together the entire face as much of the variability is contained in the previously-occluded or masked parts (as they mostly interfere with registration by introducing ambiguities).

Section 3 begins to deal with some other experiments that are not just dealing with models in synthesis mode. The same dataset is being used (with about 5,000 3-D faces), but more data gets added to it. To quote, “The dataset is composed of two partitions: the training partition (943 scans) and the evaluation partition (4007 scans). [...] The FRGC dataset was augmented by 3006 scans that were acquired using a Minolta vivid scanner in our laboratory.”

Parameters and set sizes (those which are included) get tested in very large-scale experiments that yield ROC curves. These curves help show how to set the different parameters and enable one to measure advantages of one algorithm over another. Page 13 has some comparisons to other methods from the literature, with numbers summarised in a chart.

This truly inspiring work is being partly reimplemented in order for something similar to be achieved in terms of results (see Figure 1 for pose/expression

correction). The reason for these overly specific descriptions of this paper is that it is being used as primary reference work on which to build upon, initially by mimicking an implementation.

2.2.1 Viola-Jones

A later paper from the same group [37] deals with a similar problem except the expressions and it analysis ears rather than faces (frontal). This strand of work demonstrates impressive results, but these are comparable to prior work from other groups, which show similarly good results in the region of 99% and above. The claim being made is that only one image is mis-detected (shown in table), leading to the sub-100% figures. When the gap is reaching such minuscule value, it becomes a discriminant which cannot quite distinguish between those where winning is hinged on one single image. For examples, where there is occlusion by hair, performance drops to about 50%. Robustness of such methods varies based on the assumptions that they make (e.g. expectation of structural completeness).

The method is dependent upon similar algorithms which were used for face detection. The ear does not require a resolution as high though. It is managing to detect ears within about 6 milliseconds and sometimes enabling real-time detection at a frame rate high enough for video sequencing. Performance depends considerably on the size of a given dataset, either because of galley size or the complexity of the set, whose scale affects recognition rates

too (there is down-sampling). Training took days on a cluster of about 30 PCs, so this performance has a hidden toll.

The authors are using templates and rectangular coarser frames, with Haar features and AdaBoost (Freund and Schapire). We already this in the program for the purpose of nose-finding, even though the potential of this is not being explored further at this stage (it does train on and detects faces under different conditions, but it needs transformation to 2-D).

2.2.2 Similar Work

A newer strand of work is described in a [paper from Luuk Spreeuwers \[71\]](#). This work provides a higher standard to aim for, having just claimed a 99.0% identification rate for one of the tests on FRGC v2 data. They also address the issue with ICP, which very much resembles the debate over pairwise vs groupwise non-rigid registration. Given the existing framework, plugging in something which emulates the above paper is a task requiring that we resolve all the same common issues, e.g. input preparation, then PCA. Their nose-finding method (page 8) is of interest to us too because it is more advanced than some of what we currently have ('ViolaJones', 'icp', 'sphere', 'nearest', and 'nearcentre'). Tables 9-11 do not provides assessment of speed and performance in terms of detection for geodesic distances.

2.2.3 Neutral Images as Reference

An assumption which is demonstrated to be of practical use is that by warping images into a neutral (special expression-free) frame of reference all faces are to be reliably told apart. However, as suggested by research in other areas, detection is then biased towards the selection of this one reference, which innately favours a pairwise approach and not a groupwise one. Moreover, there is presumed availability of neutral datasets, which is hard to assure, and there tends to be a difference also between distinct neutral scans of the same subject, depending on the imaged pose for example. These issues, while still separate from the main investigation domain (and thus just a secondary matter), are worth taking into account. There is a vast body of literature, e.g. regarding brain MRI analysis, about reference selection and its ramifications.

2.2.4 Similar Work on Image Sequences

In order to better understand what the reference should be, a closer look at our dataset was seen as necessary; therein, several sets contain neutral and non-neutral image captures (of the same subject/s). In the case of Mian’s group, information is available which was taken from 3 subjects only, containing roughly 1,000 images from each and involving a talking sequence *a la* Hack and Taylor[32]. This is valuable for a training phase, wherein PCA is used for intra-personal variability learning. Whether it generalises to other

subjects or not is another matter altogether.

2.2.5 PCA-based Models

Model construction with the aid of PCA is far from new and it is inspired by other strands of work [40]. In prior work which delved into face modeling Cootes *et al.* found the mean shape and restricted set of PCA axes that provide a concise description of the training set of shapes. This was used to build a 2-D model of shape and this model could also be used to generate new shapes. Let such a new shape be \mathbf{x} , generated from a set of shape parameters \mathbf{b}_s :

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s. \quad (1)$$

The matrix \mathbf{P}_s contains the eigenvectors of the covariance matrix of the training data.

The generated shapes can be constrained to be similar to those seen in the training set by constraining the allowed shape parameters, \mathbf{b}_s , to be similar to those extracted or learned from the training shapes. Typically, the distribution of training set shape parameters is modelled by a multivariate Gaussian pdf, and new shapes are generated by sampling from this pdf. To further enhance this type of models, appearance models were later developed. The appearance models encapsulate textural information about the

variation across this set of shapes. It is possible to extend the former method and construct models that encapsulates not just the variation of the shape of objects in images, but also the variation in the appearance of the object itself. Appearance models were developed by Edwards *et al.* [29]. Their greatest contribution, advantage, and essence lie within the fact that they incorporate textural information rather than shape alone. Texture is made out of grey-level pixel intensities. Incorporation of full colour is possible as well [72]. Colour can be simply thought of as an extension of the single grey-scale band. It can be divided into bands using the most common separability: red, green, and blue components⁵.

A shape model can be thought of as providing very limited information about the appearance of an object within an image, in that it describes the shape of an object, where it is implicitly understood that the shape of an object corresponds to strong edges in an image.

Appearance models describe not only the shape of an object, but the image intensities within the outline of the object as well. In the following subsections, three steps are discussed in turn: modelling shape, modelling intensity, and combined models.

The first step is building of shape models, which use a finite number of modes for representation. Shape models that are built from the outlines of objects in images enable those images to be brought into a state of alignment. We can

⁵There are different possible colour schemes, but they need not have any effect on principles of sampling intensities.

warp the shapes within an image to match the mean shape. By interpolation, we can extend this warp to the entire interior of the object. This means that corresponding parts of shapes in those images will be easy to identify and then use, e.g. in order to sample intensities. To model texture, differences in shapes are removed by morphing each training image to the mean shape⁶. A shape-free texture patch can then be estimated from the image by sampling on a regular grid and forming a vector \mathbf{g} . Statistical analysis proceeds as for shapes and it results in the following linear expression for texture

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g. \quad (2)$$

\mathbf{g} is the intensity vector. \mathbf{P}_g contains the eigenvectors of the covariance matrix of the training data and \mathbf{b}_g controls the intensity. An example appearance model is shown in Figure 2. The process is hardly different from dimensionality reduction in the case of shape. The models in Equation 1 and Equation 2 have a linear form, so they are quite compact. This is a highly desirable property which makes the models flexible and manageable.

However, at the moment, the two components of the model, namely the shape \mathbf{x} and the shape-free texture \mathbf{g} , are independent. In real images, shape and texture are not necessarily independent. One simple example to think of is an image of an individual's face. When the person changes expression, the

⁶Warps can be applied using a strategy borrowed from graphics. In all experiments described in this thesis this was achieved by a triangulated mesh generated from the landmark points and barycentric coordinates to mesh the intersections put in vector \mathbf{g} .

shape of the face changes. But the texture (i.e. positions of highlights and shadows) obviously changes too, in a way that is correlated with the shape change. Hence it is desirable to merge the shape and texture models, so as to obtain a new model that is aware of both types of variation. This combined model can then also incorporate any correlations between shape and texture.

The parameters \mathbf{b}_s and \mathbf{b}_g are aggregated to form a single column vector

$$\begin{Bmatrix} \mathbf{b}_s \\ \mathbf{b}_g \end{Bmatrix}. \quad (3)$$

The new vector is a simple concatenation of the two. However, since the values of intensity and shape can be very different in magnitude, weighting is needed. Such weighting brings equilibrium, under which both shape and intensity maintain a sufficiently-noticeable effect and impact on the model they jointly build. A weighing matrix resolves the problem introduced here and it is, by convention, named \mathbf{W}_s ⁷. With weighing in place, aggregation takes the form

$$\begin{Bmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{Bmatrix} \quad (4)$$

⁷The letter *s* stands for *shape*, as by default this matrix scales the shape parameters only. It gives logically equivalent results to these of applying the factor $\mathbf{W}_g = \frac{1}{\mathbf{W}_s}$ to intensities.

where \mathbf{W}_s is set to minimise inconsistencies due to scale. By applying another PCA step to the aggregated data, the following combined model is obtained

$$\begin{aligned}\mathbf{x}_i &= \bar{\mathbf{x}} + \mathbf{Q}_s \mathbf{c}_i \\ \mathbf{g}_i &= \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c}_i\end{aligned}\tag{5}$$

The appearance (shape and texture) is now purely controlled by the new set of parameters, \mathbf{c} . There is no need to choose values for two ‘families’ of distinct parameters. This combined model reaps the benefits of the dimensionality reduction performed, which is based on shape as well appearance. This means that this new model encompasses all the variation learned and the correlation between these two distinct components. Since PCA was applied, the number n of parameters \mathbf{c}_i is expected to be smaller than the number of parameters in \mathbf{b}_s and \mathbf{b}_g put together.

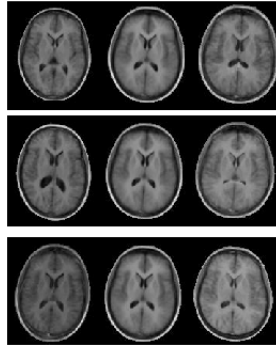


Figure 2: The effect of varying the first (top row), second, and third parameter of a brain appearance model by ± 2.5 standard deviations


The work we shall deal with will not require texture or even a complex model such as the above. Prior work, however, necessitates understanding of possi-

ble future extensions.

“Developments in medical technology have long been confined to procedural or pharmaceutical advances, while neglecting a most basic and essential component of medicine: patient information management.”

– *John Doolittle.*

3 Methods



AT the core of our methodology ought to exist a mechanism which removes expressions-imposed anatomical variance, as well as complexity such as expression representation. We address these issues with aim of acquiring serialised models. When models are built it ought to be possible to use discrepancy as a measure of dissimilarity and test systematically on a large set, then plug everything into a G-PCA/GMDS algorithm and provide direct comparisons. Here are some key components that make up the method.

3.1 Model-building

As described at the end of the previous section, a parameterised model can always be built given a set of data in a state of relative alignment, which

means that commonalities can be found and encoded in a lower-dimensional form. The key step which involves data alignment has tremendous impact on the usefulness of the built model⁸.

3.2 Data Alignment

Rigid or affine alignment of 3-D images is not a simple task, especially when those images are taken from different individuals with various different expressions where an *anatomical* correspondence – let alone a perceived one – may not exist. The method we adopt for our experiments is called iterative closest point (ICP), which will be further discussed later.

3.3 PCA

Work similar to ours predates or concurs with what we achieved by the middle of 2011.

In 2006, Russ et al. [65] produced a paper on using PCA for face recognition in 3-D, using FRGCv1 and FRGCv2. They tackled the issue of face alignment which is required for adequate sampling of signal for PCA. To quote the paper, they “achieve correspondence of facial points by registering a 3D face

⁸There is finally code in place for plotting the **Pareto distribution** of the built model, which can be used to show how much of the variation each mode accounts for and how this ratio degrades. The problem is, without resolving difficulties of fully automatic face cropping and then applying that to a large set, the data will just be noisy and the modes rather uninteresting. For small dataset (run for testing purposes) there are hardly any modes at all, in lieu with the size of the sampled set fed into PCA.

to a scaled generic 3D reference face and subsequently perform a surface normal search algorithm.”

A later paper, one from Mena-chalco *at al.* in Brazil [53], demonstrates early work that is carried out on a single subject and many acquisitions (much like the GIP dataset for expressions). The texture is incorporated too, building a model using PCA with a very small training set. This work is very different from what we do and their data sets are their own.

Using a variant of classic PCA, $(2D)^2$ PCA.A or 2DPCA, Gervei *at al.* [30] showed recognition rates of 83.3%. This deals with facial expressions too and is similar to what we already have implemented, including the pre-processing and the sampling phases. Even their charts are similar to ours (block diagrams), not just the recognition rates. They use the Gavab 3D face database which contains 540 3-D images from 60 individuals. For the sake of comparison, our June experiments use similar numbers to train the model, whereas UWA uses data of the orders of magnitude of thousands (mostly collected from 3 individuals at their lab).

3.3.1 Robust Generalised PCA

GMDS matrices should be possible to obtain and perform analysis on, but a more robust PCA process may be needed. The way everything is presently structured, using a hybrid of measures will be trivial, however the way data gets organised may matter (concatenation is just one option, probably an

inferior one to using a combined model which studies the correlation within a pair of models). In order to improve future results, a Robust Generalised PCA implementation will be attempted, where one of the early experiments will show the difference between classification rates when classic PCA is used, compared for instance to a process comprising a GPCA-Voting algorithm with Robust Covariance Estimator, Sample Influence Function, and Theoretical Influence Function. We can test different combinations of these once everything is implemented and properly tested. The proposed framework may not only accommodate experiments around GMDS but also more comprehensive benchmarks providing insight into potential steps of refinement. The improved PCA component is being implemented and GMDS too will get merged in. The concept is tested and shown in Figure 3.

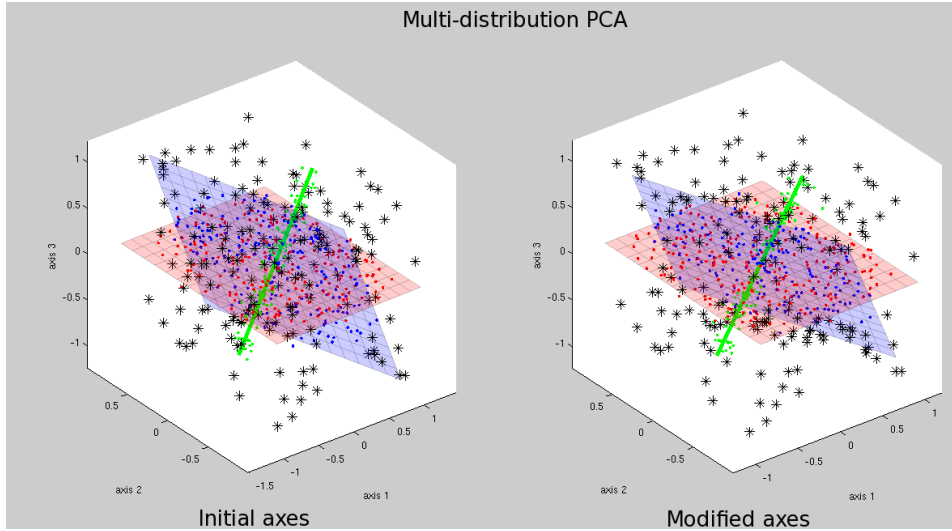


Figure 3: The analysis performed by a Robust Generalised PCA algorithm

Progress on GMDS with RGPCA will be described later.

Outlier detection would probably be needed in cases where samples deviate too much from known distributions and thus should be excluded or rejected, at least in model-building. It is not entirely clear, however, what applicability the division into groups of distributions (assuming no banana-shaped distributability) the generalisation will bring about, unless for example the task is to identify and also classify different shapes (e.g. dogs and horses) at the same time, classifying them both by type and by identify, i.e. inter- and intra-dissimilarity, respectively (even separating classification of people and expressions they have). It might this be the intended application, as otherwise we over-complicate everything. However, It could accelerate the way by which we do recognition (rather than identification).

3.4 PCA for Animation

There was a company in incubation called Genemation. It strove to achieve something similar with PCA models of faces. It is generally possible, given consistent markup (which GMDS can find and optimise over automatically), to construct reasonably realistic models with modes that each respects and represents an observed expression. The commercial value of such an application was never entirely clear and some models were shown where PCA also sampled the texture (as in AAM) to make ethnic modes of variation of even gender-related model modes. For film studios this whole process can save

a lot of work (rendering upon changes to modes) and this was one of the target markets for Genemation, along with game makers. Disney works on something similar these days [74].

3.5 GMDS vs. PCA

The modeling part of the method is loosely defined for a good reason. We wish to compare different paradigms for attaining a model. We can conceive a robust-PCA that would lead to sparse representation and eventually GMDS. It is currently being made as efficient as possible (with means of hardware acceleration too) and there will hopefully be mature GMDS-based software tools to work with quite soon, with prospects of replacing PCA with something less generic and better optimised for purpose.

PCA and GMDS have some clear similarities at some lower level. In MDS and GMDS we treat shapes in a metric space and assume that shape similarity can be reliably measured in terms of the distance between metric spaces. Dimensions correspond to variation along geodesic lines of significance, which in a sense is similar to a landmark point (coordinate) or a texture that can be handled with dimensionality reduction in PCA with all its variants/enhancements. The geodesic measure is more expressive and suitable for surfaces and some of the associated methods that require classification. The invariance tolerance that an isometry-preserving warp/deformation inflicts upon a shape is similar to the restraint in bending of, e.g. anatomy,

as judged by PCA automatically. The topology problem can exist also in PCA, especially if the sampling of observations gets automated. There are factors to be considered such as the nature of the metric which determines the sort of transformations to which topology is not an issue. It also generally depends on the shapes dealt with. Some may need additional information to resolve ambiguities. GMDS problems can be solved by optimisation to help find minimal distances for a group of different metric/dimensions, just as PCA identifies dimensions along which data can be flattened and later reconstructed, through the eigenvectors that correspond to those lines/planes. Shape descriptors are defined such that they can reconstruct a whole class of shapes that are canonical forms and if they are defined at every point on the shape, then they are too dense to provide something specific (just too generic). It is therefore necessary, just like with landmark points for PCA sampling, to pick anatomically or statistically meaningful parts, perhaps assigning weights to them instead in order to limit their influence on the model (like some PCA variants do, as does GMDS). For the measures to be insensitive to topology errors, all sorts of ad hoc methods can be used, but some require understanding the general shape or have a broader class of that shape to compare to and learn from.

The duality of this problem can be expressed visually as well as with equations, to an extent. It would be useful if we produced figures, which are probably better off prepared for use later on (explanatory purposes). It would be beneficial to do so as we envision sequence of events as PCA/singular value

decomposition (SVD) generic classification followed by GMDS robust PCA for refinement. To conceptualise this description accurately, it might look like Figure 4.

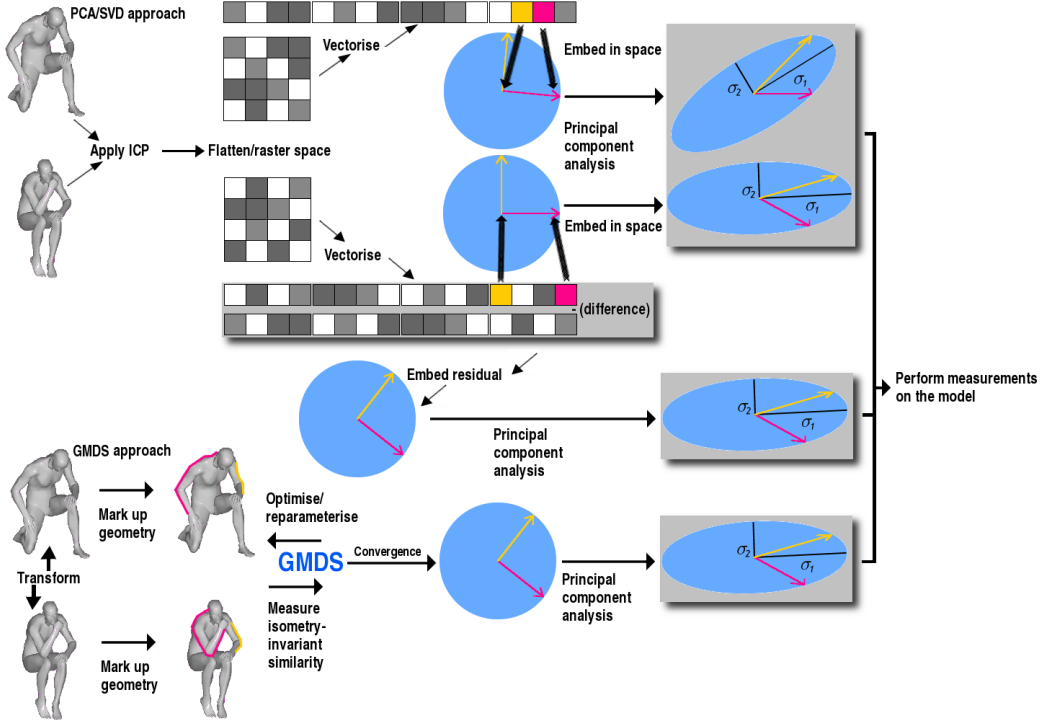


Figure 4: The proposed framework for GMDS improvement

It is possible to treat it somewhat differently, e.g. use ICP for rough alignment, GMDS for intrinsic fine alignment. If the alignment is onto a generalised face, then spectral decomposition takes place in the refined space – an Eigen functions of the generalised face. We need to refine those issues, as shown in Figure 5. Next up, we shall prepare a detailed explanation.

ICP could be considered Gromov-Hausdorff (G-H) in Euclidean space. The

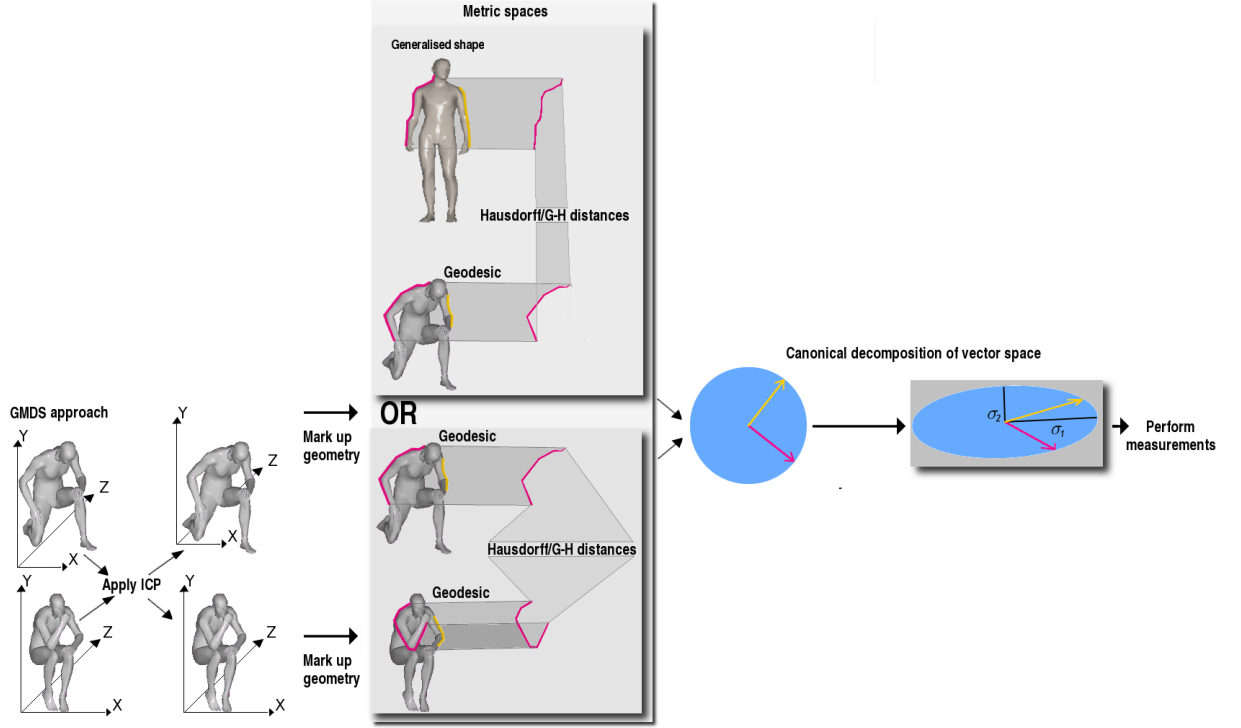


Figure 5: A closer look at the GMDS approach

G-H-inspired method strives to identify and then calculate minimal distances for a group of geometric points with commonality in a more rigid space, wherein harmonic variation occurs in inherently non-orthogonal spaces. One way to model this type of variation and then explain its nature would be high-dimensional decomposition, which evidently requires that data be represented in a high-dimensional form such as vector of coordinates, intensities, energy, or discrete/quantised G-H distances (geometric terms). Figure 6 provides an example of that subtle point.

Depending on the circumstance, different measurable attributes can be added

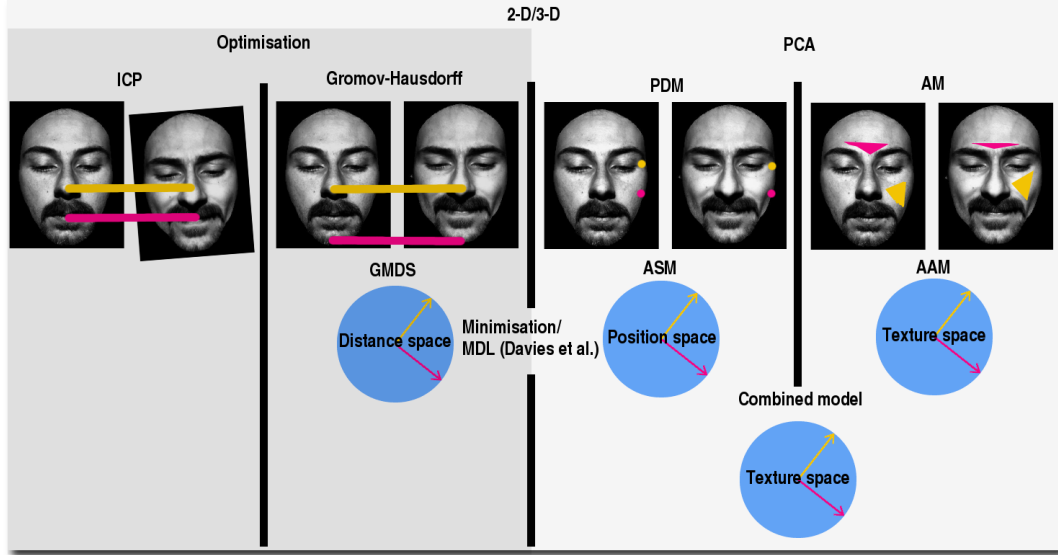


Figure 6: Crude visual example of how typical PCA and GMDS relate to one another, approach-wise

to the space, even a hybrid of them (e.g. shape and texture, so as to reconstruct/recover the relationship between image intensity and the image shape in 2- and 3-D). For synthesis of images belonging to a particular class/subspace, e.g. a canonical form (bar embedding error), one requires that the model should be specific and generic. Specific – for the fact that it need preferably not be confused with similar images belonging to another class, and generic – for the fact that it must span a sufficiently large cloud in hyperspace in order to capture the variation of all images of the same class.

As opposed to models that build upon a texture of pixels⁹, the GMDS ap-

⁹The common computer vision approach of locating and segmenting image parts, e.g. for partial similarity on a per-part basis with scoring, with or without weighting based on statistical/topological/irregular/anatomical significance – that which can more uniquely

proach largely discards the notion of geometry in the way the human eye perceives it; just as PCA/SVD facilitate the modeling of heat, quantifiable recipe ingredients, strings of letters and just about any parameterisable and measurable element, it should be possible to encode or at least translate a given image into a set of properties of some significant role; this is particularly attractive in 3-D, where the size of the given set can be vast – far greater than the actual entropy of the set. Considerable reduction in size can be achieved by considering distances between corresponding points or geodesic distances between neighbouring points, whereupon the image can be reconstructed by merely plugging in the modeled parameters and scaling accordingly. In that respect, it is a pseudo-dimensionality reduction problem. The elasticity of observed objects is implicitly modeled by a collection of metric measurements – measurements taken not in Euclidean space but in a more inherent space, more robust to the external viewer (judging an object from the outside and not relative to its neighbourhood (e.g. landmark points in its vicinity)). Intrinsic similarities are also more resistant to error due to some topological changes in the sense that, assuming there is awareness of the topological changes typically introduced (e.g. hand touches leg), it should be possible to define 'sanity' ranges within which the distances do make sense or alternatively use diffusion distance, or intrinsic symmetry tests for something like animals where preservation of this property is expected.

identify in image within a group or even externally, as belonging to one group of images and not other groups.

Euclidean distances are hardly enough for determining if two points/objects are just close to one another or actually connected.

Isometric embedding is somewhat analogous to putting an image in a reference frame from which to consistently sample parameters. As all comparisons are better off done in a spatially-neutral reference (such as a sphere onto which a more complex image is mapped/flattered), this method seems to follow the same intuition as Davies *et al.* who find parameters to model shape by (in 2- and 3-D) by mapping everything onto a sphere and then applying kernel functions to move those around consistently, for PCA to automatically use the “best” points that encode a complex shape. When dealing with modeling in this context it is typically used for segmentation, non-rigid registration, and synthesis. The problem of recognition and automated classification (e.g. telling apart extrinsic and intrinsic differences) hardly arises in this area.

In GMDS, numerical analysis and multidimensional scaling can be thought of as an iterative, concurrent search for directions and axes that better distinguish between pairs (or groups) of shapes; the general optimiser optimises over image parts and also over corresponding points, which relate to the former to an extent. When dealing with images in a metric space, the situation is merely identical or at least analogous to how image registration problems are solved, by embedment in a high-dimensional space, where the difference between them is estimated as per the vector in the manifold (Euclidean, shuffle distance, etc.), depending on the problem domain.

It ought to be possible to make the problems at hand complementary and not just mere surrogates.

When it comes to sampling geodesic distances for PCA, **furthest point sampling** would do the job of meaningful sampling[55]. It is 2-optimal in sense of sampling. $d_{GH(S,Q)}$ where distances are Euclidean is like ICP. Roughly speaking, PCA is like SVD, being a classical MDSearch dimension, i.e. moment by value. Thus, PCA over the geodesics matrix (meaning MDS) could serve as initialization for the GMDS. We need to use Euclidean distance as one feature and geodesic distance as another. We do need a geodesic distance calculator and the same goes for GMDS.

For cases of mis-identification we may need robust PCA variants that leave out widely different data and improve the overall model. The idea is, the distances being measured will better encompass data associated with the surface of the skin, not just mere points in a locality. Areas that are not typically changing much will be associated with low distances variation (intrinsic to the subject), whereas those subjected to expression variation will not only be expected to have high variation but the type of variation too, as measured in terms of distances, will be possible to measure and use.

3.6 GPCA

In the 2004 paper introducing GPCA [82], page 11 shows and explains an experiment reminiscent of Bronstein & Kimmel work (which was misrepre-

sented somewhat by UWA, based on a toy experiment from an IEEE journal in 2006). It ought to be possible to implement a GPCA approach alongside the existing framework, i.e. without interfering with existing methods and thus facilitating performance comparisons.

Ah...agreed. I will try to get multidimensional scaling results within days.

3.7 Algorithm

The section on implementation provides details which are implementation-specific rather than generalised to a method (still being actively pursued but not finalised). The graphical user interface is also explained therein.

The graphical user interface required an investment of time which will ease later operation and make anyone with interest in the tools more oblivious to the underlying code. The code is annotated and divided sensibly nonetheless.

The following Candès *et al.* text, which comes from a technical paper (not peer reviewed, just an informal 40-page manuscript about ongoing work from the end of 2009, December 2009 to be precise), relates to a problem which degrades the quality of our current results due to ICP failures. In one of the latest experiments, for example, decomposition was very obviously broken and it could easily be seen when the graphs detailing the model were shown, after hours of processing. Plots and ROC curves were produced nonetheless, actually showing that the major error was not too fatal for performance but noticeably problematic still. Any images that move out of line can domi-

nate the signal and therefore spoil the model, which otherwise can give a recognition rate higher than 90% (for the whole NIST database, assuming further tweaking). Technical Report No. 2009-13 addresses the practical application relating to faces, namely lighting imbalance mitigation by low-rank approximation, L.

On Robust PCA: "PCA is arguably the most widely used statistical tool for data analysis and dimensionality reduction today. However, its brittleness with respect to grossly corrupted observations often puts its validity in jeopardy – a single grossly corrupted entry in M could render the estimated L arbitrarily far from the true L_0 . Unfortunately, gross errors are now ubiquitous in modern applications such as image processing, web data analysis, and bioinformatics, where some measurements may be arbitrarily corrupted (due to occlusions, malicious tampering, or sensor failures) or simply irrelevant to the low-dimensional structure we seek to identify."

Addressing 2-D face recognition: "It is well known that images of a convex, Lambertian surface under varying illuminations span a low-dimensional subspace. This fact has been a main reason why low-dimensional models are mostly effective for imagery data. In particular, images of a human's face can be well-approximated by a low-dimensional subspace. Being able to correctly retrieve this subspace is crucial in many applications such as face recognition and alignment. However, realistic face images often suffer from self-shadowing, specularities, or saturations in brightness, which make this a difficult task and subsequently compromise the recognition performance."

Surely it does refer to work that can be generalised to 3-D, but pursuit for literature on that subject carries on. If this was not done before, perhaps it is worth exploring, demonstrating in particular an advantage over a simpler approach it's derived from (SVD) – something we already have implemented anyway, and can refine further to a satisfactory level.

3.7.1 Related Work

From half a decade ago comes this long talk about [Generalized Principal Component Analysis \(GPCA\)](#).

Further literature review shows work on GPCA in the context of face recognition. There is very limited amount of work on the subject which applies to 3-D sets however. In their 2005 paper, Kong *et al.* [43](Singaporean group) looked at the application of GPCA to 2-D face analysis, having sought to overcome the curse of dimensional while not accepting too much noise in their training set. “In this work,” they explain, a “[k]ernel-based 2DPCA (K2DPCA) scheme is developed and the relationship between K2DPCA and KPCA (Scholkopf *et al.*, 1998) is explored. Experimental results in face image representation and recognition show the excellent performance of G2DPCA.” The acronym G2DPCA stands for Generalized 2D Principal Component Analysis (G2DPCA) and the “K” stands for kernel. There are comparative graphs there, supposedly showing the different performance of the different ‘families’ of PCA-based algorithms as tested on various datasets,

e.g. ORL, UMIST, and Yale databases (experimental results correspond only to the former in some of the cases, but all three datasets are used for experimental purposes). Quoting the short conclusions section: “A framework of Generalized 2D Principal Component Analysis is proposed to extend the original 2DPCA in three ways: firstly, the essence of 2DPCA is clarified. Secondly, a bilateral 2DPCA scheme is introduced to remove the necessity of more coefficients in representing an image in 2DPCA than in PCA. Thirdly, a kernel-based 2DPCA scheme is introduced to remedy the shortage of 2DPCA in exploring the higher-order statistics among the rows/columns of the input data.”

In a later paper, this one from Wu *et al.* in 2006 [85], Generalised PCA is applied to virtual faces to achieve face recognition in 2-D. The Yale Database is used for experiments and a recognition rate of 81.5% is reported, compared to 59% for PCA on its own and 68.5% for Fisher faces.

3-D Application

After hours of searching and researching, we were unable to find any work which utilises GPCA for interpretation or separation between 3-D datasets, e.g. finding trajectories that distinguish between individuals, bar facial expressions variation. Search was not restricted to just face recognition, either.

One possibility we have is to implement GPCA, demonstrating its advantage over a purely PCA approach (which can be further refined implementation-wise). Later on, a GMDS-based measure can be introduced or embedded

into the framework, hopefully showing in an empirical fashion what would be considered an analytical correlation.

In GPCA, vectors perpendicular to points that represent lines or other elongated distributions (whose principal axis at the given dimension defines this line's direction) are used to determine the separability between an unknown number of different clouds, e.g. a set of shapes belonging to a common person/expression. In GMDS, the relationship being exploited is that of distances between analogous points in surfaces. If each sample was to be incorporated and formally defined by the assemblage of aggregated distances, for example (alas, ordered meaningfully), then the dimensionality is defined consistently such that each dimension in hyperspace corresponds to an innately meaningful distance in 3-D space. Since the points move in harmony on a face, the real dimensionality (not that reached and formed by concatenation) is actually a lot lower. PCA allows us to automatically find an analogous set of axes that capture the variation and decompose this effectively, sorting everything in an ascending order. For instance, we expect to find that when the mouth is opened there is a particular expansion in several dimensions and if the training set exhibits this relationship, then the description length of the model (a la MDL) will remain small and similarity therefore accordingly high, which is exactly what we want as it may be the same person and consistent with a purely expression-imposed difference. The main relevance of the generality of PCA (the G in GPCA) is that it facilitates capturing and then clustering groups of faces matching similar criteria already inherent in

the model, thus occupying less space. For that to work as a similarity measure we can adhere to calculation of the product of the eigenvalues, which is a fast approach already proven and tested. The true importance of geodesic distances in this context (and work with isometrics) is that they provide an anatomically meaningful set of measures on a given surface, even a partial one (which presents a challenge to more primitive sampling on a grid if there is limitation on dimensionality. We cannot sample every single point as an observation as it makes vast matrices of 480x640 dimensions. Only a subset of that is truly essential and the more compact the signal, the better.

3.7.2 Wavelets (Texas University)

An interesting strand of work comes from text which summarises key work from Texas University – work which explores the use of wavelets and an approach which can be found in "New Approaches to Automatic 3-D and 2-D+3-D Face," a thesis at [their repository](#). The work uses an interesting dataset, and in the case of the paper this seems limited to just a recently-acquired set matching particular requirements/protocol. From one description it emerges that "T3FRD is the largest free and publicly available database of co-registered 2-D and 3-D face images that is suitable for separate evaluation of the recognition task. For its construction, 1196 pairs of high resolution range and colored portrait images were captured from 116 adult subjects at the former Advanced Digital Imaging Research (ADIR) LLC (Friendswood, TX) using a MU-2 stereo imaging system made by 3Q Technologies Ltd.

(Atlanta, GA), under contract to NIST."

The work seems novel enough, as it combines the data embedded and encoded in the form of texture and also the underlying geometry. The core or the 'engine' is quite simple to grasp since it mostly relies on compressibility of a finite set of features, or rather their mutual entropy. It is similar to work I did with NRR, where a bunch of wavelets are used to assess misregistration through inherent correlation, using compressibility as a measure of similarity, or entropy estimation (somewhat related to mutual information). Quite a few people tried this before with limited levels of success. It is a 'cheap' way to get similarity working, with libraries that are used by JPEG for example.

The explanations are well organised and the opening parts of the thesis are clear. They provide a good, broad overview of existing methodologies. The latter sections show the authors using their own data, matching the requirements from NIST, preprocessed and aligned using ICP. This also corrects pose and scale, which makes the T3FRD database dependent upon other methods. Although it is smaller than the FRGC database, it is said to be a lot simpler to deal with (for instance, the poses in there are more limited). The texture gets exploited as well and this too gets refined in order to remove effects of noise and annual comparative analyses where rigid or affine stages act as distinguishers at the expense of the latter stages, which are typically the more interesting ones and those that are more actively researched.

The author is looking at rigid areas of the face, so recognition rates are

understandably high. Comparisons, performance-wise, should preferably involve the T3FRD as the only data to work on. The results reported when applied to the older-but-commonly-used Face Recognition Grand Challenge (FRGC) should be treated and referred to an altogether separate benchmark with different levels of difficulty (increasing based on the semester in which the images got acquired). This perhaps limits the number of methods actually compared in this paper, as not so many studies were done with the same database. An important point that gets raised is that "researchers can make fair comparisons between competing algorithms based on recognition capability only without biases introduced by pre-processing," which makes T3FRD "an attractive alternative to the older FRGC database." It does, however, limit the scope of benchmarks based on the literature and there is a short description of the hole-filling and median filter used (with 3x3 kernel). A lot of space is being dedicated to justifying the use of T3FRD rather than FRGC, even though it does not have many subjects in it, based on the Results section (116 subjects enrolled for the experiments, only 18 used for training)). This can make the recognition problem simpler. Gupta *et al.* published similar results in IJCV last year, proving or at least validating to a degree the quality of these results. Perhaps the text can be made more concise by tightening the description of the dataset which makes it seem slightly promotional at times and not exceedingly relevant to the methods presented by the paper.

Addressing the methods at hand, there is merit in adopted approach, which

takes wavelets (barycentric in this case) and multi-scale (coarse-to-fine) ideas to get locations, extract wavelet coefficients, and sometimes use PCA to then model the variation as characterised by those succinct, localised descriptors. It actually pools together three recognition sub-systems that are aggregated with LDA to perform a decision on several levels. For landmark detection, Gabor-based methods are utilised, identifying eye corners and the nose tip. A region ensemble where the faces are treated as a set of regions would work well provided the division into regions is accurate. The multitude of regions provides extra robustness, in case one discriminant is overly weak or misleading.

The use of Gabor jets is not so new, but it does provide compelling starting points that are fiducials (1 millimeter away from manual landmark for the nose tip and about two for eye corners). The descriptions in the remainder of the text are rich in words and light on equations which would otherwise help formalise the process or methodology, before leaping to results. It is not entirely clear by this stage (until the later block diagram and composition of methods) what the novelty of this work is. A figure this showing a this diagram helps visualise the proposed framework and had it been shown earlier it would have helped foresee the structural aspects of the problem domain. This is what several IJCV papers in this area sometimes do. The separation between methods and experiments is not clear because the results section keeps introducing new methods or variants of these basic, pertinent methods. ROC and CMC curves show impressive performance, but one must bear in

mind the quality of the data, the set size, and multi-modality of sorts (not just range images). Some of the compared-to studies (including one from Le Zou *et al.*) are from the same institute and Mahoor *et al.* from Miami University is perhaps a better one to compare to, although there may be the bias due to using another university's database. McKeon *et al.* (Robert McKeon and Trina Russ of Digital Signal) train on the FRGC database and get competitive results, very much comparable to those from the group in Texas University.

Overall, performance shown in this work is high, the degree of difficulty is hard to assess as not many other studies were done with this dataset (not outside Texas University), and there is novelty in the way measures are combined to attain a powerful discriminant, rooted primarily in 3 limited regions of the face (rigidly registered to begin with). The work would benefit from having a set of results from experiments applied to the FRGC dataset, in order to demonstrate performance from another frame of reference.

Addressing more specific points, while the paper has real importance, there are concerns and it also comes with caveats because of the lack of adequate benchmark comprehensive enough. The work contains original work. Somewhat dated, a bit dependent on similar work, but original nonetheless. Notable absence of more tests definitely stands out. A lot of space is dedicated to advocating the use of this data, whereas a more useful thing to do would be to explain the methods. The abstract does highlight the limitation. Methods could be described better, especially due to the order of presentation. It

is mostly acceptable, but there is room for improvement. The grammatical quality of the text is high with few exceptions like typographical errors, many formatting inconsistencies in the references (e.g. page numbers, commas, et cetera), and some issues that require ironing out. Among drawbacks of some of the covered methods is that they are also slowed down/performance degraded non-linearly. Nothing is being said about performance in the paper, e.g. time taken to run experiments. Normalisation is well defended in the text which explains composition of regions. Related to this there is work in the International Journal of Computer Vision [31] ([online version](#)). It is prior work on geodesic distances for recognition, courtesy of S. Gupta, M. Markey, and A. Bovik," Referring specifically to their use of geodesic distance, they write: "Lastly, we develop a completely automatic face recognition algorithm that employs facial 3D Euclidean and geodesic distances between these 10 automatically located anthropometric facial fiducial points and a linear discriminant classifier. [...] We develop a successful 3D face recognition algorithm that employs Euclidean and geodesic facial anthropometric distance features and a linear discriminant analysis (LDA) classifier. [...] As features for our proposed Anthroface 3D algorithm, we employed 300 3D Euclidean distances and 300 geodesic distances between all of the possible pairs [...] We computed geodesic distances along the facial surface using Dijkstra's shortest path algorithm (Dijkstra 1959; Tenenbaum *et al.* 2000). Besides 3D Euclidean distances, the motivation for employing geodesic distances was that previous studies have shown that geodesic dis-

tances are better at representing ‘free-form’ 3D objects than 3D Euclidean distances (Hamza and Krim 2006). Furthermore, a recent study suggested that changes in facial expressions (except for when the mouth is open) may be modeled as isometric deformations of the facial surface (Bronstein *et al.* 2005). When a surface is deformed isometrically, intrinsic properties of the surface, including Gaussian and mean curvature and geodesic distances, are preserved (Do Carmo 1976). Hence, algorithms based on geodesic distances are likely to be robust to changes in facial expressions. From among the 300 Euclidean and 300 geodesic distances, we selected subsets of the most discriminatory distance features, using the stepwise linear discriminant analysis [...] Using this procedure we identified the 106 and 117 most discriminatory Euclidean and geodesic distance features from among the 300 Euclidean and 300 geodesic distances, respectively. We pooled these 106 Euclidean and 117 geodesic anthropometric distances together, and using a second stage stepwise linear discriminant analysis procedure, we identified the final combined set of 123 most discriminatory anthropometric facial distance features."

Figure 2 shows an example of a simplistic implementation we have implemented, as will be shown later.

This is rationalised by claiming that "the Anthroface 3D recognition algorithm (Sect. 3.3), with Euclidean and geodesic distances between 25 arbitrary facial points (Fig. 2) instead of the 25 anthropometric fiducial points (Fig. 1). These points were located in the form of a 5×5 rectangular grid positioned over the primary facial features of each face (Fig. 2). We chose these

particular facial points as they measure distances between the significant facial landmarks, including the eyes, nose and the mouth regions, without requiring localization of specific fiducial points. A similar set of facial points was also employed in a previous 3D face recognition algorithm for aligning 3D facial surfaces using the ICP algorithm (Lu *et al.* 2006)."

"For all faces in the test data set, the 123 most discriminatory anthropometric Euclidean and geodesic distance features x were first computed. They were projected onto the 11D LDA space as $y = Wx$ that was learned using the training data set."

They are using the T3FRD database, which makes things a lot simpler due to the aforementioned factors.

3.8 Outline/Thoughts About Operation

Remaining tasks that will be taken care of over time are hard to name very specifically because they depend heavily on progress and results which this progress brings about. The plan going forward is to complete (by coding) the missing pieces of our conceptual framework at an appropriate capacity, put all the datasets (pre-processed) in a suitable frame of reference using the ICP2 implementation, then feed the data from the images into PCA and start building models for feasibility tests. Then, data may need to be classified based on criteria such as neutral and non-neutral, in order to achieve some sort of separability. For the time being, this division is pre-supplied, which

simplifies everything.

Once the pre-processing step works reasonably well with many arbitrary images, running the algorithm on thousands of images would be worthwhile, with all images then stored offline (saved to disk) for quicker experiments to be performed on them later. We only need to open images and make the experiments a dual-phase process (data preparation separated from modeling).

With an existing implementation of model-building, dealing with large sets should be possible, albeit it can consume a lot of computer resources. It is desirable to plan very carefully what sets of experiments are wanted here. Should we reproduce the experiments from Mian's group or branch out to exploring different aspects of the problem, perhaps building a hybrid of algorithms by fusing in some homebrew code that makes use of what previously worked well and therefore yields unique work that has a more local 'flavour' rather than a reproduction of what's presented by IJCV? Novelty is required for papers, but a plan needs to be outlined along with an hypothesis. By now, some of these questions have been answered and will be further explored later. These constitute a documentation of early discussions.

In terms of timeframe, things have progressed reasonably well so far, despite major limitations in terms of resources (overly occupied server cores), no fixed computer in lab, no local access to MATLAB at home, except GNU Octave. With all sorts of accounts-related issues that consumed a lot of time and with all the data now in place, things should progress more smoothly

from now on. The literature is also well understood and images with known properties are in place (all ~ 100 gigabytes of them). This is beyond the scope of this document though. Some areas already addressed or still being addressed are:

- ▷ automatic classification of neutral or not neutral
- ▷ testing efficiently for spikes and holes handling (see Figure 9)
- ▷ convert to centimeters and make code resistant to scale changes by making it more adaptable to given measurements
- ▷ replicate the results of prior work, if possible (may require over-occupation with work that has already been done)
- ▷ acquiring sufficient computing resource (Amazon, Google, local servers, clusters, etc.) for larger experiments to come

3.9 Systematic Experiments

Comparing a PCA approach to a GMDS approach was the original goal of our work, primarily utilising statistics, e.g. 3-D facial expression interpretation through statistical analysis. With the goal of validating and comparing face recognition methods, we can embark on the following path of exploration. The data to be used needs to be of different individuals and the datasets must be large enough to enable model-building tasks. As such, the

data specified in Experiment 3 of FRGC 2.0 should be used for both training and testing. It needs to be manually classified, however, as groups that previously did this have not shared such metadata. It would be handy to select hundreds of those that represent expressions like a smile and then put them in respective loader files, alongside data with an accompanying neutral (no expression) image. It ought to be possible to set aside 200 such pairs, all coming from different people. Identification in such a set ought to be quite challenging, without texture (which is in principle available in separate PPM files).

The experiments can have the set of 200 pairs further split into smaller groups for repetition that takes statistics into account and can yield error bars. Dividing into 5 groups of 40 pairs is one possibility, even though a set of 40 individuals is becoming a tad small. In order to train a model of expressions it ought to be possible to just use the full set.

When approaching this problem the goal would be to pair a person with an expression to the same person without the expression (or vice versa), attaining some sort of gauge of expression-resistant recognition. The gallery is the set of all faces in the set. Similarity measures being pitted for comparison here can include the 4 ICP methods we have, plus variants of these and different selection of parameters. Different measures resulting from ICP and the region being compared (e.g. all face versus nose, versus forehead and nose) are another area of exploration. There ought to be separation between the idea of cropping for alignment alone and cropping or binary masks for

the sake of computing difference as well.

What we may find is, by cropping out some parts of the face recognition will improve considerably. But in order to take the deformable parts that change due to expression into account, something like an expression becomes necessary. Then, there is room for comparison between expression-invariant model-based recognition and recognition which is based purely on alignment. The type of alignment too, e.g. the implementation of ICP, can be compared in this way.


To summarise this more formally, we take $N=200$ pairs 480×640 3-D images acquired from N different subjects under various lighting, pose, and scale conditions, then register them using 4 ICP methods, in turn (potentially with variants, time permitting), using a fixed nose-finding method. As the first experiment we may wish to apply this alignment to a set of cropped faces, ensuring that they all lie in the same frame of reference. A model is built from the residual of all 200 pairs, in order to encompass the difference incurred by an expression of choice, e.g. smile or frown. In the next stage, 5 sets of $M=N/5$ images are set as a gallery G and a probe p goes through all images in G , attempting to find the best match best on several criteria such as model determinant or sum of differences. To measure determinant difference it is possible to add the new residual (between p and any image in G), then concatenate it to the set of observations that build the model. This is how it is implemented at the moment. Subsequent experiments can extend to compare other aspects of recognition using the same framework/pipeline. Measure-

ment of performance should be easy if the correct matches are recorded for a random permutation of the set and then paired for some threshold (or best match) based on the gallery.

“The trust of the innocent is the liar’s most useful tool.”

– *Stephen King.*

4 Data



THE data which we work with constitutes one large datasets which is depicting a single subject and another large dataset which is the aforementioned FRGC v2.0 dataset (will be referred to as just “FRGC” from here onwards, for the benefit of brevity). It can be used for direct performance comparison as many papers reported findings based on this data.

4.1 FRGC

For the sake of face recognition benchmarks and general research we have the widely-used FRGC data (see example face in Figure 7), however additional work is required to make it more workable for the following reasons.

4.1.1 Imperfections in Signal

Any datasets which already have reliable refinements/enhancements such as removal of holes/spikes and smoothing would have been best to work with.



Figure 7: Example face from the FRGC dataset

No matter how many cycles are spent parsing through the data we have from the FRGC¹⁰, there is usually some pesky hole or spike left in some images. The noise interferes with parts of the application such as PCA, distance/density checks, and it sometimes seems like a never-ending make-or-break task because when the code is tweaked to perfect the processing of one image, there will usually be another (unseen) special case among the set on which the algorithm falls short. For the framework to be completely automatic it needs to adapt to many different situation without any user intervention, which is exceedingly hard. It's nearly impossible to tell apart

¹⁰We do vectorise the code as much as possible, for the sake of speed. Alas, it is still cumbersome and slow.

erroneous spikes from real ones, without false positives. To start adopting anatomy-specific checks would be computationally expensive (e.g. identifying eyebrows based on local statistics).

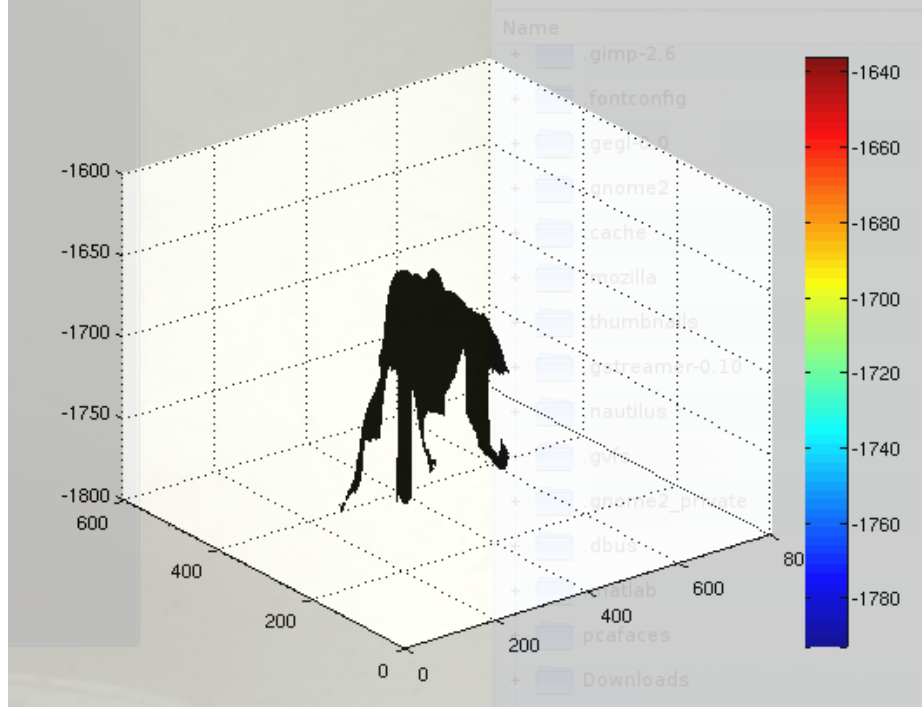


Figure 8: 3-D Image example from the FRGC dataset, demonstrating points on the side of the face – points which need to be removed

4.1.2 Segmenting Parts of the Face

The problem is multi-faceted in the sense that it reinvents the wheel when it comes to refining surfaces (surely there is a lot of literature on that part alone) before one can even start to segment the eyes, nose, eyebrows (if any), etc. Any error in detection anywhere in the set can result in mis-detection and therefore pollute shape residual vectors. Typical problems involve mis-location of the nose tip, especially in the expressions dataset with long hair (this problem was largely overcome for the FRGC dataset). When approaching this problem, contact was made with various people who may have already written code that addresses similar problems, at the very least so as to avoid reproducing – poorly – the same type of code. The cropping phase is already reliable enough and pose correction should not be an issue. Actually, by trying for example not to correct/compensate for rotation we may be able to model also the movement of the head, even though with the addition of expression it would make modes rather fuzzy and the sources of variation less isolated. For n images we'd have $n - 1$ modes, hopefully ones that capture the principal expression changes and not a mixture of structural and positional changes. Instincts suggest that the smarter route to follow is to focus more on model-building and not delve too deeply into segmentation, which can meanwhile be assisted by manual work, e.g. with `ginput()` for 100 images.

Model building is currently more important than grappling with segmen-

tation, so modeling is what we should go for. However, without proper separation between signal and cruft (see for example Figure 8) it is hard to guarantee good results. In face recognition, attempts at recognition by parts was very popular at a point and by manipulating the coefficients within the GMDS or mapping to some average shape, one could properly align the scan (normalise in a way) and from there the path would be much easier than with other forms of normalisation.

4.1.3 Isolated Faces

As figures starting with Figure 9 help show, extracting surface data is made more or less possible, however some blemishes remain and this cannot be allowed in experiments which deal with large datasets and therefore cannot rely on corrective human judgment/input.

Work remains to be done on properly isolating all the faces from irrelevant parts of the image.

4.2 GIP

A large set which is intended to help study and test facial expressions (e.g. validation, recognition, training) is made available in V3R GIP format, along with scripts for reading these GIP-specific formats, e.g. displaying a single frame from a video, parsing file headers, and much more.

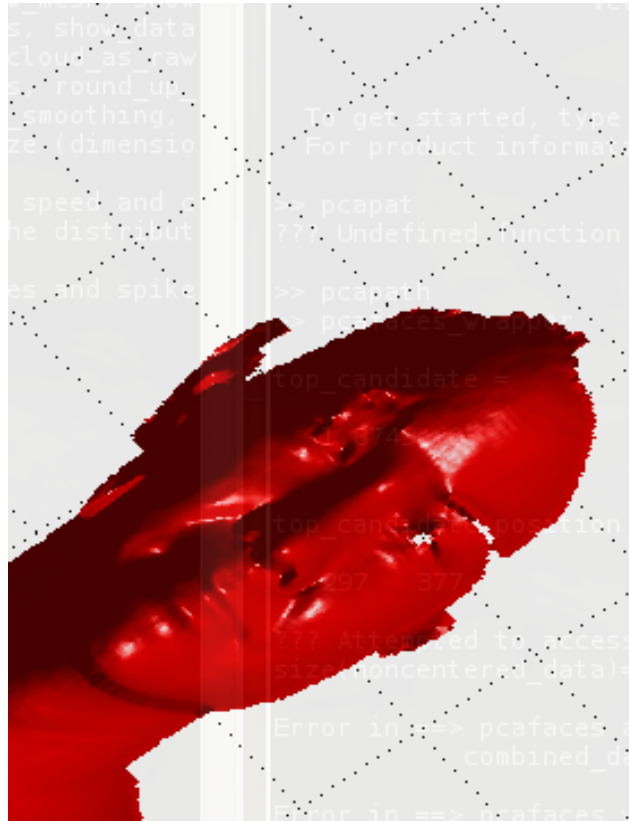
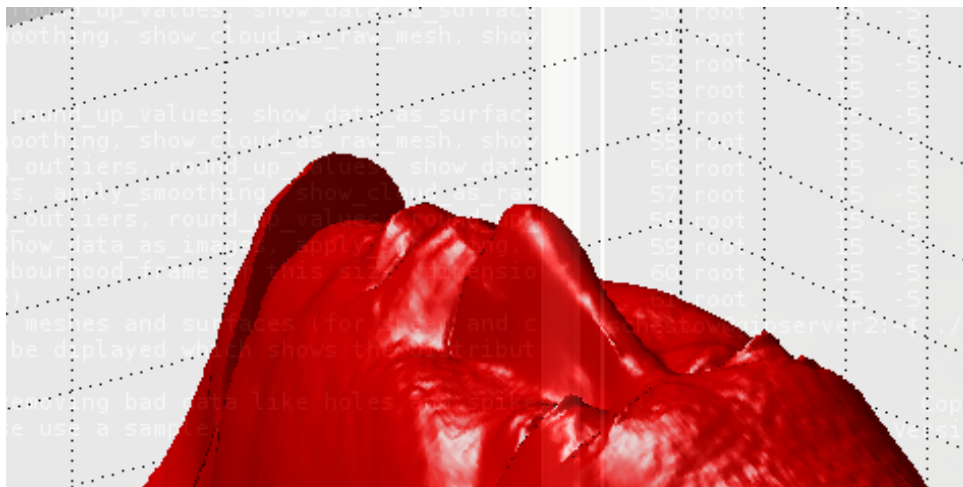


Figure 9: Example face with holes remaining in the data

The expressions dataset is accompanied by different ways to extract the raw 3-D data, remove improper signal using sanity checks/thresholds, etc. Each expression file comprises a few seconds of video, meaning about 10-15 3-D single frames. The easiest way to use that is to pick the frames and extract just the faces from the video. After that it is a relatively easy task to load the files into MATLAB (using GIP-supplied scripts).

Regarding the nature of the imaged subject, it is laid out in consistent locations that cross-frame analysis can help predict. That seems like the right



thing to be using. In fact, building a model for each sequence can be interesting too (although the set might be too small given the density of the sample points, assuming full surface is sampled without increased spacing on a grid).

The nature of data from the expressions dataset suggests that it was acquired differently from the FRGC dataset. In fact, the noise cancellation and cropping algorithms as there were developed to handle the FRGC datasets cannot cope too well with the expressions set; to be specific, there are very many spikes everywhere, which smoothing alone cannot eliminate and the outliers removal phase also needs considerable adjustment. Regarding smoothing, especially since there are spikes, we decided to try median-based filtering, which worked reasonably well. In fact, any version should work fine, separable on x and then y , or vice versa (or even one of them). In practice, median is consid-

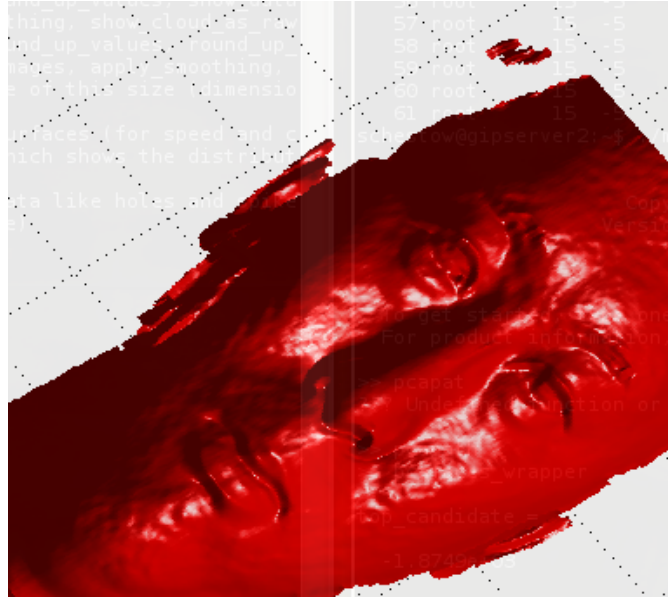


Figure 11: Same as above, different angle

ered along both dimensions and the mean of both values is then taken as the spikes-resistant value. Median is what it is implemented with at the moment, but the one other problem is threshold setting, where the principal drawback is blindly deciding that one spike is noise (e.g. salt and pepper) and another is real signal that should be kept in tact. Having image-specific factors/issues in the GUI would lead to confusion as manually optimising to deal with one image (or image type, based on acquisition parameters) rarely generalises to all (unseen) images in the same set (glasses, hair, and impairments are unpredictable). Spending many hours dealing with pre-processing would probably detract from progress made on the 'meat' of this work, but then again, without automatic, intervention-free pre-processing, there is modeling based on very poor training data, which in turn will yield unimpressive results (poor

breakdown into modes) and raise exceptions due to edge cases. Optimally, getting polished images with noise removed would be nice, albeit it would entirely miss the point of presenting an automatic method (or fully pipelined framework) which can handle data in its raw form.

4.2.1 Experimental Framework for GIP Dataset

The framework for the first experiment – one which builds the expressions model – is now more or less complete, although there are some deficiencies. To name problems which need to be overcome:

- ▷ Despite compensating for GIP/FRGC anomalies such as distance units, scale, noise levels, etc. (support for both datasets is needed for future experiments, where one builds a model and another has it validated and then assessed with an unseen dataset), the hair in the GIP dataset still poses an issue/difficulty. Although the vaguely-described nose tip identification method was implemented (it locates a peak by slicing the image horizontally and then considering tip candidate by measuring intersection of perpendicular line with a sphere), it remains hard to always find the nose without false positives. Smoothing or other filters – median-based for the most part – are very localised, so they cannot reliably eliminate false signal which resembles a nose and lacking nose recognition which is reliable, ICP rigidly/affinely registers non-correspondent parts. For non-rigid methods that taken into account

dense data in its entirety see, e.g. [64, 51, 50].

- ▷ The loaders of GIP datasets extract what appears to be unaligned sequences of images, where one image does not overlap its predecessor because there is an imperfectly-sliced stream of them. To annul this effect and not be confused by misclassifications-causing cruft, additional steps are made necessary. These preparatory steps are intended to remove distractions and biases as it is necessary to have a guesstimate of scale inside the image (for parameter setting), not just nose location (otherwise cropping might fail). It is clear that the majority of the time so far was spent dealing with these issues rather than addressing the more novel parts, notably decomposition and expression expression (not a typo) in a lower-dimensional space.

4.2.2 Additional Data on Demand

We have considered the possibility of scanning more/other expressions, at least if it is required. At the moment we have a database of less than a dozen different facial expressions. From GIP-organised acquisition we have a dataset of some distinct expressions, such as smiling, anger, disgust, etc.

4.3 Synthetic Data

We may also generate some synthetic data for the purpose of testing the algorithms on data where the correct solution is known. This is valuable in

debugging.

4.4 GIP Data Localisation

We now have an adapted image loader for the 90 gigabytes or so of GIP data, much of which (volume-wise) comprises video sequences. The usage of this raw data is unclear at this stage because a systematic set of experiments needs to be decided on based on the utility of this data. For instance, one application might involve determining who is imaged based on an image sequence rather than standalone 3-D images. In prior work from Hack and Taylor there was a Ph.D. thesis on modeling (in the PCA sense) the sequence of talking faces in 2-D, basically predicting the likely subsequent frame based on a training phase. For purposes of straightforward face recognition we have less grainy images that also constitute more individuals and a wider variety of environmental conditions.

The important thing is that GIP data is readily available for loading shall this be required and the data is mounted from a remote storage server rather than stored locally on a computational server.

“All art is exploitation.”

– *Sherman Alexie.*

5 Implementation



PREPARATORY stages, such as pre-processing of images, can be done separately from the main experiments as these stages do not change once the algorithm is deemed acceptable. However, the body of the work too can be made more modular and thus explained by its modules or its layers of operation. Taking stock of what we have and organising it visually gives us an overview image, as in Figure 12. It hopefully helps keep track of the said process, implemented mostly in according with the IJCV reference description.

The work on code can be sub-divided into different stages which can also be treated separately in order to make the experimentation pipeline more manageable. We shall classify the stages as follows: preprocessing, modelling, validation, and benchmarks. The key part is about PCA [38], which MATLAB implements with `princomp`; it is essential for constructing statistical models, via decomposition of face characteristics as derived automatically from the dataset.

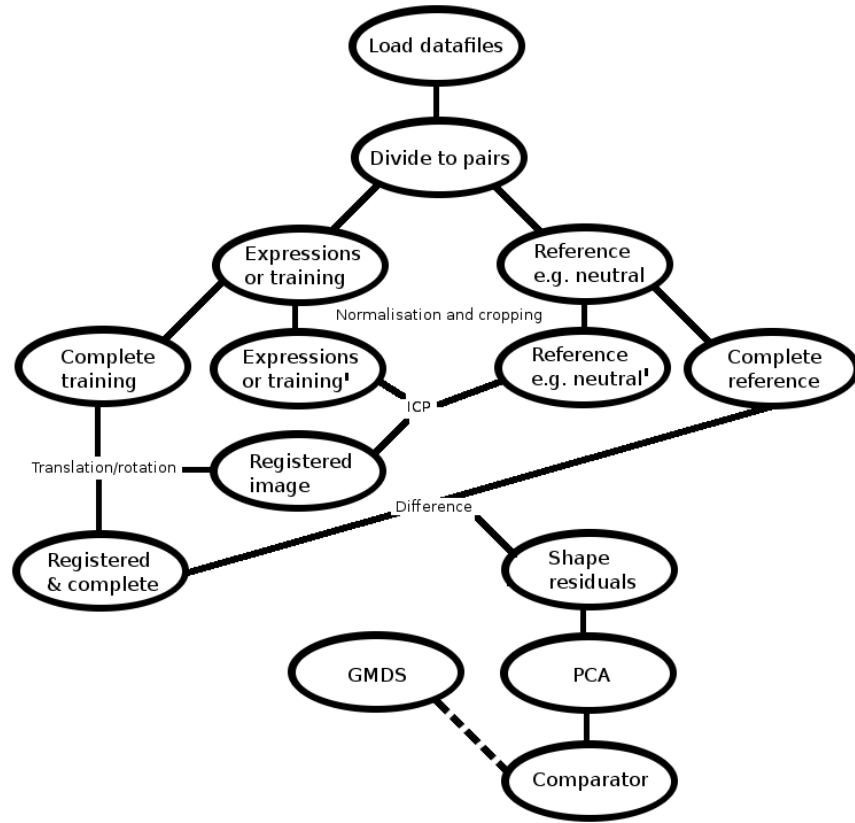


Figure 12: Program steps broken down into an overview-type flowchart

Shown in Figure 13 is an annotated version of the original figure from the paper. The overview is simplistic in the interests of abstraction and elegance. For instance, methods of hole removal are described only as "local statistics". This leaves room for multiple competing implementations with different results depending on *ad hoc* refinement (ours was about 3,000 lines of code at the start of April).

The functionality is two-fold; one major part is modeling and the latter, which shares many pertinent components, does the matching. The program

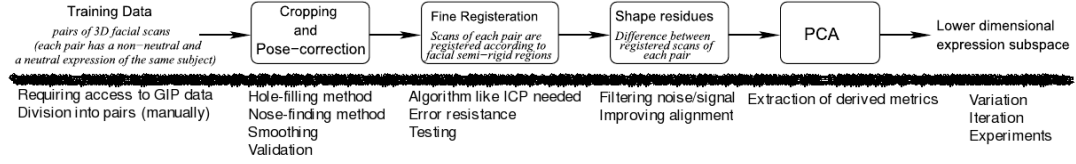


Figure 13: A replotted block diagram of the components in the IJCV paper (top) and our proposed extension/modifications (bottom), and the already-implemented procedures

first takes as input the training set from which to build a model to be used as a similarity measure (done by Schestowitz *et al.* for NRR [67, 68, 66]), which is an overkill that assumes infinite resources like time and RAM. The latter part, which can be scripted to run standardised benchmarks, takes as input a probe and gallery which is assumed to contain just one instance of the same person as in the probe. Initially, two similarity measures are taken. One is residual mean and the other is distance from the N utmost principal axes of the EDM.

For experiments, a control file loops around the algorithm with different datasets and parameters as input. It is worth repeating that the quality of results will largely depend on one's ability to clean up the data – removing the wheat from the chaff (spikes, noise, irrelevant parts of faces) – as that very much determines what an expression is modeled as.

This portion of the document explains to a limited degree the computational methods and puts a lot more emphasis on the software tools.

5.1 Preparation and Preprocessing

Bash scripts were written to locate all of the relevant files in the FRGC dataset¹¹. About 5,000 3-D faces are contained inside. The packages comes with information and programs of interest to those who may find themselves working with 70 GB of data and some accompanying metadata. Given the [Face Recognition Vendor Test \(FRVT\) of 2006](#) and [overview of the FRGC](#) it should be possible to know what is available and where. The latter is an official FRGC Web site. The large package comes with associated applications and scripts written in Java, C++, Perl, etc.

An existing MATLAB/GNU Octave implementation identifies all 3-D images in the dataset. These are located/scattered in many different paths, dependent on time/place. Scripts were adapted to decompressed the files and cycle through them, e.g. to pre-process them in series.

Code was written to perform the pre-processing steps as specified in the corresponding paper (which describes an entire Ph.D. thesis from Australia¹², and very densely so).

¹¹`find | grep .abs | wc` allows us to count and handle all images by taking them one by one and then just handling them one single image at the time. There are many in the current collection, 4950 to be precise; all of them are compressed, as `find | grep .abs.gz | wc` helps reveal. To get a list of the 3-D faces, one can run, e.g. `find | grep .abs.gz | awk '{print $1}' 1>~/files_list.txt`, which yields something like the following: `“./nd1/Spring2003range/04334d218.abs.gz; ./nd1/Spring2003range/04419d182.abs.gz; [...] ./nd1/Spring2004range/04936d102.abs”`.

¹²Faisal R. Al-Osaimi completed his degree in 2010. He was Mian’s first Ph.D. student.

The code was made modular with dozens of options to control what is done and how it is done (through a settings files containing all the parameters), as well as how the data is presented to the user throughout runtime.

1. Each image is taken in turn while the program is performing some analysis that includes a histogram (no manual selection as it would be laborious for thousands of data instances) and visualisation in 2- and 3-D.
2. The image is studied to separate a person from the background and remove all data points associated with the background.
3. The remaining sets of points are made more uniform by filling holes (using local statistics), removing spikes, and smoothing the surface using one among a set of possible methods. We identified better ways of eliminating holes as well as spikes (more generally just noise) on the face surface and then tested the results on a larger sample of 3-D images (the majority is handled perfectly well).
4. The tip of the nose is found and the image is normalised by making its \mathbf{Z} coordinate (depth) zero, then centering it by shifting XY space such that the tip of the nose is at $(0,0,0)$. In order to normalise – so to speak – what remains visible before ICP is invoked to align the data, 3 methods were implemented to select only a region which can be consistent across data sets. The ears and hair, for example, are not

wanted for statistical analysis, so they can be removed by discarding all points associated with them. See Figure 14 for a visual example in 2-D (although rough, there is a similar screengrab shot in Figure 15).

5. We have measured the density in \mathbf{Y} and \mathbf{X} in order to normalise distances, such as the distance from the nose to the chin. More options were added to the control file/wrapper (nearly 20 at the time of writing) and additional function now deals with cropping the face using one of three methods. The best method is capable of isolating the face surface irrespective of the size of the head, which is being centred and brought into alignment at the front. There are then options which define how the face gets cropped to maintain just rigid surface such as the forehead, nose, and eye area (assumes no blinking and eyebrow-raising expressions). The algorithm sets everything necessary – to the extent possible – for ICP to nicely deal with alignment to a common frame of reference.

5.2 Normalisation

Normalisation (with scale) of the face as a geometric structure would be a mistake, so it is worth specifying what we refer to as “normalisation”. While rotation may be fine, scale is a trickier transformation because one must think about the fact that the distance from nose to chin could in fact be used to identify a person. So, the normalisation applied is in no way modifying

anything in the image itself. In fact, nothing is being rescaled or even rotated. The image is being translated only, so as to bring into alignment the many noses, assuming no-one's nose is behind the chest (easy to check for these special cases), the chin, or the forehead (requiring the person to look up or down, although that too can be checked to avoid misclassification). The use of the term "normalisation" refers not to any real transformation but rather to the acquisition of additional data, which allows the algorithm to determine:

- ▷ How much of the face to crop for PCA/GMDS
- ▷ How low a surface beneath the eyes and the nose should be treated as "rigid" for the sake of reliable alignment
- ▷ Suffice to say, this may also come handy when applying ICPv2.

One could, in theory, use normalisation for decomposition (PCA) and then plug back the normalisation scale to compensate for an aforementioned and previously-applied scaling. That is not the method being adhered to, however, because scale can be treated just fine as long as the sample point are selected correctly (with reasonable correspondences marked up); PCA can overcome scale anomalies.

As the next sub-section explains, concerns remain.

5.3 Expression Models

5.3.1 Reproducibility Concerns

Mian's group does not seem to provide enough details about the 3 cases/subjects on which they train. Understanding of this requires careful reading and some extrapolation around the very dense text (it is a whole Ph.D. that's described therein). Currently, trying to get one's head around which images they actually used to do their experiments at the lab is hard. They claim to have had 3,000 images acquired at the lab, but no examples are given and the descriptions are vague at best. To a sceptic, any such thing means that obscurity implies deficiency; this strikes a nerve because if they basically test on their own data and use just three unique physical faces (with variation), there is no guarantee that the results can be generalised, so to speak. Moreover, there are repeated admissions of weakness and argument for the training of person-specific models, which seems to be a realm more capable of easy handling. At one stage it is explicitly stated that the PCA applied to 300 FRGC instances (enough to accommodate hyperspace of limited proportions) was rather useless, which led to thinking, "might we expect to reproduce these results at all, despite the fact that the original researchers themselves had reported difficulties?"

5.3.2 Apparent Limitations

Any data we may have of few individuals and a breadth of facial expressions from them may be useful for a comparable demonstration of results. Had we had no access to their (apparently proprietary) data, experimenting with it is a walk through smoke and mirrors. Open Access and Open Data are particularly important for this reason, at the very least as complementary, auxiliary material in one's paper. In many cases, code too should be made available for audit (at all level), in order for one to defend the results.

Considerable time/attention should be dedicated to ensuring that the training phase is done with data which is known to have yielded good results; we do not quite have that from Mian and based on a quick survey of the stock of FRGC images, there are rarely cases where one individual was imaged more than a handful of times. Without the ability to prove that good eigenvectors can be derived from the set, leaping towards ROC curves and systematic experiments would be very premature and clearly time-consuming.

The intention here is not to defend GMDS by scrutinising a counterpart's work; rather, it's about understanding what it is exactly that they show and how it was achieved. For example, were comparable databases tested on? And if so, how were the models trained? Is there a preparatory phase the outsider if not being informed of? Models can surely be refined by studying variations 'off-line', without delving into large sets with significant variation.

5.4 Registration

ICP has a working implementation and is now capable of delivering back the 6 degrees of freedom which define a shift to apply to non-reference face data, typically non-neutral one (conventions may vary). What's required is a set of neutral plus non-neutral pairs, organised suitably for a data structure which treats pairs as even and odd items in the index of an array. In its current form which is extensible, the program takes pairs as input, applies the necessary pruning, and then aligns the two at the centre. Rotation and translation (no scaling) are applied and the non-neutral image is fit to the neutral one based on closest points. This process helps determine the resultant shape residuals, which it retains until the entire set of pairs is processed. As described in Mian's paper, the residuals are then vectorised; the program reduces the scale and complexity of PCA (to prevent running out of memory) based on 4 variables which define a sub-frame to sample points from. Each residual instance is uniformly serialised and then fed into a principal component analysis algorithm that yields very many eigenvectors. At this stage, with test data that deals not with real pairs but with pseudo-pairs, the output of PCA is not particularly fascinating. The IJCV paper does not show anything too spectacular either, but decomposition of the variation does capture – although roughly – movements of the mouth and then descends to less interesting properties. If a good model is built, then removing expressions by using it 'in reverse' ought to be doable, which in turn neutralises some

variation that otherwise impedes identification. Implementation was sped up considerably while coding in a way that reduces spurious computations and vectorises some expressions to dodge excessive looping (which cannot be optimised on the go).

5.5 Modelling

Generating faces with different expressions should be possible in both 3-D and in 2-D, e.g. for the sake of testing the method. Upon defining an experiment the group is able to make available a set of galleries (a couple of subjects, multi-expressions for example).

We now have a working implementation of ICP. Having spent a reasonable amount of time viewing FRGC galleries, it was hard to find anything non-neutral; that's probably by design. The first experiment currently being implemented requires pairs of the same individuals, with and without a smile (or some other expression for that matter). Then, ICP can be applied and given enough shape residuals of pointclouds (PCA won't tolerate any less than 100 instances as a very low bound for some meaningful eigenvectors), it ought to be possible to build a model, synthesise from it, and animate (videos are fun). However, whilst all the pairing code was put in place, there is no suitable data and the implementation is too slow, especially all the preprocessing. It is harder than one would imagine to reliably remove all the holes/tears/folds/spikes in every single dataset, including unseen ones

(almost 5,000 of them). Once the algorithm runs sensibly – without need to backtrack and refine – making 'offline' copies of reduced sets should be a high priority; this way they can be pulled from file and used 'on the fly', just like models whose covariance matrix can be loaded instantaneously from file. A reasonable path forward might be to look at some example data, analyse accuracy/consistency/robustness and then prepare all the data for large experiments.

5.5.1 PCA

MATLAB provides a simple implementation of principal component analysis – an implementation which Octave does not yet have. This permits basic correlation finding, which in turn yields several things that can be measured. There is a lot of information out there about PCA.

5.5.2 GMDS

Pioneered by the Bronstein duo, GMDS is described in their many recent papers, e.g. [12, 8, 11, 13]. As explained in prior parts, GMDS can be used for a lot more than face recognition as its application can be further generalised (e.g. to analysis of non-rigid shapes). GMDS deals with isometric embedding, where the intrinsic metric structure of some given surface can be represented using another surface, which in turns yields some correlation between these two. The numerical framework proposed enables the finding of

correspondences. Measures of similarity can also be derived from the average metric distortion and the Gromov-Hausdorff distance advances matching of surfaces, using this exact same framework. In the context of our work, GMDS should be used for (dis)similarity in an objective function we ought to define.

5.6 Control Files

Although deprecated due to the graphical user interface (GUI), experiments are still possible to script and to run from a set of files where batches of tasks are specified. These are entirely external to any of the actual functions involved. The same goes for the GUI, which is separated and fits nicely on top of the program's functionality.

5.7 Graphical User Interface

Rather than leap to the experiments while a suitable wrapper is not yet available, in the month of March we started building an interactive GUI that can complement control files (potentially daunting to a command-line-*apathetic crowd*). So, once we made a GUI for the program (see figures 16 and 320) we resumed work on experiment design, which a later section will elaborate on.

When we had completed and polished the GUI, then added the expressions loader to it and added a visual mechanism by which to view 2 pairs of images

while they are being processed (a progress bar was removed as it would provide no indicators of much value; see screenshot) it was a lot more user-friendly and yielded visual output in a more organised or cohesive fashion (including 3-D images). Figure 320 helps show how 2-D image representations are supported, even though it is out of date and abundantly spurious to those who are just interested in algorithms.

5.8 Remote Access to the Program

As agreed by peer after a testing session and a one-hour conversation, we have modified the code such that it should enable any person to run the program from my own home directory (no need to pass any of the files around), even without being the owner of the user account. All that should be needed is a small change which percolates down to all the code. Without losing elegance, the change is already applied, so the program as it stands at the moment ought to work. I cannot test this as I have no account other than mine.

```
schestow@server:/home/schestow~$ ./mb
```

I removed (commented out) the HOME= directive as we found that it generally confused X and led it into the wrong ~/.Xauthority, yielding authentication errors. The advantage of this approach might be that instructions can be provided for any user with an account on the server to reproduce the results for oneself using the GUI I created. Given the circumstance of work from afar, this is a handy utility provided it's easy enough to follow (running

the shell script above ought to invoke the GUI, regardless of who executes it).

“The choice of C is the only sane choice.”

– *Linus Torvalds.*

6 Experimental Framework



A lot of the necessary work for constructing an experiment is described in prior sections (loaders, GUI, etc.) and now that there is handling of data with aim of input alignment¹³ we look forward to some experimental work and results that support the main hypothesis, which is about GMDS having an upper hand over PCA, at least for faces. At the moment, until pairs of images are set up for the initial, proof-of-concept experiments which validate the approach by showing extraction of biologically sound eigenvectors (we have rudimentary code for eigenfaces too), the least we can do is conceptualise and write down expectations.

While we look forward to seeing something similar to the pictures in the IJCV paper, in which the authors attempt to synthesise expressions via the eigen-structures, it may be necessary to vary another set of parameters when dealing with GMDS.

¹³This is very much needed for landmark points and thus correspondences to be identified, otherwise we model nonsensical examples.

6.1 Validation

6.2 Residual

Included now is a routine hole-filling code. It's part of the program, which also shows shape residual for debugging purposes, as shown in Figure 18 for example.

Mian's group vaguely alludes to some difficulties and workarounds that help tackle special cases of mis-detection, whereupon they use thresholds to rule out error or even the mere possibility of error (i.e. real signal which seems suspect based on algorithmic judgment).

I have begun working on the matching algorithm – that which pairs images to galleries of possible matches and then optimises over model parameters in order to find a match which best fits for each and every image. The convenient assumption is that only an expression-removing transformation will give a global minimum (or minima), but maybe it is a tad *ad hoc*, in that sense that it works in practice even without always yielding what it says on the tin. Either way, the ROC curves compare performance of rigid ICP-based algorithm to that of an equivalent, expansive non-rigid approach, which obviously will show the latter doing better; before dealing with the GC set the goal is to produce ROC curves for GIP data. But the matching part is more tricky than it may first appear; there is open admission in the paper that spikes and cruft creep in, so only some hacking around can help

in ensuring that valid signal is exclusively preserved.

The **residual** images show why: 1) the nose tip detection is crucial to modeling, otherwise the example must be discarded (standard PCA is not resistant to outliers); 2) there are many sharp edges at the borders and they must be removed, otherwise they will dominate the signal in the training of a model and thus become principal components which we do not want. Figure 19 shows a residual as it typically appears (with spikes) and Figure 20 shows what happens when the algorithm fails to crop the nose at the right positions.

At the moment, over 90% of the images in GIP data have the nose (and subsequently the whole structure of the face) detected correctly. The progress made so far is encouraging given that we are able to gradually reproduce a whole Ph.D. project at the capacity of just 50 hours per month (like one week of full-time work). The milestone which is ROC curves can hopefully be reached without much in the form of technical peril.

The recently-composed (and not thoroughly tested yet) code enables the separation between the training and matching phases, which in turn makes future experiments a lot faster (data offloaded to files). Several more experiments were run to test the ground and iron out a few more artefacts like noise. The framework in mind is one that will, as prescribes, morph out expressions using the model that we already build, applying the search to all images in the gallery and then assuming the pair with least dissimilarity to be of the same person.

Untold problems with the modelling phase are not technical issues which are associated with the methods; rather, they are to do with the nature of the data, especially the way it gets normalised and considered within a particular rigid frame of reference (and faces cannot be explained rigidly when pertinence parts flex and move about). In the interest of progress, we ought to press on and implement all the method, then refine them (later on when time permits). The reason for this is that comparison – namely between PCA and GMDS – is more important than absolute results for the time being. Mian *et al.* further improved their methods (separate branch of their algorithm) which they then demonstrate in the ROC curves. So, being hopeful that comparable results will be obtained is reasonable, but expecting a rudimentary implementation to fall short is only realistic.

Shown in Figure 21 are some of the residuals that would give a hard time and demand a lot of testing to make elegant.

6.2.1 Residue Filtering

The implementation incorporates a filter component which carves out shape residue based on intuition and not a clear description of a method elsewhere, e.g. in the IJCV paper (there is none specified in the paper, just the observation that it's essential). They appear to be using a hybrid of tricks to achieve this, so we start with threshold that is permissive enough to accept natural variation. Next, a special case to remove border differences will be

implemented. Figure 22 shows a real example as it is being processed in phase 1.

In order to improve models that are being built automatically, i.e. without human intervention, effective removal of noise and outliers is done with spatial thresholds (depth and boundaries perpendicular to the camera). There are more examples in Figure 23. These are examples of outlier stripping in the shape residue, without further cleaning of differences near face borders.

6.2.2 Further Data Preparation

In addition, due to cases of mis-detection of the nose (about 10% of the time in the case of GIP datasets with 3 examples shown in Figure 24), improvements were made to the cropping algorithm, which now has 4 methods implemented. The latest one yields something which best resembles the IJCV paper. This ought to help future implementations involving not just PCA, which is largely implemented but requires good data to operate on.

Images are shown in Figure 25, which demonstrates what cropping tends to look like.

6.2.3 Binary Masks

A set/groups of binary masks, such as the one shown in Figure 26, are being used to remove unwanted parts of the image residue, in addition to other alternations to the data – alternations that are intended to give models which

can be presented for meaningful variation. Had a method for mask synthesis and data making been formally described, it would have been a lot easier.

Looking at the same type of transformation but combined (both thresholds and mask) from another angle, it looks something like the examples in Figure 27.

6.3 PCA and Projection

The pairing/match assessment phase gets implemented with a regular principal component analysis (PCA), where there is no particular scaling applied, meaning that it is problem domain-neutral. The program requires just over 1.8 GB for a small model to be saved, about 3GB of RAM/swap to run depending on the size of the training set (the weight is dominated by the amount of data stored in memory rather than offloaded to disk). For better performance and handling of larger experiments, a redesign will be needed because, as it stands, the computational servers already stretch to the swap partition, which slows them down considerably.

Looking at the reference description of the implemented method (whose work ours is conceptually derived from), I found some small error in their paper and I can think of better methods that would achieve better results given enough time. The usage of PCA, for example, could be improved by undertaking a proper model-fitting task – a stage for treating the whole problem as one of optimisation, where the varied parameters are the high-ranked Eigen

coefficients. The original work uses the projection from which squared errors are extracted (could use Hotelling's T-squared statistic instead). It is too optimistic an approach that incorporates hacking around it with truncation, binary masks, or zero assignments *en masse* (removal of what fits poorly, without clear description of how, just why). Evidently it does give some decent results, but there is room for suggestion and/or belief that it's far from optimal. They do speak about the artefacts we too get at the borders, they name a threshold in millimetres, but there are no formal and specific details about the method being used (this is not the only example of missing details about more opaque parts of the said methods). We can, in due time, reverse-engineer – so to speak – their pertinent set of algorithms, but for the time being there is some guessing and generally an incomplete pipeline, especially that which is associated with polishing the residuals. It is similar to the problem we used to encounter and then tackle when it came to pre-processing images – a problem which is largely resolved now but still needs more polish (time-consuming and counter to measurable progress).

The code necessary to generate ROC curves is almost complete (projection as an objective function needs further development), but then it becomes just a brute-force routine, which would go to waste if there is still a lot of false signal (or noise) in the processed data. From a paradigmatic point of view, the important pieces are nearly in place and they are packaged in a user-friendly GUI and easy-to-follow functions, which are properly interfaced too (needed for script-based looping and automation). The syntactic side of

things has room for improvement, e.g. vectorisation as means of replacing loops that MATLAB won't optimise away.

6.4 ROC Curves

We shall be looking at ways of designing experiments whose final step helps in producing error charts and then ROC curves. The comparative results will hopefully demonstrate an advantage when GMDS is used. The plan is to produce initial ROCs/EERs that support Mian's results and we will try to reproduce their experiments. Failing that, we may try to contact the group for help or for missing data. It is defensible to suspect that things are more complicated than they seem on the surface (pun intended).

The types of experiments that would make sense to run are:

1. Neutral to neutral identification comparison (easiest case)
2. Training set versus neutral (including non-neutrals)
3. Everything available versus neutral
4. Arbitrary non-neutrals versus neutrals (hardest case)

In order to get the results of the expressions comparators in low-dimensional space experiments need to be designed such that they include everything we have compared to all of the available expressions (i.e. use all training

including expressions) as a final step. Alternatively, we could generalise expressions from neutral to all. The above 4 steps should come first anyhow. What one hopes to show is that by aligning a fully-correspondent set and then creating an expression deformation model, one gets particular achievable results (with GIP dataset) and with GMDS one can attain better results, perhaps even under difficult conditions which one method is more resistant to than its counterpart. There seems to be bias (by data-fitting) in the original paper and by reproducing some of the experiments we can hopefully show that the opposite of what was claimed there is true.

In order to compare ROC curves (corresponding to what Mian's group had attained) it seemed reasonable to construct similar tests. There are still missing bits of code; for example, visualising the resultant eigenvectors, getting proper alignment all the time (very crucial), reliable and consistent cropping (a black art of trial and error), and most importantly comparators of model fit to examples. The problem is, rushing towards getting results – any results – without improving dependent parts won't serve towards getting something which is workable. If the best one can do is show a reproduced algorithm performing at rate of – say – 70% rather than 90% detection rate, then no strong claims can be made about one being inferior. In these cases you have two options; 1) the said paper was cheating or 2) those who try to mimic their counterpart's algorithm intentionally implement it poorly so as to get the desired performance gap. Suffice to say, it is harder to cheat when there are standard tests one must conform to, but in any event, the plan is to show

a similar framework in an apple-to-apple comparison, where basically both methods are implemented the same way with the only distinguishable difference being GPCA [87, 49] and GMDS swapped¹⁴. Then, rather than adhere to comparing methods on an absolute basis a relative comparison can be made, with a paper demonstrating that an MDS- or GMDS-based approach works better than PCA for human faces and more specifically anatomy of expression.

We finished building a GIP EDM and expected to see what experiments based on model criteria can be run now. It will be interesting to get a sense for TP/FP rates and then beat that by changing similarity measures and cleaning up the data a little further (automatically, not manually). Ron said, “Regarding the 77% figure they are right, but this was a very pessimistic test aimed as proof of concept rather than a face recognition tailored one. There was no much intelligent pruning, no alignment, no special treatment of missing parts, etc. I am in fact surprised we got the 77% altogether.

“Moving from PCA/MDS to GPCA/GMDS would introduce (I expect) some

¹⁴In one of their original papers, the authors’ abstract states that they present an “algebraic-geometric solution to the problem of segmenting an unknown number of subspaces of unknown and varying dimensions from sample data points. We represent the subspaces with a set of homogeneous polynomials whose degree is the number of subspaces and whose derivatives at a data point give normal vectors to the subspace passing through the point. When the number of subspaces is known, we show that these polynomials can be estimated linearly from data; hence, subspace segmentation is reduced to classifying one point per subspace. We select these points optimally from the data set by minimizing certain distance function, thus dealing automatically with moderate noise in the data. A basis for the complement of each subspace is then recovered by applying standard PCA to the collection of derivatives (normal vectors). Extensions of GPCA that deal with data in a highdimensional space and with an unknown number of subspaces are also presented.”

flexibility to the modeling that would enable capturing small variations about the given expressions and thereby better recognize the identity of a person...”

6.5 Benchmarks

6.6 Extensive Work

The long term goal is to eventually replace the PCA with G-PCA, or more accurately, with GMDS. I.e. use the existing result for alignment and then measure the discrepancy between faces by embedding one extracted portion of the face to another and use distortion as a measure of similarity. Under the assumption that G-PCA is generalised PCA, yes, it’s easier to see where this is going. Prior work on GMDS proved viability of distortion as similarity measure, but that cannot be compared on a like-with-like basis against the other group’s results.

By applying shifts to the average shape (of either a person-specific or group-specific) corresponding to a residuals model they can show synthesis and by minimising an objective function it ought to be possible to do fitting, too. The parametric space is very high-dimensional though and it is hard to believe fitting is something which was done in this context before (Cootes *et al.* have methods for enabling it, even when texture is added to a combined model). One might wish to demonstrate GMDS-driven fitting algorithm.

We used the functions as they were given to us and leveraged them to get

video sequences loaded and sorted for the PCA experiments 28.

We eventually built and saved a relatively large EDM from GIP datasets. It can be used later on. Figure 29 shows a decomposition of that.

6.7 FRGC Experiments

As template experiments with data that was collected from distinct subjects are needed (in order, for example, to perform recognition tests and use a classifier for ROC curves), experimental protocols from FRGC were sought, finding a suitable training set and target/query/matching sets.

This process varies from learning some of the metadata and experiment scripts/SGML/markup/data to looking at some sample data in 3-D, which is non-trivial to browse/navigate without some special/ised visualisation tools. For FRGC 2.0, the master file in

`FRGC-2.0-dist/BEE_DIST/FRGC2.0/metadata/FRGC_2.0_Metadata.xml`

contains a suitable index and experiments from previous years can also be run using the older protocols at

`FRGC-2.0-dist/BEE_DIST/FRGC1.0`

Early tests with the data (see screenshots of the development and experimentation framework in Figure 30, as well as figures 31, 32, and 33) suggest that we need to make some new binary masks for residuals.

“The right to be heard does not automatically include the right to be taken seriously.”

—*Hubert Humphrey.*

7 Ongoing Progress and Results

THE resultant models and their performance when used in generative mode were the initial results we got for the aforementioned framework. But over time we moved on to exploring other areas, which also fall under this massive section, for reasons of convenient.

Having overcome most of the barriers associated with pre-processing of GIP data, models were built using a couple of separate sets, each containing a sequence of expressions of a certain type. These two were compared in the discrepancy sense, which in turn was modeled by PCA and yielded Figure 34 (experiment from 9/4/2011 with nose tip search near the centre, with smoothing, expression of surprise compared to sadness).

Fear compared to surprise 35.

Values of 0.0018746, 0.0016749 and 0.002639, accounting for 38.9452, 34.7974, and 26.2574 percent of the observed variation.

We have run experiments on the largest expression datasets we have, weighing 440MB and 880MB (see Figure 36). This algorithm works a lot more reliably now, however matching/scoring remains to be done. We expressions-free datasets to register to?

Next, matching criteria based on the model will be used to score for recognition rates; larger experiments can then be engineered to produce ROC curves. In order for such large experiments not to require reruns, however, it will be desirable to further refine the existing implementation while improving the GUI, the documentation, and also ensuring the route taken is widely accepted and not an enormous effort embarked on in vain.

Figure 37 shows the percentage of the explained variation for sets of 9 images; “fear” to “surprise” yields 24.0210, 18.9342, 13.2081, 12.6385, 10.1781, 9.1887, 7.0255, and 4.8060 percent, corresponding to the magnitude of the 8 modes of variation.

At the point where we are prepared to run larger, systematic experiments there might be use for GIP’s implementation of ICP. The detection has come to the point where it can quite reliably capture the right parts, at least based on some early observations.

All control of probe-to-gallery experiments was moved to the GUI side such that iteration will be simpler and require less manual work or future coding. The implementation was made somewhat more elegant by merging similar bits of code and ensuring deprecated parts are removed or hidden away in

secondary options. Assuming that the data is reasonably clean by now¹⁵ (for GIP data it is a lot more manageable now), we are ready to run experiments and then vary parameters to extend these to ROC curves, as planned. EDMs can be constructed for pairs of expressions or neutral/non-neutral. The 3-D portion of GC data is said to contain just smile shots and neutral shots for each subject, so perhaps for these experimental results will be dependent upon whether the model is trained with just smiles or all sorts of different expressions. Perhaps the GC models should be treated in total isolation from GIP ones, as the curves adhere to a certain standard which is independent of locally-acquired data.

GIP data contains data from one subject (at least for expressions), so coming up with a way to test recognition of a one subject among many (inter- or cross-personal) is a non-starter. Preliminary results ought to show only feasibility, so these will be run as a side task while we return to GC data for experimental validation.

7.1 Visualisation

As visualising the eigenvectors should require a reshape back into image form it ought not be hard to demonstrate some results more visually. What is being passed to PCA is a vectorisation of image data, where the density of the

¹⁵In page 12 of the IJCV paper they state that "about 2% of the probe scans" were misdetected due to these errors that we have, but they have had more time to work on these. It's the less finely documented part of their work, which presents itself as trivial by hiding the creases.

sampled points is a variable recently added to the GUI sliders (MATLAB runs out of RAM if everything is sampled) and for the time being a compromise is also made, where NaN values are cast 0 (ideally should be the result of interpolation). It's not too clear how to model/treat those because they cannot be skipped; the `reshape()` function will expect some sort of value in there and not all faces are alike in the sense that they cannot overlap one another perfectly.

7.2 Statistical Analysis

Section 8 will cover that in depth, using example.

7.3 Detection

Same as above.

7.4 Automation

In April we were putting the final touches, improving the program's automation options (there is a sub-window/child window for that). We developed this to run future experiments by specifying ranges of files to perform comparisons with, specifically for object matching. The program writes to standard I/O at the moment, but it can also be made versatile enough to assemble results in static files (e.g. CSV-formatted).

7.5 Similarity Measures

At the moment there are 3 similarity measures implemented: one checks the absolute differences (squared if that is deemed required) applying the residual, one adds the residual to the PCA model and assesses the changes this imposes upon the model, and the third (which is a slow one to run) is a family of methods that warp any residual corresponding to a pair into EDM space and then checks the differences, assuming there is nothing suspicious in the data which made the models (Mian's group too had problems with this). The PCA modes that we currently have have been improved considerably by reworking two phases; one is preparing the data for alignment with ICP and one filters the data for subsequent stages that compare pairs.

7.5.1 EDM Versus Pure Residual Approach: Correlation Between PCA-based Approaches

In page 13 of the IJCV paper there is a head-to-head comparison between GMDS- and EDMs-based similarity – one that takes what's said to be verification rate of 77% at 10^{-3} false acceptance rate (FAR). Their recognition rate is not too high, about 93%. The subsequent work involves manageable experiments as measured in terms of scale, using lab-collected data to construct person-specific EDMs and then plot recognition rates as ROC curves. Given GIP data from different subjects, this is reproducible.

Working with groups of similar expressions from the same subject (screen-

shots available), e.g. by taking residuals computed from

'~/Facial-Expressions-Recognition/fear.v3r'

to

'~/Facial-Expressions-Recognition/Suprise.v3r'

and then also considering the diffs from

'~/Facial-Expressions-Recognition/Suprise.v3r'

to

'~/Facial-Expressions-Recognition/Joy.v3r'),

we do have agreement between match score for the model-based approach which yields (for the first six pairings):

$1.0e - 04 * [0.0839, 0.1897, 0.1435, 0.1322, 0.1971, 0.1946]$

And for the non-model-based similarity measure (mean of differences in the residuals):

$[0.0107, 0.0190, 0.0169, 0.2056, 0.2825, 0.2562]$

The best match so far is image #1, based on both measures that agree (the model based and pure residual/difference-based).

7.6 Performance

Now come the CPU-heavy parts, requiring longer runtime, larger covariance matrices (the memory footprint of the program sometimes exceeds 3 GB

of RAM because we sample as densely as possible), and needing further debugging, albeit not necessarily in code but in tweaking of parameters that improve input data.

7.6.1 Performance with GPU Boost

Read some more papers and got access to Tesla GPU-powered server this month, operated via SSH with KDE as a guest environment, as before. Even though our code is not compiled for Compute Unified Device Architecture (no SDK available here, either), I thought I'd at least change the default password as it had been sent unencrypted over E-mail. I seem to lack a homedir on the server and cannot create one without being root

7.6.2 Dataset

To avoid over/under-fitting, size of database matters a lot and our performance depends on the quality of the data used for training. We could try just running it on the NIST dataset, but it would not reproduce what the IJCV paper states – a paper wherein they basically worked on 3,000+ 3-D images taken from 3 individuals at their lab. They trained their statistical models based on this vast dataset before applying it to an unseen test set.

Finding a conceivable way to access many data files or pass them to/mount them on the computational server (as done with NIST data). The program is ready to run experiments, it just needs 3-D data from different individuals.

7.6.3 Data Deficiency

The codebase is in a state where it branches out to accommodate and handle different data types without the user needing to perform manual adjustments. Having constructed and tested binary masks for GIP datasets, similar binary masks needed to be made and tested for the data in the GC experiments, where scale variations are more common but can be modelled nonetheless. See Figure 38 for 2 examples of masks we can use (assuming roughly fixed scale).

The datasets in the experiments' protocols are not sufficient because, as noted earlier, more data (from fewer individuals and as little as 3 should be enough) is required for training. To quote from IJCV (page 10 or 311 in the published book):

The FRGC dataset was augmented by 3006 scans that were acquired using a Minolta vivid scanner in our laboratory. 1 The 3006 scans belong only to three subjects (1000 non-neutral scans and 2 neutral per subject). The non-neutral scans were acquired while the subject was talking or reading loudly and with the intention to produce as diverse facial expressions as possible. The FRGC dataset have scans for a large number of subjects but we also need a large number of non-neutral scans per subject for experiment no. 3 (Sect. 3.5). In addition, some of them are used in model training (Sect. 3.2) as it turned out that the expression

deformation model requires large training data.

3.2 Model Training Data

The training partition of the FRGC was not sufficient for training our Deformation Expression Model. Insufficient training data can result in noisy eigenvectors (of the model, Sect. 2.3), especially those with lower eigenvalues. Also, a smaller training data may lack enough instances of facial expressions of different people. Consequently, the model may not perform optimally during face recognition. To test how the size of the training data affects the performance of the deformation model, three Expression Deformation Models were trained using training data of 400, 800 and 1700 scan pairs. Then, they were used in non-rigid face recognition of 400 unseen probes under non-neutral expressions with an appropriate subspace dimension of the deformation model (the dimension is 55, see Sect. 3.3). The identification rates in the three cases were 89%, 93%, and 95%, respectively. The rates have increased for larger training data sizes.

In the following experiments, the Expression Deformation Model which gave best results (the one which is trained by 1700 pairs) was used as the generic deformation model. The 1700 pairs were formed among the training partition of the FRGC dataset (943 scans), 597 non-neutral scans from the evaluation partition (leaving about 1000 non-neutral scans for testing) and 500 scans from

our acquired data.

to reproduce the results we may need access to this data. They seem to be reliant upon it and the GIP datasets are a tad different in nature, at least in the sense that they provide a lot of expressions from the same subject or many neutrals from a lot of different people, which makes the problem similar to that of running experiments with orphaned GC data and protocols (although it would not be identical to it, which only complicates this further).

It is clear, based on explanation given in the text, that we need to have very large (and augmented) training sets to be training a model from which to take the 55 utmost dimensions. Using the Perl scripts provided in BEE to parse the XML files would not be sufficient to reproduce the models and replicate the results. Upon failure it could be argued that we had not followed the documented procedure.

In order to plan analogous experiments we ought to check if their results are reproducible based on the details given in the paper and granted, if there is difficulty in reproducing these results, the authors can be contacted for pointers or perhaps be contested.

Reasons for scepticism in this case may seem unfair, but there are points of weakness that are only found along the way, as all of these things are being put together and then highlight issues which must be overcome – issues that are only mentioned/alluded to in the text but not properly addressed in a formal, technical sense (not even by reference to prior work). All this wiggling

room is the reason framework mimicking has been slower than expected. Mian's prior work (he received his Ph.D. under Bennamoun's supervision in 2007) is basically the fundamental work his first student's work is based upon. It is an extension of his rigid, ICP-based implementation. This means that we need to decipher some of this prior work and build upon it the extensions that Osaimi put in place until his graduation last year. It's not a monumental task, especially if we aim to only disprove their parts about GMDS benchmarks, which seem unfair as they compare apples with oranges. We contacted the first author regarding availability of their dataset which they augment FRGC 2.0 with.

7.7 Benchmarks

[FRGC Data Set](#) or [as a paper](#) (CVPR 2005)

The FRGC data distribution consists of three parts. The first is the FRGC data set. The second part is the FRGC BEE. The BEE distribution includes all the data sets for performing and scoring the six experiments. The third part is a set of baseline algorithms for experiments 1 through 4. With all three components, it is possible to run experiments 1 through 4, from processing the raw images to producing Receiver Operating Characteristics (ROCs). The data for FRGC consists of 50,000 recordings divided into training and validation partitions. The training partition is de-

signed for training algorithms and the validation partition is for assessing performance of an approach in a laboratory setting. The validation partition consists of data from 4,003 subject sessions. A subject session is the set of all images of a person taken each time a person's biometric data is collected and consists of four controlled still images, two uncontrolled still images, and one three-dimensional image. The controlled images were taken in a studio setting, are full frontal facial images taken under two lighting conditions and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions; e.g., hallways, atriums, or outside. Each set of uncontrolled images contains two expressions, smiling and neutral. The 3D image was taken under controlled illumination conditions. The 3D images consist of both a range and a texture image. The 3D images were acquired by a Minolta Vivid 900/910 series sensor.

The FRGC distribution consists of six experiments. In experiment 1, the gallery consists of a single controlled still image of a person and each probe consists of a single controlled still image. Experiment 1 is the control experiment. Experiment 2 studies the effect of using multiple still images of a person on performance. In experiment 2, each biometric sample consists of the four controlled images of a person taken in a subject session. For example,

the gallery is composed of four images of each person where all the images are taken in the same subject session. Likewise, a probe now consists of four images of a person.

Experiment 3 measures the performance of 3D face recognition. In experiment 3, the gallery and probe set consist of 3D images of a person. Experiment 4 measures recognition performance from uncontrolled images. In experiment 4, the gallery consists of a single controlled still image, and the probe set consists of a single uncontrolled still image.

Experiments 5 and 6 examine comparing 3D and 2D images. In both experiments, the gallery consists of 3D images. In experiment 5, the probe set consists of a single controlled still. In experiment 6, the probe set consists of a single uncontrolled still.

7.7.1 FRGC 2.0 Experiment 3

In FRGC 2.0, Experiment 1 and Experiment 2 contain just clusters of 2-D data for training and targets. Partitions of volumetric data are available through Experiment 3, with XMLed links to non-existent files containing the `.sfi` and `.t.sfi` suffixes/extensions, even though these are now stored in `.abs` files (initially the Spring 2003 range), with numbers differing and textures stored in loadable `.ppm` files (their number is one above the `.abs` files, they are potentially valuable for future experiments, but irrelevant to

the current work and thus discardable). We wrote script files to find and uncompress all the files, then grepped and neditted the XML files to give a list of the files that we want (32 GB of it in total), eventually piping them into MATLAB-style data structures.

Experiment 4 uses a large still training set and there is not much to be found in Experiment 5 and Experiment 6. The 4 experiments from FRGC 1.0 have the same deficiency and therein we find even more of a binary element rather than XML for the BEE framework. So, we are definitely left with just 3 'sub-experiments' which are derived from the third and it ought to be enough as a baseline.

Considerable time was spent trying to find patterns in classification of facial expressions in the 3-D datasets. Out of 4950 images, about 1000 are used for training and the rest are targets (no separate query set). This is the largest 3-D face database available out there and it is said to contain 4007 shape data instances collected from 466 individuals in the gallery (acquired with a Minolta Vivid 900/910 series sensor), the rest being probes from a laser scanner, not an optical one. This ought to help increase the difficulty of the problem, e.g. by reducing consistency in the signal. In case it helps, also given are the manually-marked up coordinates of the noses, eyes, and chins. Huang *et al.* [36] show examples taken from the same subject (top row in Figure 39) and difficult cases with holes and occlusion (bottom row). To train a model there needs to be consistent mapping for separation between neutral and non-neutral instances. We have that for GIP data, but the set is small

and not identical to what was used in the published paper from Australia. They might also have expression classification which is needed (a lot of work to redo, so querying Houston University too might be worthwhile).

The exploration and navigation around the existing data ought to have helped remove doubt about existence or absence of necessary data, to remove the possibility of something already being available but hidden away. It does not seem like there is much other than more software tools that are irrelevant to us. This was already clear from the documentation of the FRGC, but looking at each experiment for future insights and understanding of syntactic characteristics was worthwhile.

ROC curves derived in Experiment 3 are conventionally referred to as ROC I, ROC II, and ROC III. These measure the face recognition performance for target and query from the same semester's data collection session, collection in the same year but in two different semesters, and then collection taking place in different years, respectively. The latter represents the most challenging of the three due to all kinds of changes in the environment (e.g. background type, location of face, and intentional lighting variations). Literature survey by search does not bring up many experiments which use these protocols for 3-D, not as strictly described for Experiment 3 anyway. It does appear as though Bennamoun *et al.* [54] have used this as a yardstick for quite some time. For example, regarding prior work, which was multi-modal, they summarise: "We present a fully automatic face recognition algorithm and demonstrate its performance on the FRGC v2.0 data. Our algorithm is

multimodal (2D and 3D) and performs hybrid (feature-based and holistic) matching in order to achieve efficiency and robustness to facial expressions. The pose of a 3D face along with its texture is automatically corrected using a novel approach based on a single automatically detected point and the Hotelling transform. A novel 3D Spherical Face Representation (SFR) is used in conjunction with the SIFT descriptor to form a rejection classifier which quickly eliminates a large number of candidate faces at an early stage for efficient recognition in case of large galleries. The remaining faces are then verified using a novel region-based matching approach which is robust to facial expressions. This approach automatically segments the eyes-forehead and the nose regions, which are relatively less sensitive to expressions, and matches them separately using a modified ICP algorithm. The results of all the matching engines are fused at the metric level to achieve higher accuracy."

This paper helps explain some of the otherwise-unexplained bits from the later papers, e.g. nose-finding approach. There are a lot of dependencies among these disparate bits of work from them, which is cumulative in the algorithmic sense.

"Encyclopedia of Biometrics" (Volume 2), a book by Stan Z. Li and Anil K. Jain, has some more valuable details about the dataset in question and the protocols one must adhere to. A group from the University of Houston describes its work there (first author is Professor Ioannis A. Kakadiaris, Director of the Computational Biomedicine Lab). They also divided the set into neutral and non-neutral, which is curious because such a division is not

pre-supplied, which leaves bias to the assessor (not standardised). I will make contact to inquire about it.

Some more manual work may be required for splitting the data as in the identification case; the dataset needs to be split into a gallery which contains just one face of the same subject as the probe, which means that only one correct match would be possible. This can be achieved by taking the first image of any individual and then treating it as the gallery part, the rest being probes which need to find/match it. The cumulative match characteristic curve can then be plotted. For verification, measuring the fraction of datasets which are returning "positive" would also be needed; the false accept rate (FAR) measures how many of few are classified as positive given a threshold, e.g. 10^{-3} FAR.

The organisers of the Grand Challenge only have this to say about the neutrality of faces (in CVPR '05 [60]): "The controlled images were taken in a studio setting, are full frontal facial images taken under two lighting conditions (two or three studio lights) and with two facial expressions (smiling and neutral). The uncontrolled images were taken in varying illumination conditions; e.g., hallways, atria, or outdoors. Each set of uncontrolled images contains two expressions, smiling and neutral."

There is a proper database which allows images to be fetched based on modes, where the attribute 'mode' can have the following values:

mode="Standard"

This mode only cares about a target/query pair of recordings if the query recording was taken after the target recording. In these cases it is considered a match if the subject IDs are the same, and a non-match if they are different.

```
mode="FRGC_2.0_ROC_I"
```

This mode only cares about a target/query pair of recordings if the target and query recordings were taken in the same year, and the query was taken seven or more days after the target. In these cases it is considered a match if the subject ids are the same, and a non-match if they are different.

```
mode="FRGC_2.0_ROC_II"
```

This mode only cares about a target/query pair of recordings if the query was taken seven or more days after the target, regardless of year. In these cases it is considered a match if the subject ids are the same, and a non-match if they are different.

```
mode="FRGC_2.0_ROC_III"
```

This mode only cares about a target/query pair of recordings if the query was taken in a later year than the target. In these cases it is considered a match if the subject ids are the same, and a non-match if they are different.

```
mode="Identity:/home/user/filename.mmx"
```

This is a special mode that simply copies an existing match matrix file to the output file and ignores the target and query

signature sets passed in on the command line. The file to copy must be specified after the colon and can be an absolute or relative path. This is the only mode that does not require interfacing with bBase-Lite.

There are also 3 masks that are mapping matrices and go along with the ROC ones, but these do not correspond to expressions. Today and tomorrow I will start building models with the data we have, even though for the EDM experiments Bennamoun and his group concede that it is insufficient. They added their own. Figure 40 shows the sort of images which need to be dealt with.

If we have most modules from previous papers implemented, there is a baseline for comparison between existing methods and novel ones. Again, the first that comes in mind is after the ICP refinement of the alignment via GMDS (either with geodesic distances or even Euclidean ones, which is a minor modification to the ICP method). Guy and Dan at the time were converting their ICP into a black box everyone could use (at the time of writing it was unusable).

Currently, we have two ICP implementations (pluggable methods, interfaces requiring output as xyz for translation and 3x3 matrix for rotation, although Euler and Cartesian would work too). Alas, we have not done any systematic experiments to compare the performance of each. Actually, I ran some overnight experiments that build a model from the entire 3-D training set

(~900 images) after a lot of planning last night, but the program crashed just one hour into it, so I need to debug and plan for the next night. It will be useful to see if computing resources, especially RAM, are sufficient for doing so. If not, then the program needs to be rearchitected. Then, assuming that Kakadiaris' or Bennamoun's groups get in touch, we may also need additional data (Open Access) like neutrality/expression semantics and datasets for augmentation, as prescribed by them.

In Encyclopedia of Biometrics (Volume 2), a book by Stan Z. Li and Anil K. Jain, I saw your group from the University of Houston describing its good work. You divided the set into neutral and non-neutral I would like to inquire about access to such data. I currently work with Prof. Kimmel and we need this type of classification which raw FRGC data does not make available.

7.8 Full Model (EDM) for FRGC Data

Some examples are innately more complicated than others, e.g. the case where borders intersect with part of the region to be sampled or where X and Y data has holes. Figure 51 shows one example where the camera zoomed in or the subject approached it, leading to the chin breaking some constraints and raising exceptions. For a fully automatic approach which deals with thousands of images like this, a solution must exist in code and not be sought through manual processes, not even rejection of the data instance.

7.8.1 Newly-built EDMs

We can, in principle, work on the ROC curves right now, however they would not be impressive and this also requires a lot of work that can be avoided if we get sent additional data from groups that plotted these curves (ROC I-III).

The good news is that after some failures I did manage to build an EDM (expression model) from the whole training set of FRGC. That's a set comprising 943 3-D datasets in total. The bad news is that I have not heard back from Faisal, so I ended up contacting his Ph.D. supervisor. We need their data to reproduce their results. They do not use only the FRGC training partition.

In the mean time I am adding more similarity measures and then using these, testing them with some galleries (we still need more metadata such as expression or subject ID, e.g. in order to split properly into subsets).

Today I delved into visual inspection of the model as visualisation leaves bare the coarser elements of it, which result not so much from misregistration as much as from the difficulties inherent in data filtering, bringing justification to further tinkering with parameters such as thresholds and binary mask types (there are 4 type of these right now).

At first we encountered some improper cropping as shown in Figure 42 and later I resolved that with better code (on the GIP dataset it already does this reliably enough, but there is low diversity there, being intra-personal

and acquired in one location with one modality). What we are getting right now into the model can be seen in Figure 43 and Figure 44. The model takes several hours to build, which is not too bad given that it is not a compiled program and not a dedicated server, either.

got access to the ICP code. This server has always been hard for me to access as it's not UNIX/Linux-based. While waiting for neutrals, gallery, and ROC I-III metadata we have been looking at the models I built from FRGC data. With access to UWA's missing data, better EDMs can be built, but in the mean time we looked at PCA space and visualised it a bit (see Figure 45), also ironing out some bits of code and bugs along the way (mostly UI-related).

"Unfortunately," explained to us the group of the original EDM paper, "the volunteers (from whom the data was collected) did not give us permission to distribute their 3D face data." We understand this fully and will try to acquire similar datasets at our lab. Sadly, we cannot have the data they used to perform benchmarks, but it seems safe to suppose that similar experiments can be designed with different data. The EDMs would be different.

More EDM work was used to explore the newly-built models. Looking at the models that lack the necessary training data from UWA, Figure 46 shows the distribution of modes' weight based on the model built from the FRGC datasets.

Figure 47 reveals interesting circular patterns that show point-to-point corre-

lation along the 10th utmost principal axis (similar to the previously shown images in Figure 45).

Figure 48 is similar to the previous ones, but it is a representation of scores rather than principal modes of variation.

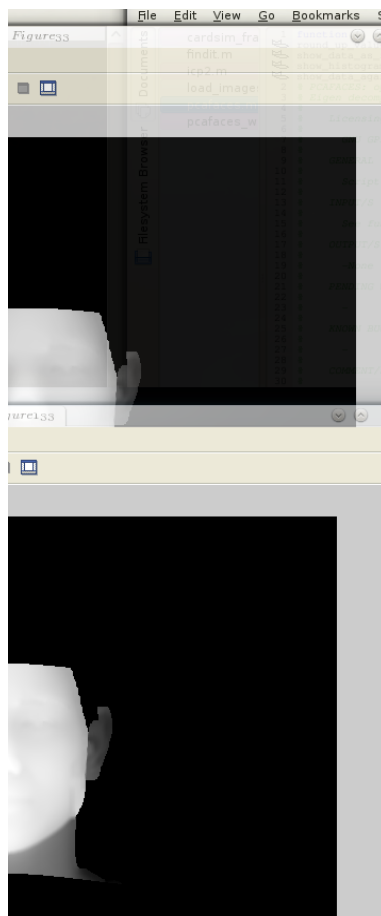


Figure 14: Translation of the given (cropped) face applied so as to position it with the nose tip at the front and at the centre

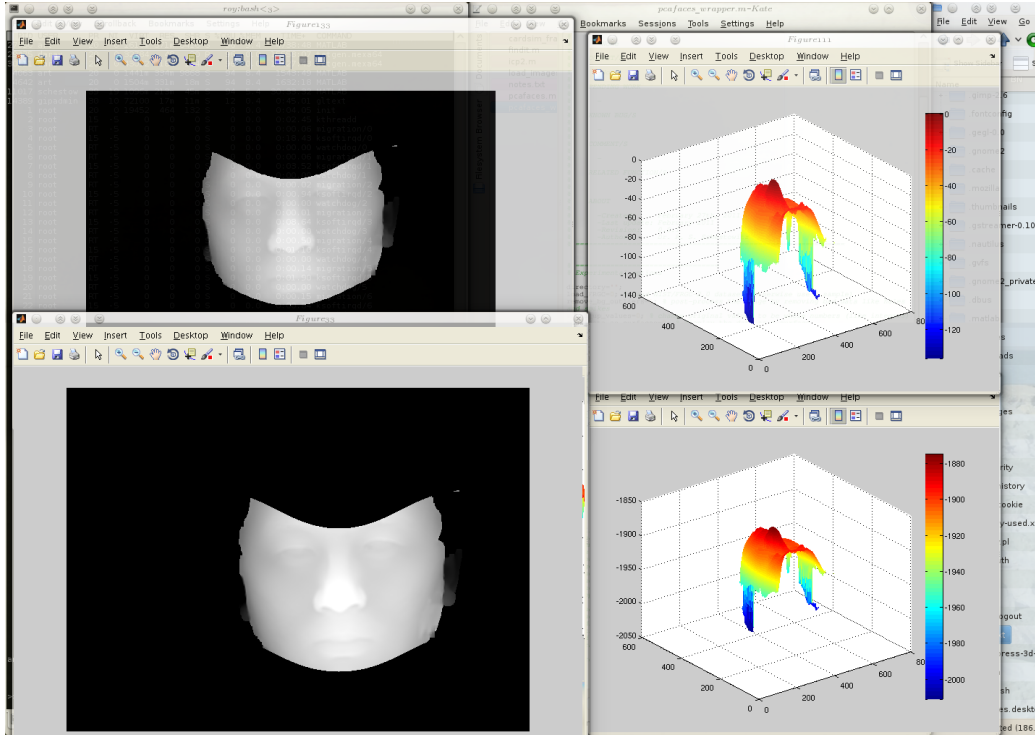


Figure 15: A before/after overview

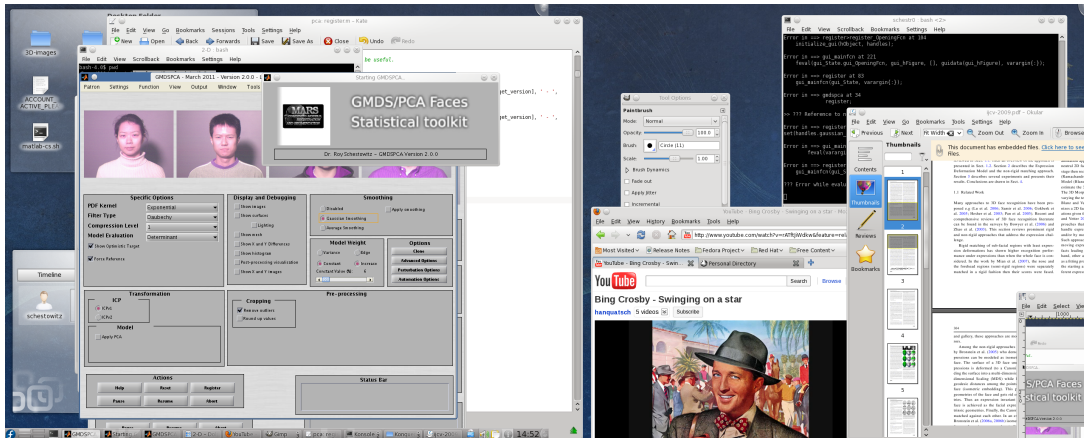


Figure 16: Early prototype of the GUI

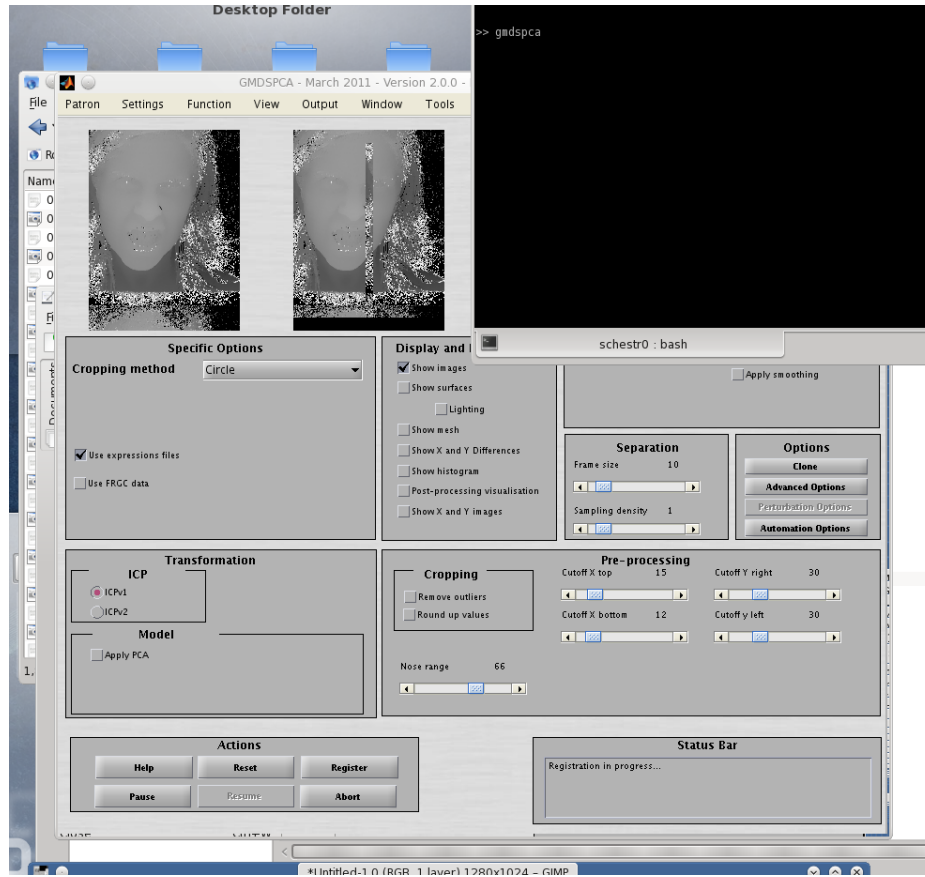


Figure 17: The same GUI at a later date

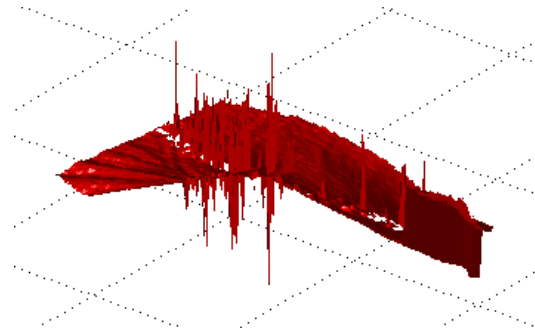


Figure 18: The shape-residual extracted from two different images of different people, where the faces are aligned so as to fit a common frame of reference.

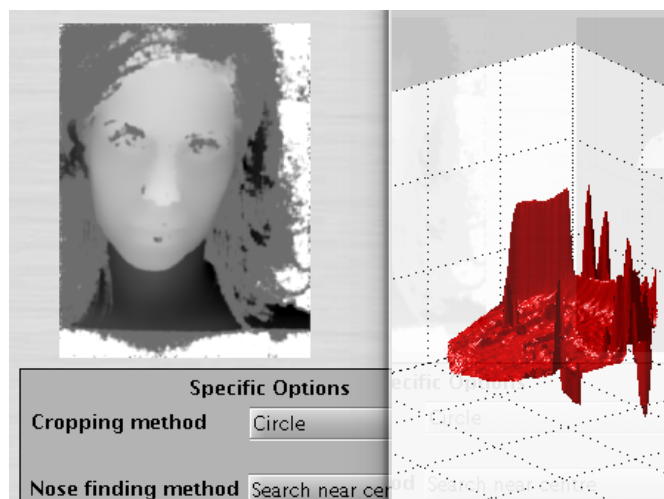


Figure 19: A sample image and a corresponding residual wrt to another (unseen) image

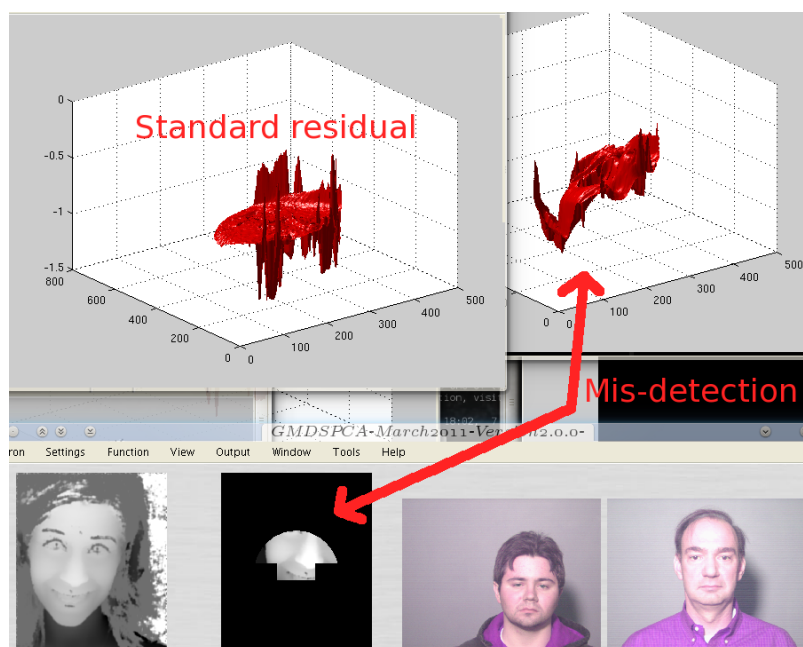


Figure 20: Example of what happens when the nose is incorrectly identified

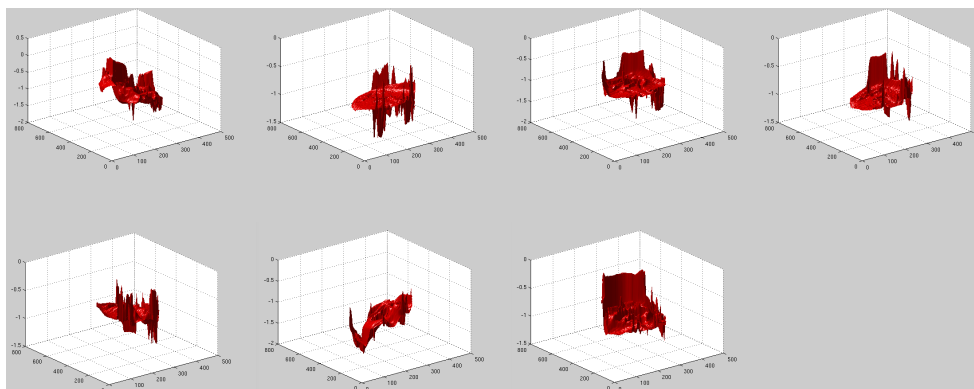


Figure 21: Examples of challenging residuals that have a lot of noise

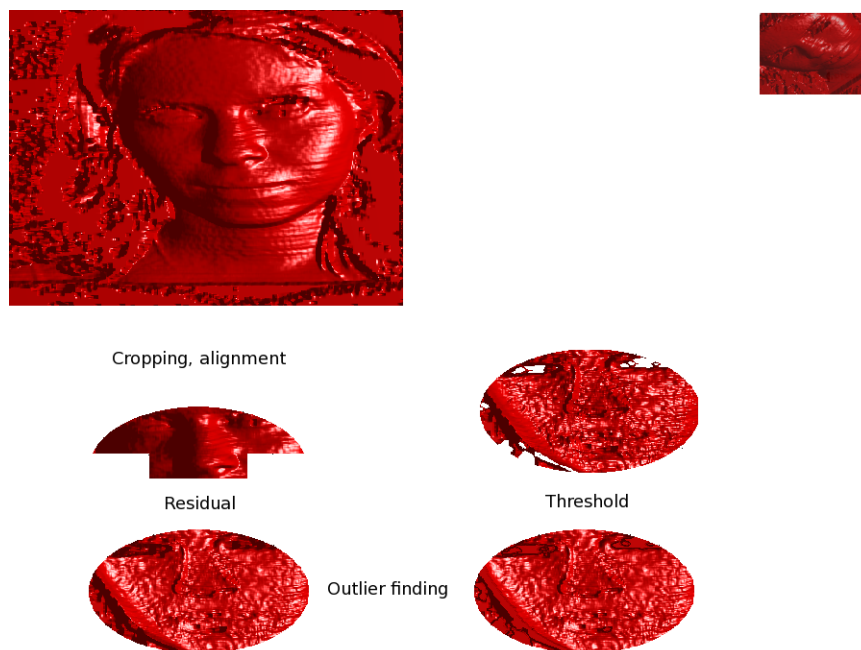


Figure 22: Process of cleaning up the residual of two images (GIP data)

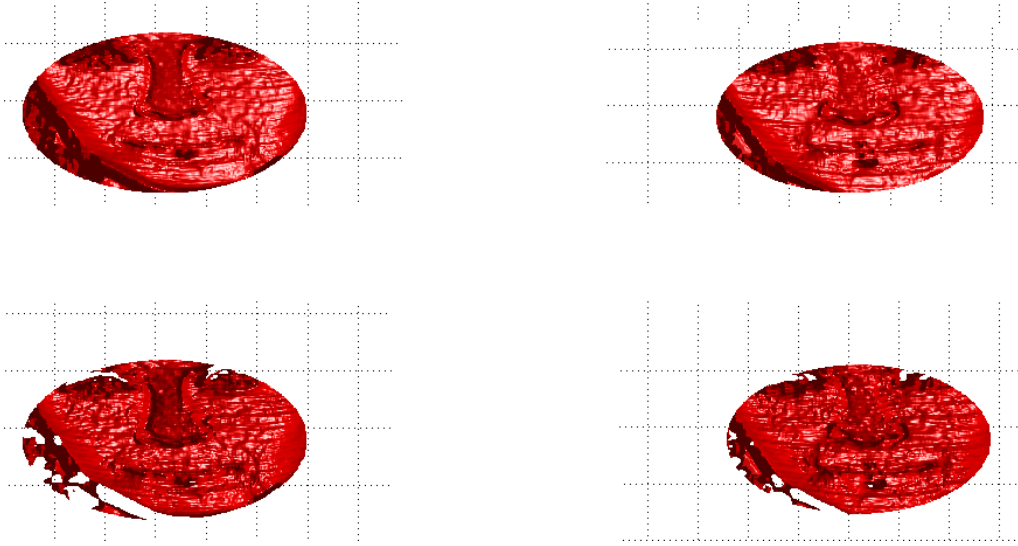


Figure 23: Examples of residual images before and after outliers removal

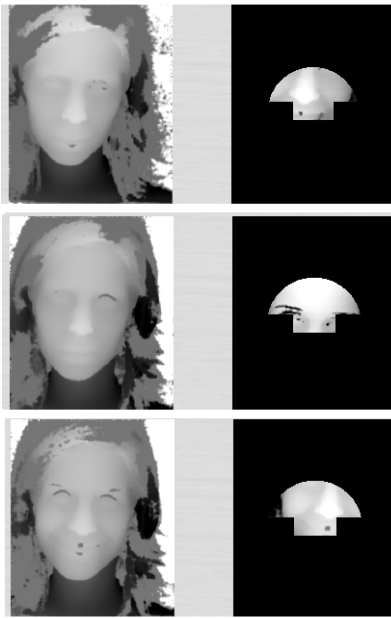


Figure 24: Examples of faces that, given the default set of parameters, do not detect the nose correctly (with old-style cropping)



Figure 25: Improved cropping of faces that takes spatial measurements into account

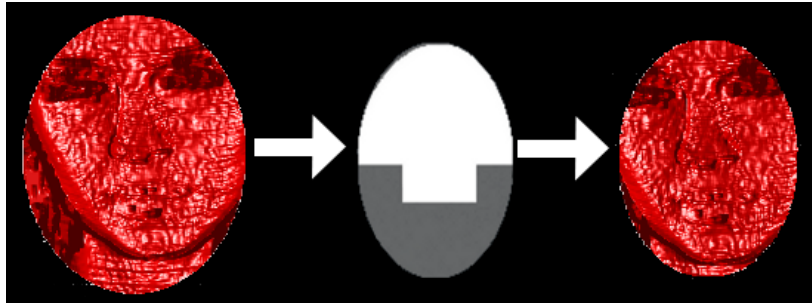


Figure 26: Example of binary masks being applied to image residue

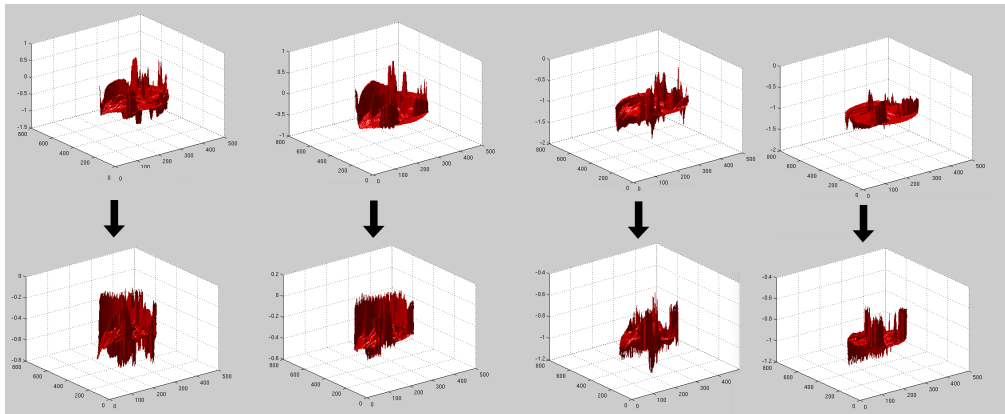


Figure 27: Examples of 4 raw residuals being reduced (notice the Z scale) using thresholds and masks

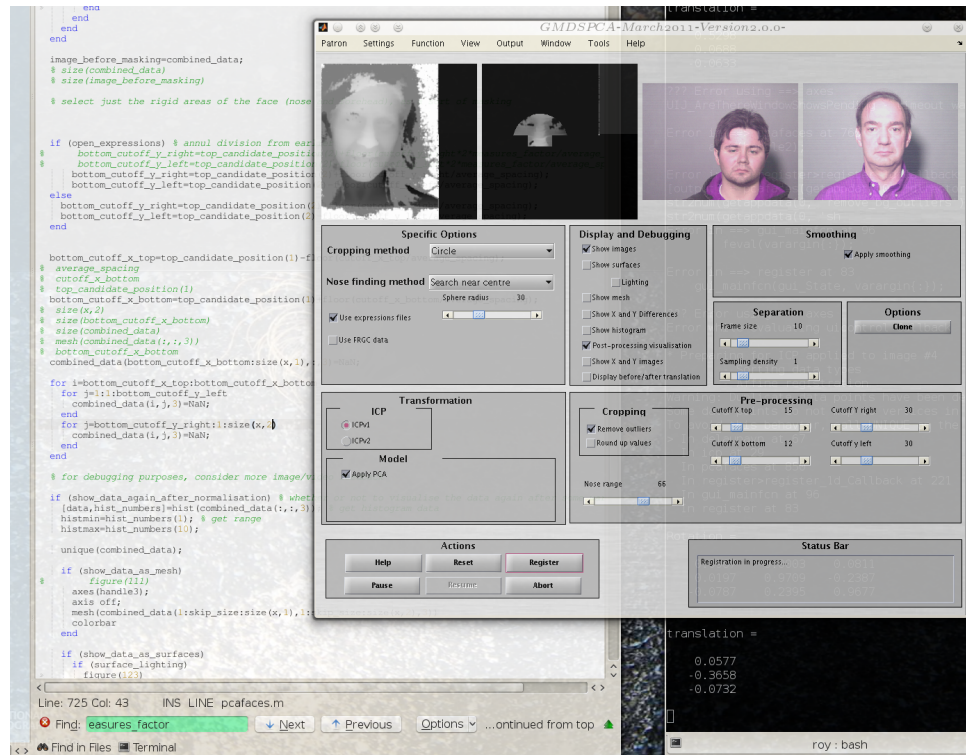


Figure 28: Cropping of GIP data shown on the top left

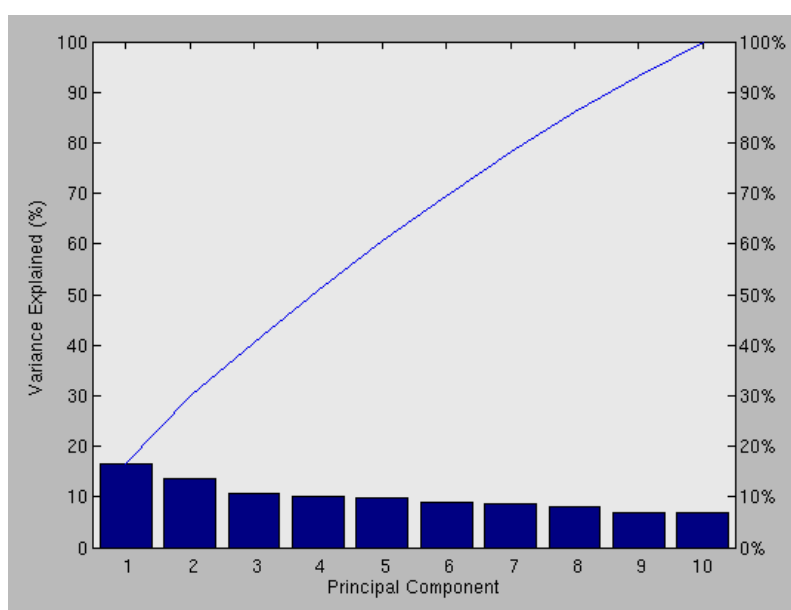


Figure 29: Model modes decomposed for a couple of GIP datasets (abbreviated to account for top 10 modes alone)

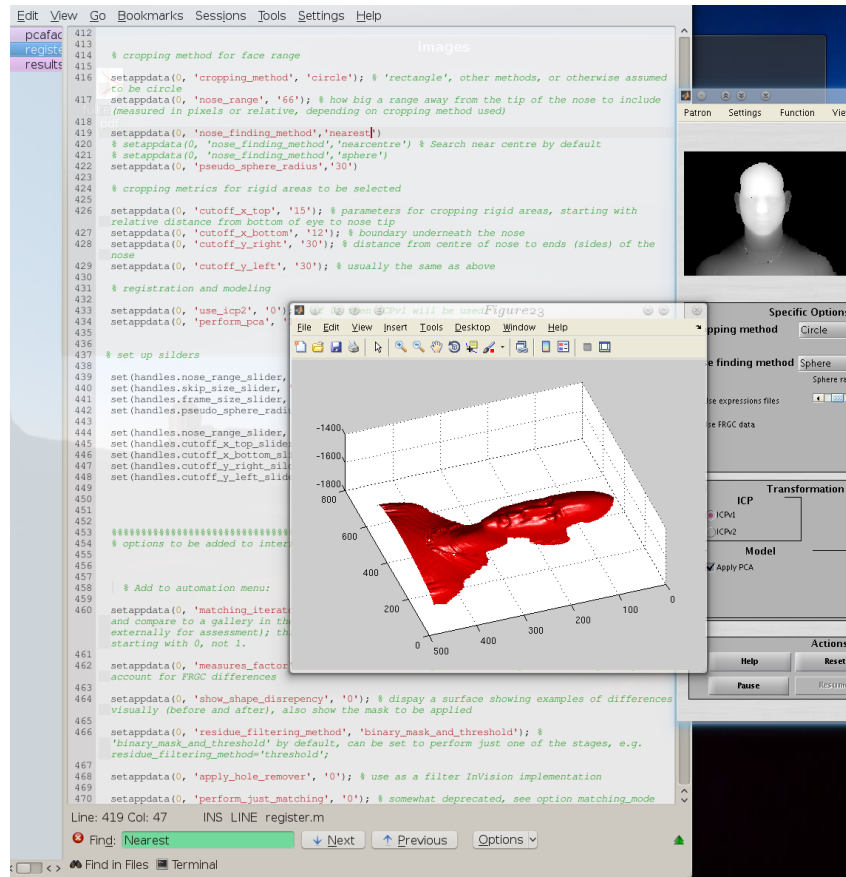


Figure 30: Example 3-D representation of an arbitrary face image from FRGC 2.0

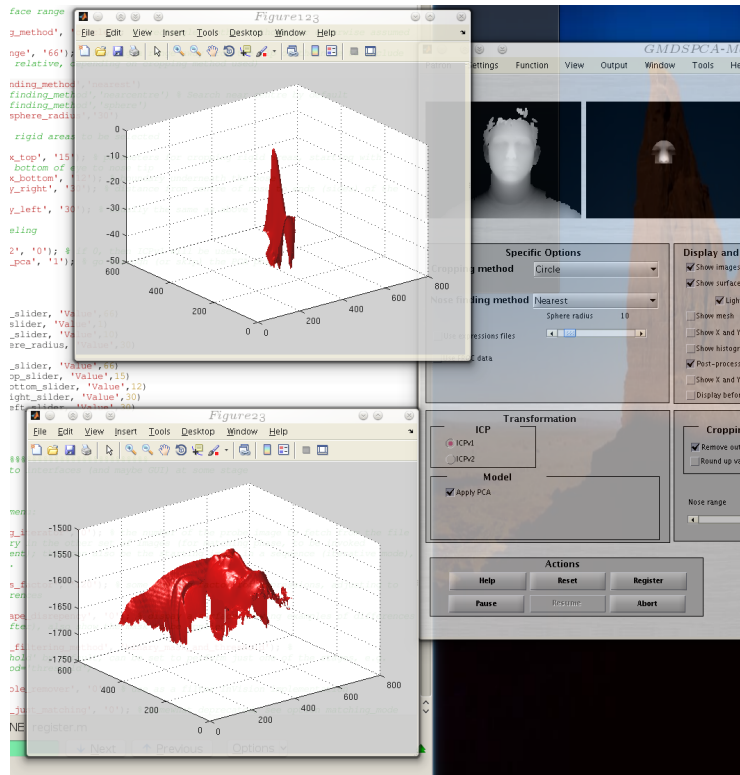


Figure 31: Example of alignment and cropping of the rigid parts of another face (mind axes scale) with the result shown at the top surface and right image inside the GUI

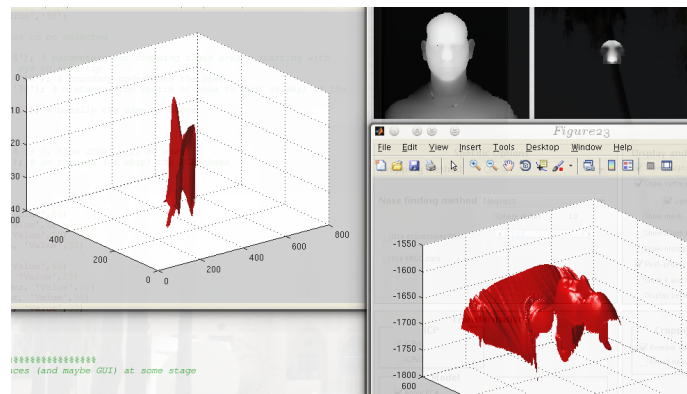


Figure 32: Example of alignment and cropping of the rigid parts of another face with the result shown at the top surface and right image inside the GUI

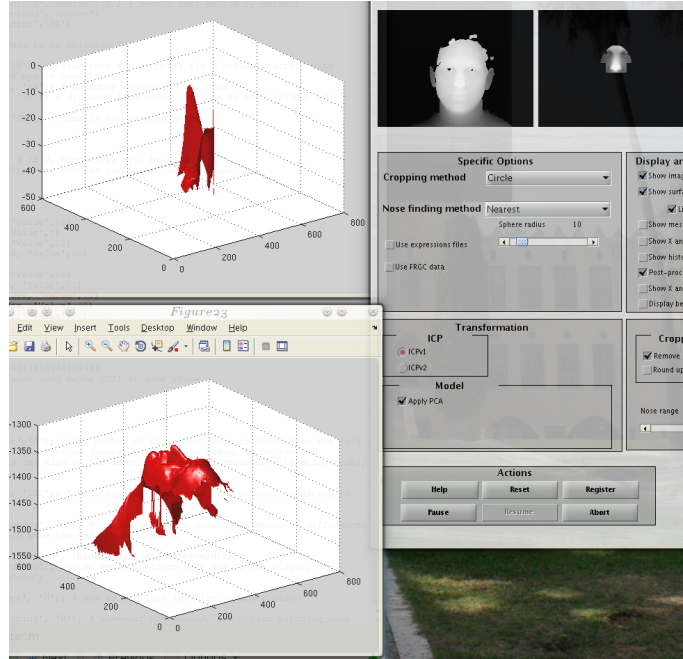


Figure 33: Example of alignment and cropping of the rigid parts of another face there there is some noise and hole that pose a challenge

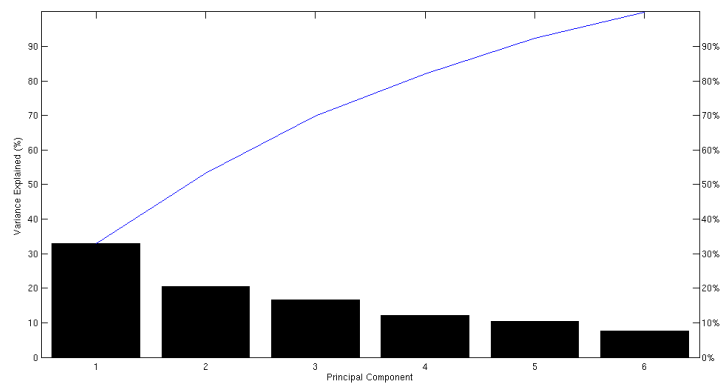


Figure 34: Decomposition based on sample GIP data (Pareto)

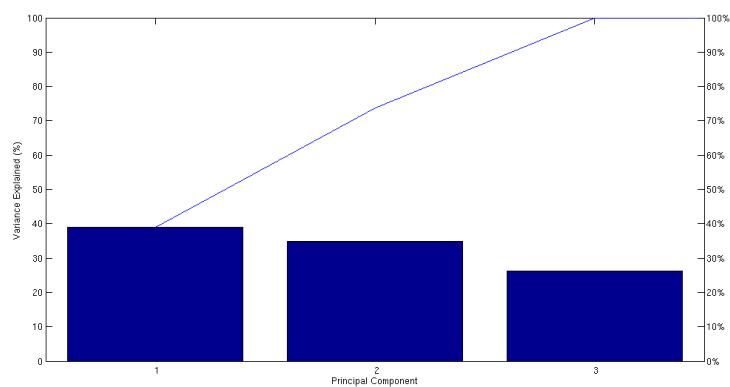


Figure 35: Decomposition based on just 4 registered faces with different expressions (Pareto)

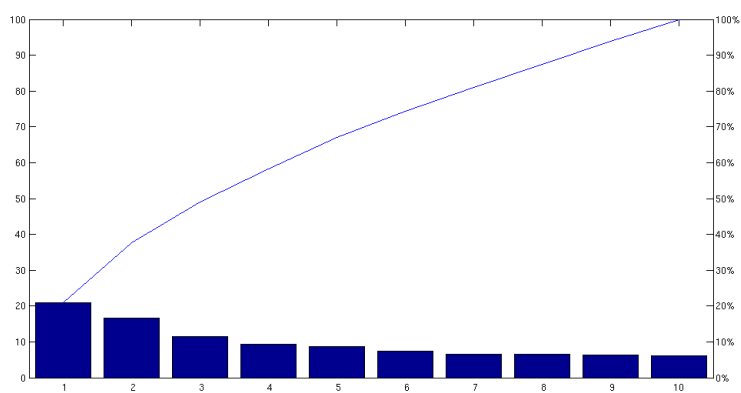


Figure 36: Same as prior figures, but with 90 images in the set

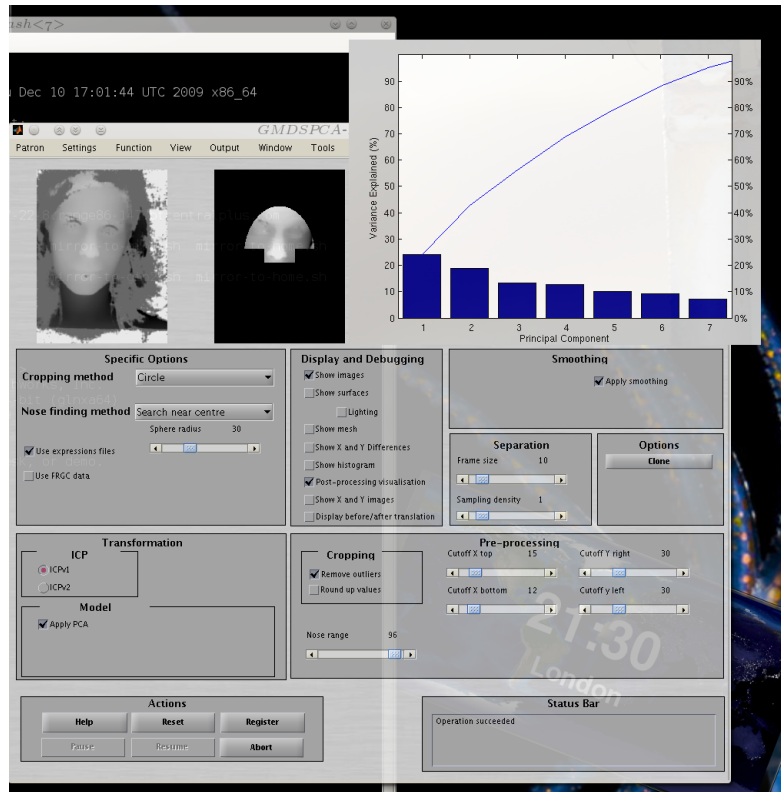


Figure 37: Overview of an experiment dealing with expression-to-expression variation model-building



Figure 38: The masks used to crop residuals in the FRGC dataset. The left-hand one is more restrictive and selective in the sense that it omits some of the data associated with the face near the edges.

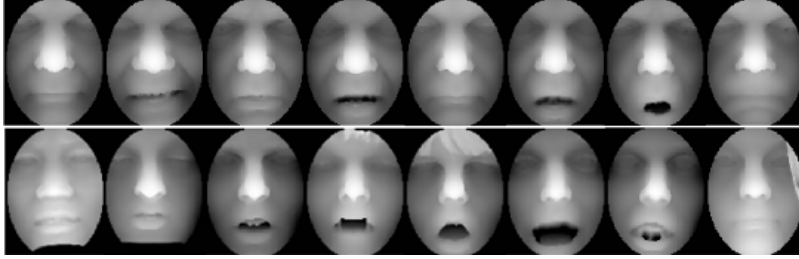


Figure 39: The top row shows images of the same subject and the bottom one is a group of hard cases (image from Huang *et al.*)

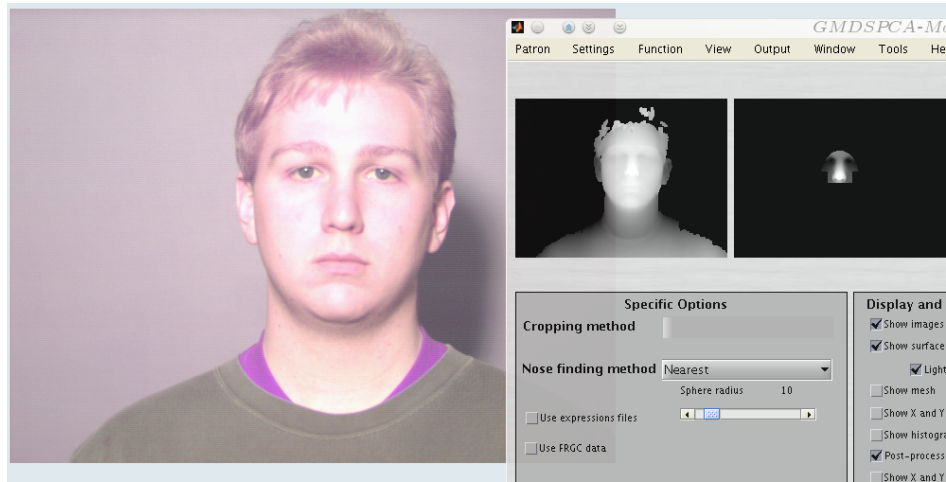


Figure 40: Left to right: Texture image mapped to 2-D, 3-D representation, and cropped parts (for alignment)



Figure 41: Example of an image where the face does not fit the image frame, unlike the example at the top right

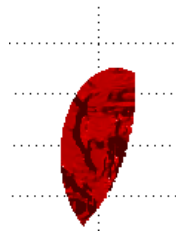


Figure 42: Image residue incorrectly cropped by a binary mask

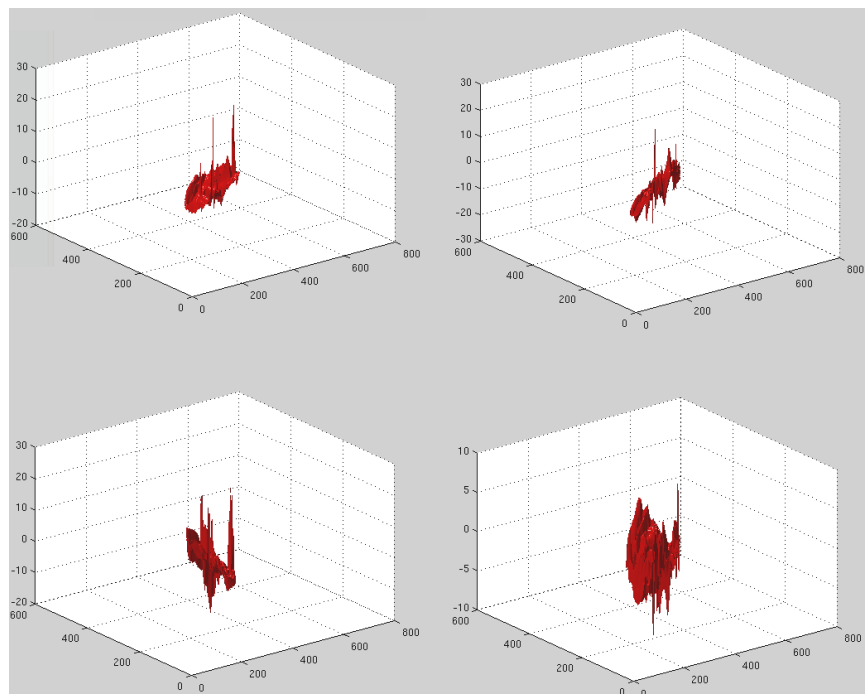


Figure 43: Examples of image residues from the FRGC datasets

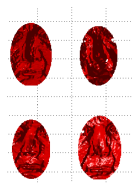


Figure 44: Image residues from the FRGC datasets shown from frontal angle

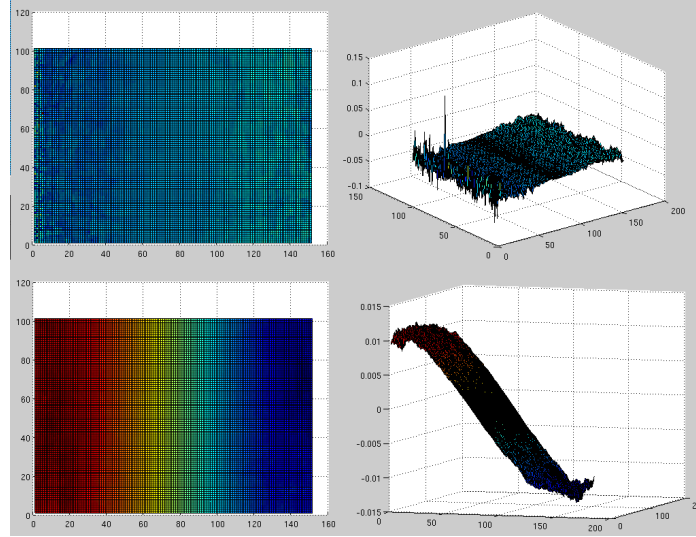


Figure 45: The image shows a matrix corresponding to two things; on the left there is a top-down view of what is shown on the right. The top part shows how the 15,000 sampled (cloud)points get distributed after dimensionality reduction and the bottom part relates to the magnitude of the principal components, where the red parts show higher deviation from the mean. It is fairly smooth.

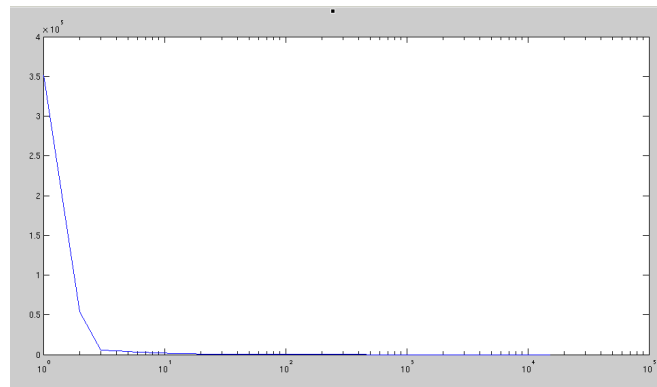


Figure 46: Principal components as a function of datapoints (log scale)

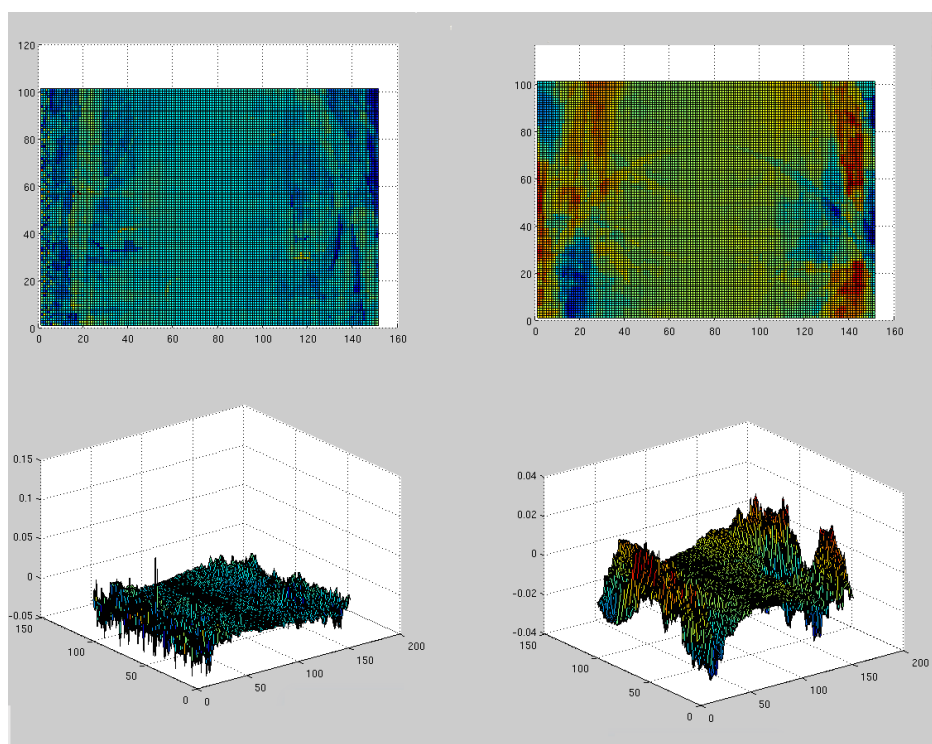


Figure 47: Point-to-point correlation along the 10th utmost principal axis

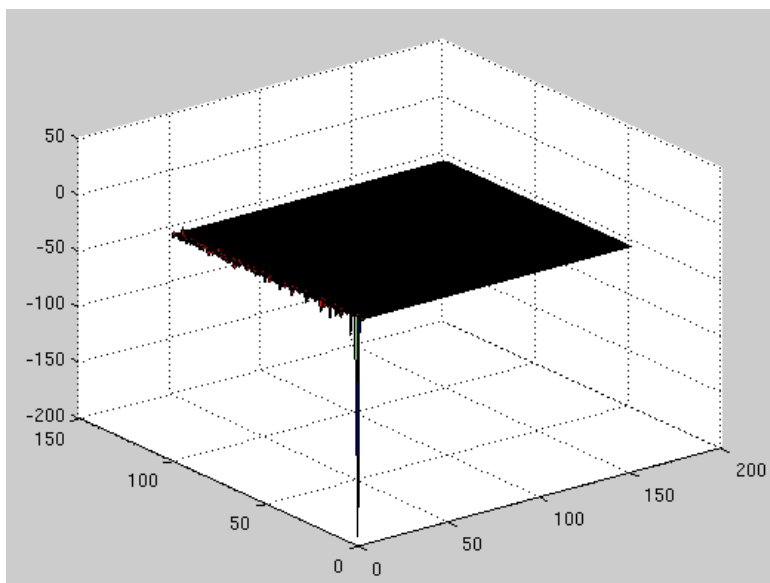


Figure 48: Score per sample point, mapped back from vectorised form to the image grid

We have asked all 3 authors of the IJCV paper if we can get access to their data in order to replicate their methods and ensure these are reproducible on the same data. I was polite and forthcoming, but I have received no response yet. Nevertheless, it is worth looking around the model for attributes which can be exploited in a novel way as a dissimilarity measure. This relates to some of the figures shown earlier. From what has been built so far (encompassing a very comprehensive set of images) it is non-trivial to identify one single area that serves as a fast-to-compute similarity measure. The implementation used in UWA is simplistic, probably for reasons of speed. But it would be valuable to reassure ourselves that it does work without any caveats. One method we now have implemented basically compares an EDM

built with the gallery and one which is built with the gallery and the probe. It is assumed that – under the model space transformation hypothesis from EDM – the observation most similar to that of the probe in fact corresponds to the same person. This can be tested as a new method that has not been explored before. It might even be publishable.

It would be preferable to use exactly the same datasets to compare results, e.g. the rather unfair GMDS vs EDM benchmarks.

With no proprietary data available from the original authors, we decided to move forward. If we can define an experiment for our lab to execute in the lab, people would be happy to contribute our faces to science.

Given large enough datasets (as are already available in sufficient quantity under gipmain), there is a lot that can be learned from systematic variation of methods, parameters, and datasets. We can, for instance, define a protocol for experimenting with PCA in a more cunning way than the group from UWA, e.g. searching along the lines of particular modes of variation or determining similarity based on the determinant of the covariance matrix. The pieces are already in place (for the most part, including multiple pluggable paradigms like Viola-Jones') and I previously published several papers that adopt this line of work at Cootes' and Taylor's group. For ICP we can promote the photometric version which getting to grips with – either in binary or interpreted form – is something we work on.

The advent of PCA is not fully exploited in previous work, so there exists

an opportunity to show work which truly builds upon prior work rather than imitate it poorly. Moreover, our methods for nose detection and ICP are more advanced. Their maturity in the literature makes them less ripe for adoption, though.

A rational experiment to perform would validate a PCA-based measure of choice and subsequently validate it using the GIP datasets. Then, the same can be done with GMDS. When validation is shown to be giving us a monotonic curve, e.g. dissimilarity as a function of the number of different people (this ought to work correctly one way or the other) in the set or level of perturbation applied to the dataset (noise, or better yet, diffeomorphic warps) we can run benchmarks on the data corresponding to Experiment 3 of FRGC 2.0. Overall, this would help demonstrate novelty in

1. Photometric ICP as applied to large datasets comprising faces
2. PCA applied to 3-D, e.g. to be used as a similarity measure and thus a classifier for face recognition of part or an objective function for group-wise 3-D registration (not just faces)
3. Comparison performance- and speed-wise between G-PCA and GMDS, hopefully showing the latter's upper hand and thus promoting the exploration of geodesics for this type of purpose.

It remains to be decided how important to us the "E" in EDM (expression) really is because the tricky, time-consuming part is separating very large sets

of images based on human-inputted classifications. UWA has used over 3,000 such images, which would weigh tens of gigabytes for standard resolutions and formats. Given that this data came from just 3 people inside their labs, it is regrettable that these "volunteers (from whom the data was collected) did not give us permission [...] distribute their 3D face data." Why spend so much time and disk space collecting data on which there is a monopoly and no chance for outside auditing? Anyway, I digress...

Shown in the images below are some examples of EDM projections. Figure 49 shows the tenth mode of variation, whereas Figure 50 shows a projection.

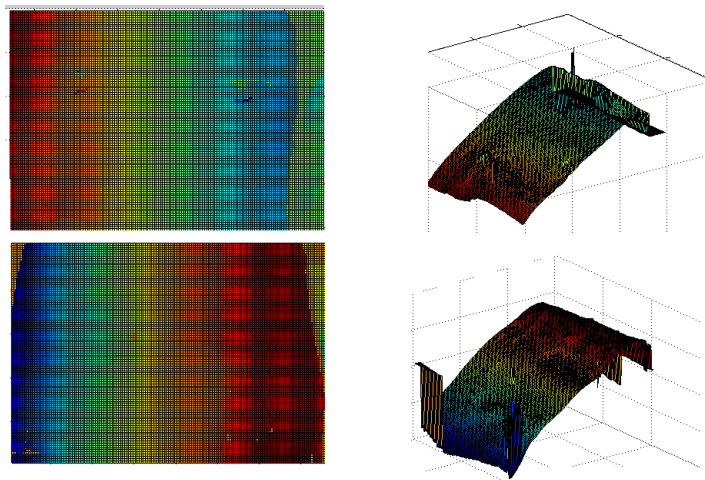


Figure 49: The tenth principal component derived from PCA, as visualised based on the reshaped (previously vectorised) image residues

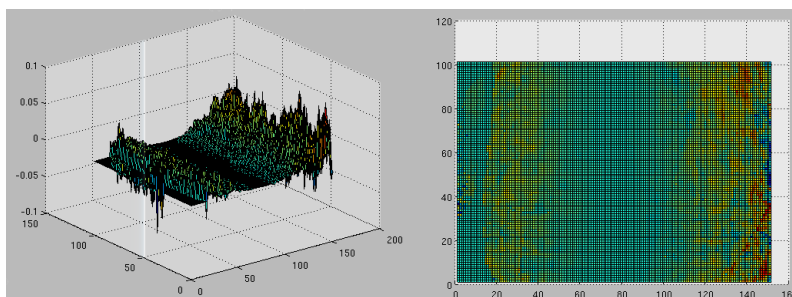


Figure 50: The tenth image residual from the Experiment 3-built EDM projected onto the EDM space’s most significant component

7.9 ICP

It ought to be possible to compare 3 ICP methods based on their groupwise performance, as judged for example by the determinant of the covariance matrix. Clustering of faces is important for Down syndrome detection, but more relevant data is needed.

Based on face statistics one is hoping to correctly detect on a Boolean basis attributes like gender, abnormality/syndrome, ethnicity, etc. There is some work on this in 2-D, where basically one can build models for groups of subject belonging to one group and then build appearance models of shape and intensity for those. The appearance/texture which includes colour makes classification simpler and automated. It performs AAM fitting to a target and scores the match.

We have begun running basic experiments which we reran using the photometric ICP methods (Figure 51), altering bits of code to make it better adapted at the interfaces level. There are at least 3 types of ICP methods at

our disposal now. It's an opportunity for benchmarks.

We have tried different parameters and options; we also attempted different methods, but the C-coded part of the program keeps crashing, even though it's run on a 64-bit machine:

```
uname -a
```

```
Linux 2.6.31-17-generic #54-Ubuntu SMP Thu Dec 10 17:01:44 UTC 2009  
x86_64 GNU/Linux
```

I took a look at the function (`ann.m`) to see if there's a non-mex option but could not see any such option. Did you build the binary on machines with the same specs?

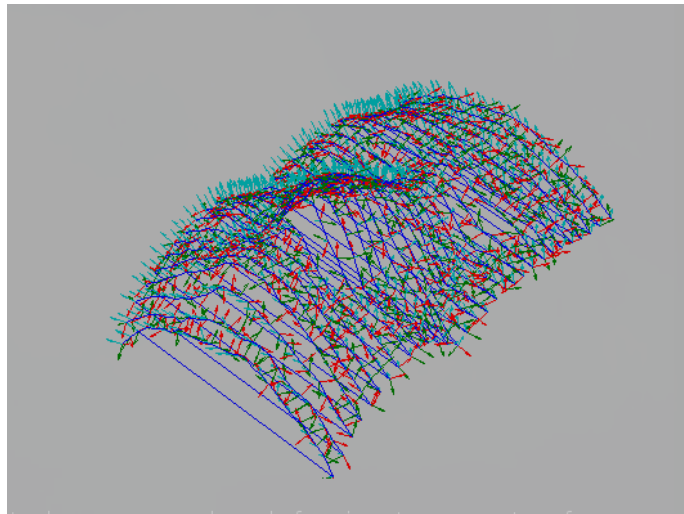


Figure 51: Example of an evaluation experiment for photometric ICP

Typically about 15,000, but spacing can be varied using a slider such that they are sampled from a lesser dense grid, e.g. 10x10 (which would reduce

this by two orders of magnitude to about 150 points). I have spent another hour or so trying all sorts of other points, but it might be a compatibility issue in the binary. It crashes everything in my session, without exception.

We still grapple with segmentation faults in my use of photometric ICP. The binary being accessed just crashes the whole MATLAB session and the shell session too become defunct (it gets intercepted as an exception at a lower level, probably for security reasons), but there may be a substitution for this binary. It ought to be possible to perform systematic ICP benchmarks on large sets once this is working, the sets being either FRGC (e.g. Experiment 3) or GIP data.

Dan wrote: “Download the ANN code from the SVN. In `buildmex.m` you will find a matlab code that does the same (for verification). Just use it instead.”

I can reproduce the error. Following recompilation of the ANN code the same crashes persisted, whereas the interpreted implementation worked as expected, by default. By changing the parameters, e.g. $k = 15000$, the same session crashes occurred:

```
ANN: ERROR----->Requesting more near neighbors
than data points<-----ERROR pure virtual method
called terminate called without an active exception

./mb: line 1: 3887 Aborted nice -n19 matlab -nodesktop
-nosplash -r "addpath(genpath('~/.pcafaces')); cd ~/.pcafaces;
```

```
system('w'); system('uptime'); gmdspca;"
```

And there it hangs.

So despite speed considerations, maybe it's worth using a different set of points or use a non-C implementation if one is available. We gave that a go. Since we did not write the ANN we do not know why it crashes on the production system. So we needed to choose one of three options: 1. write one on one's own. 2. find another one on the web. 3. use the slow/matlab version.

Tests where the function calls are made whilst increasing k from 3 to 30 and then 300 (whereupon one could reproduce the error and realise that it's an unhandled exception-raising issue) help in debugging the issue. It seems like a programming issue and not a compilation issue. Improper number of neighbours where one tries to find the nearest neighbors of 100 query points in the collection of just 10 points is when the same error (and nasty crash) can be reproduced.

We have been trying to shoehorn the photometric ICP code into the program of choice by modifying some incompatible functions.

We eventually inserted it to the SVN repository. It is possible that other functions are missing. This function is, part of an attempt to add spatially-regularized, partial weights.

We have been trying to grasp the workings of the different methods by going through the code and adjusting it slightly so as to accommodate for different

data structures. The code in its current form has been somewhat challenging to work with although it is well structured.

This module was under construction at the time – it still needs tweaking (of the parameters, to say the least), definitely for real data. We’re going to add also simpler measures of point rejection. It is satisfactory, but it’s not there yet.

The data we are working on comprises faces with a well separated (i.e depth) background.

Additional functions which were missing from the code repository are being brought together to make possible the operation of more cunning ICP code. A lot of small adjustments to the code are needed, either because some traits/attributes do not exist in the raw data or because the scale of the problem exceeds that of proof-of-concept/synthetic data/manageable scale experiments.

We have spent nearly 10 hours learning some of the code and modifying bits of it, but as a whole it is still very hard to use like a black box and simply run to get translation and rotation parameters. Some of the tests contain references to non-existing fields or functions (some might be deprecated, some belong to other parts of the SVN repository). If there exist a way of modifying the function to accept \mathbf{X} \mathbf{Y} \mathbf{Z} matrices and rigidly align two surfaces, that would be excellent. Exploration of other areas might meanwhile make more sense for progress (having recently had a similar setback when UWA said

they could not share their experimental data with a lot of expressions for testing purposes).

Geometric ICP is good enough and the photometric part would be an extra. And despite the fact that geometric should be enough, there is less novelty in it and with other directories like `icp`, `icp-quaternions`, and `icp_lihi`, these might there be worth trying too. We already have 2 ICP implementations in the program that operates on face data.

Regarding **photometric** implementation, I eventually got a fresh copy from the repo with `wget`. The `/Aux` directory was gone and I could see some duplicate functions in other ICP directories, along with nice demos that work. I tried to make use of these.

7.9.1 ICP Experiments

By exploiting more information in this problem domain we can demonstrate various things:

- ▷ ICP based on advanced geometry and richer characteristics can yield better registration performance based on the resultant model built with it. By varying parameters in ICP graphs can be produced help select better value/s for particular data of greater extent. Shown in the graph we may choose to have a level of distribution – however we may choose to approximate it – assuming quite rightly that better registration will yield more concise descriptions (Occam’s razor principle).

- ▷ A trickier thing to do, either for technical reasons or for purely computational limitations, is to use models as a similarity measure in an objective function for face analysis. This can be tested on coarser representations of faces, perhaps icon-sizes ones at a resolution far lower than the original.
- ▷ Expression recognition or expression-agnostic face recognition can be done using the above tools, which generally require further refinement. Data for this is already available. However, the exact method of choice for similarity must be strictly defined and tested systematically for compelling validation.

By making alterations and putting them back together into the code it was made possible to run several older variants of ICP algorithms, incorporating them into the pipeline of the program. Older implementations (even yours from 2008) can now be compared based on face data.

Their assessment is to be done with PCA that estimates complexity; the drawback of this approach, however, is that it becomes slow when the dataset is large. In the past, sets as small as 10 could be sufficient for an objective function in non-rigid registration. Figures 52 and 96 show the type of data we deal with.

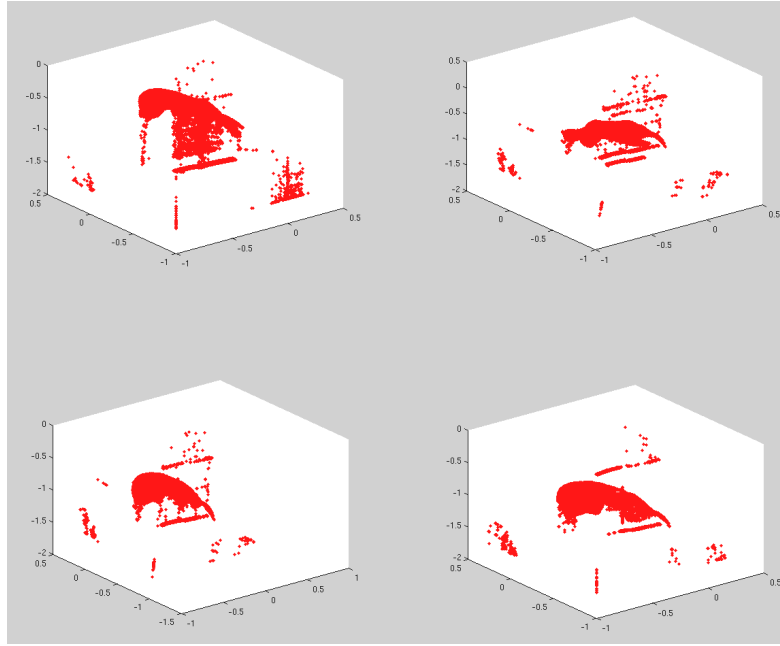


Figure 52: ExamExample points cloud for ICP to register

7.10 Systematic Experiments

With the goal of validating and comparing face recognition methods, we can embark on the following path of exploration. The data to be used needs to be of different individuals and the datasets must be large enough to enable model-building tasks. As such, the data specified in Experiment 3 of FRGC 2.0 should be used for both training and testing. It needs to be manually classified, however, as groups that previously did this have not shared such metadata. It would be handy to select hundreds of faces that represent expressions like a smile and then put them in respective loader files (manual work), alongside an accompanying neutral (no expression) image. It ought

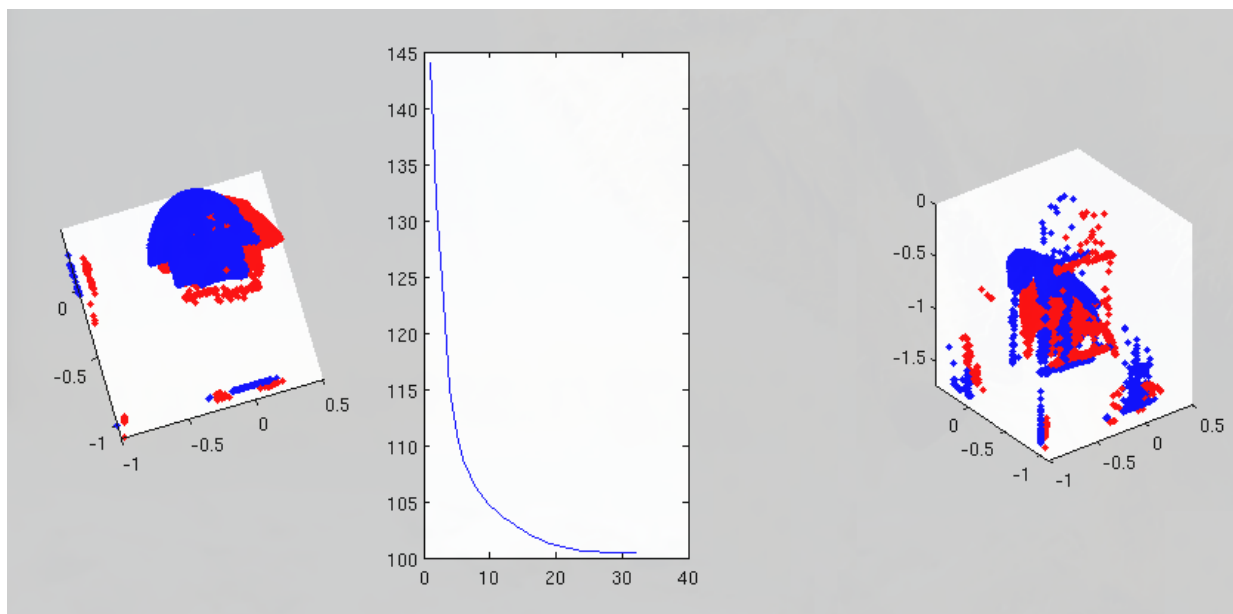


Figure 53: On the left: two faces (with binary masks cropping them for rigid parts like nose and forehead) overlaid for ICP; on the right: same from another angle

to be possible to set aside 200 such pairs, all coming from different people. Identification in such a set ought to be quite challenging, without texture (which is in principle available in separate PPM files).

The experiments can have the set of 200 pairs further split into smaller groups for repetition that takes statistics into account and can yield error bars. Dividing into 5 groups of 40 pairs is one possibility, even though a set of 40 individuals is becoming a tad small. In order to train a model of expressions it ought to be possible to just use the full set.

When approaching this problem the goal would be to pair a person with an expression to the same person without the expression (or vice versa), attain-

ing some sort of gauge of expression-resistant recognition. The gallery is the set of all faces in the set. Similarity measures being pitted for comparison here can include the 4 ICP methods we have implemented, plus variants of these and different selection of parameters. Different measures resulting from ICP and the region being compared (e.g. all face versus nose, versus forehead and nose) are another area of exploration. There ought to be separation between the idea of cropping for alignment alone and the strategy of cropping or using binary masks for the sake of computing difference as well.

What we may find is, by cropping out some parts of the face, recognition will improve considerably. But in order to take the deformable parts that change due to expression into account, something like an expression becomes necessary. Then, there is room for comparison between expression-invariant model-based recognition and recognition which is based purely on alignment. The type of alignment too, e.g. the implementation of ICP, can be compared in this way,

To summarise this more formally, we take $N=200$ pairs of size 480×640 , where all of them are 3-D images acquired from N different subjects under various lighting, pose, and scale conditions, then register them using 4 ICP methods, in turn (potentially with variants, time permitting), using a fixed nose-finding method. As the first experiment we may wish to apply this alignment to a set of cropped faces, ensuring that they all lie in the same frame of reference. A model is built from the residual of all 200 pairs, in order to encompass the difference incurred by an expression of choice, e.g. smile or frown. In

the next stage, 5 sets of $M=N/5$ images are set as a gallery G and a probe p goes through all images in G , attempting to find the best match best on several criteria such as model determinant or sum of differences. This is how it is implemented at the moment. To measure determinant difference it is possible to add the new residual (between p and any image in G), then concatenate it to the set of observations that build the model, rebuilding it rapidly (coarse-to-fine approach if needed). This is how it is implemented at the moment. Subsequent experiments can extend to compare other aspects of recognition using the same framework/pipeline. GMDS can also be added for comparison with G-PCA. Measurement of performance should be easy if the correct matches are recorded for a random permutation of the set and then paired for some threshold (or best match) based on the gallery. The most time-consuming task is organising the data for this set of experiments. That may sound plausible enough as a starting point.

===

If the experiments work as hoped, performing them on larger sets ought to be trivial. To proactively remove allegations of the set being too easy to deal with (picky-ness in peer review), the most difficult partition when it comes to acquisition quality is taken. Figure 55 shows some examples of pairs that are being used after being selected as not many images contain expression variation. The selection process of very tedious as very few 3-D images exist with expressions in them, especially ones from the same person (required for consistent training assuming intra-subject residues are alike for common expressions).

About 5 hours were spent classifying the NIST datasets for future experiments. An initial subset of it is put in loader files. From the whole 3-D data of the Face Recognition Grand Challenge, one can only find a few hundreds of distinct individuals. Not all of them have an acquisition with a smile. I found just over 80 by manually browsing everything and some will be hard to work with due to obvious cases of degraded signal. The criteria was that all parts of the face (mouth upwards) must be visible and the expression one of happiness, not necessarily a smile.

The new methods of ICP are applied to target data such as the above. The program works reasonably well (see Figure 55) with GIP implementations of ICP (there are two main ones from GIP) and the new data which comprises 86 pairs, or 172 images in total.



Figure 54: Examples of the faces used for training and recognition, with neutrals on the left and smiles on the right

We have gone through nearly the entire group of images for the purpose of reassurance, ensuring not so systematically that without any intervention or modification the face is detected and brought into alignment, first by segmenting its parts and then by translating it to better fit the pairing (a companion image which is non-neutral). So far we have found just one problematic image, where some guy's broad locks are mistaken for the parts of his face. This will need to be corrected somehow, without treating it as a special case. It is actually surprising, given the variation inherent in this set, that the vast majority of images will be detected so easily.

If we need a single or a couple of people with various expressions, that we define somehow, it will be possible to generate in the lab. We could define the required expressions by pictures.

We already have a group of a dozen expressions, all acquired from the same young lady. In order to perform a comparison where recognition is not a binary selection problem (UWA had three subjects in their sets) we thought it would be preferable to try this algorithm on a cluster that builds a model and then also manifests galleries from the training sets. If a model-based approach can be shown to be superior to a purely geometric approach, it would concur with some previous work I did on human brain in 2- and 3-D.

The use of broad galleries of expressions taken from smaller groups would be immensely useful for experiments that check our ability to detect expressions rather than detect the person, bar expression. The paper from UWA dealt

with the latter problem, wherein they present a probe with a set of possible matches and by mitigating expression differences, metrics like point-to-point distances become more meaningful (resistant to elastic deformations).

Shown in Figure 56 is the difference between the images from our expressions set and resultant corresponding images following ICP-found translation. This uses a 2008 implementation of ICP even though we have a newer one which works.

Several isolated images that only account for about 1% of those selected from admittedly difficult sets cannot be deal with, at least not short of major improvements to the algorithm, which then jeopardises handling of all the other images. To avoid falling into such a cycle of refinement/overhaul which is set-optimised, it is reasonable to consider cull-out. Images such as the ones shown in Figure 57 (the only problematic ones found thus far) pose too much of an issue to be usable due to the hair, which stands out and fits within the frame looking for parts of the face. Therefore, to simplify the experiments, these images get dropped. It still leaves faces from over 80 different people (distinct anatomical characteristics, scale, gender, and so on).

To give examples of some of the faces we deal with (and the algorithm deals with painlessly), see Figure 58 which merely shows the first 6 rather than cherry-pick good examples (everything in the current set is handled correctly).

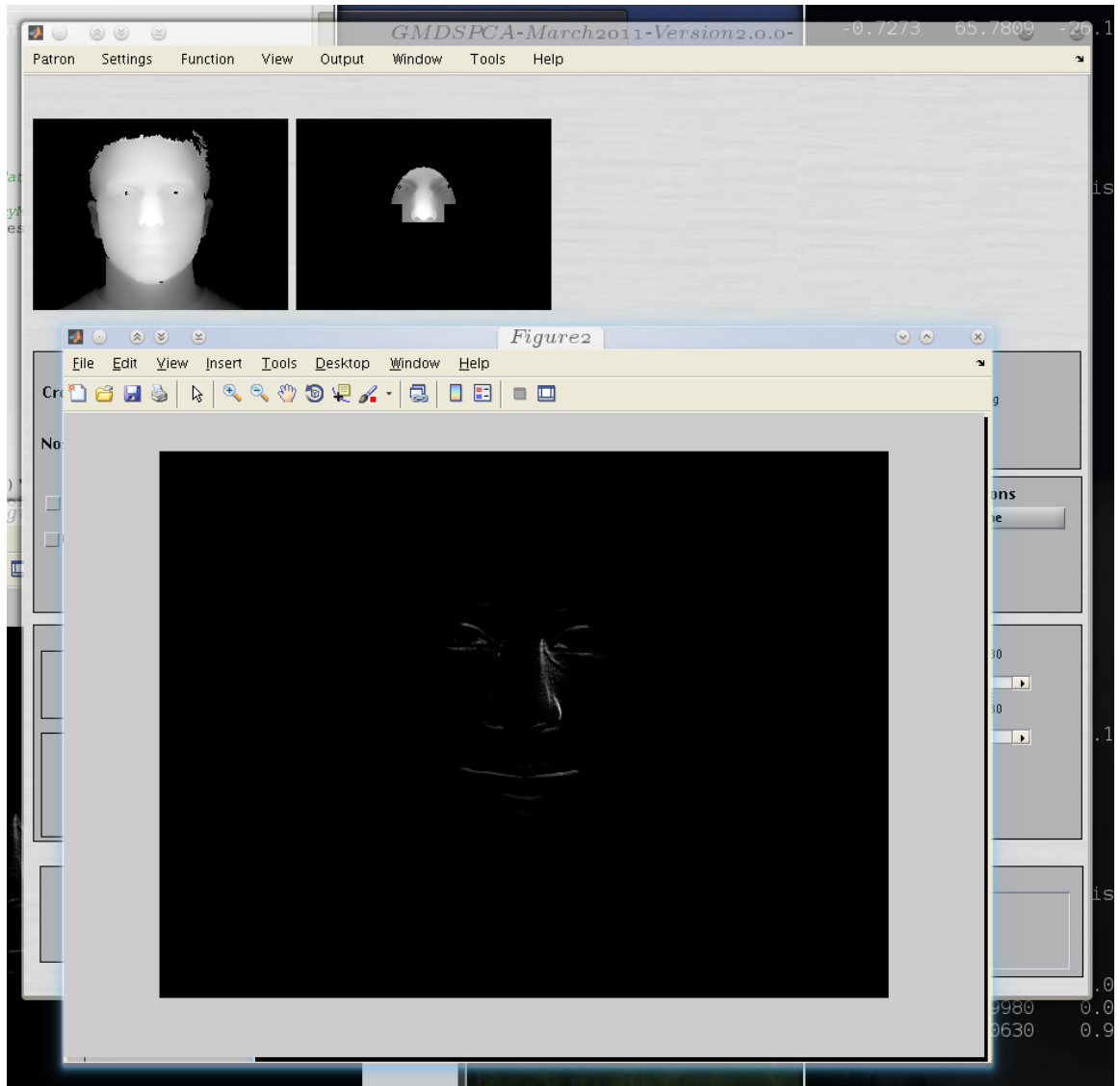


Figure 55: Examples of the program with the new data and methods in place

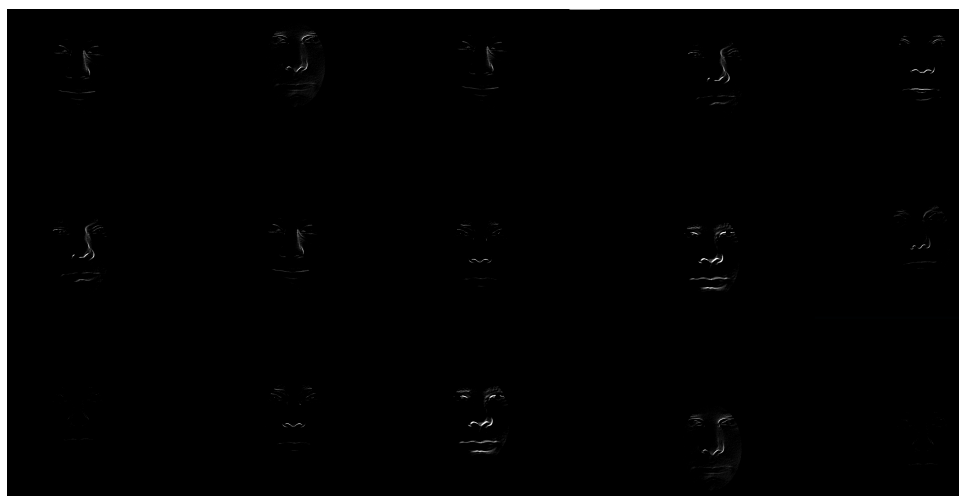


Figure 56: Pixel-wise difference between the images from our expressions set and resultant corresponding images following ICP-found translation

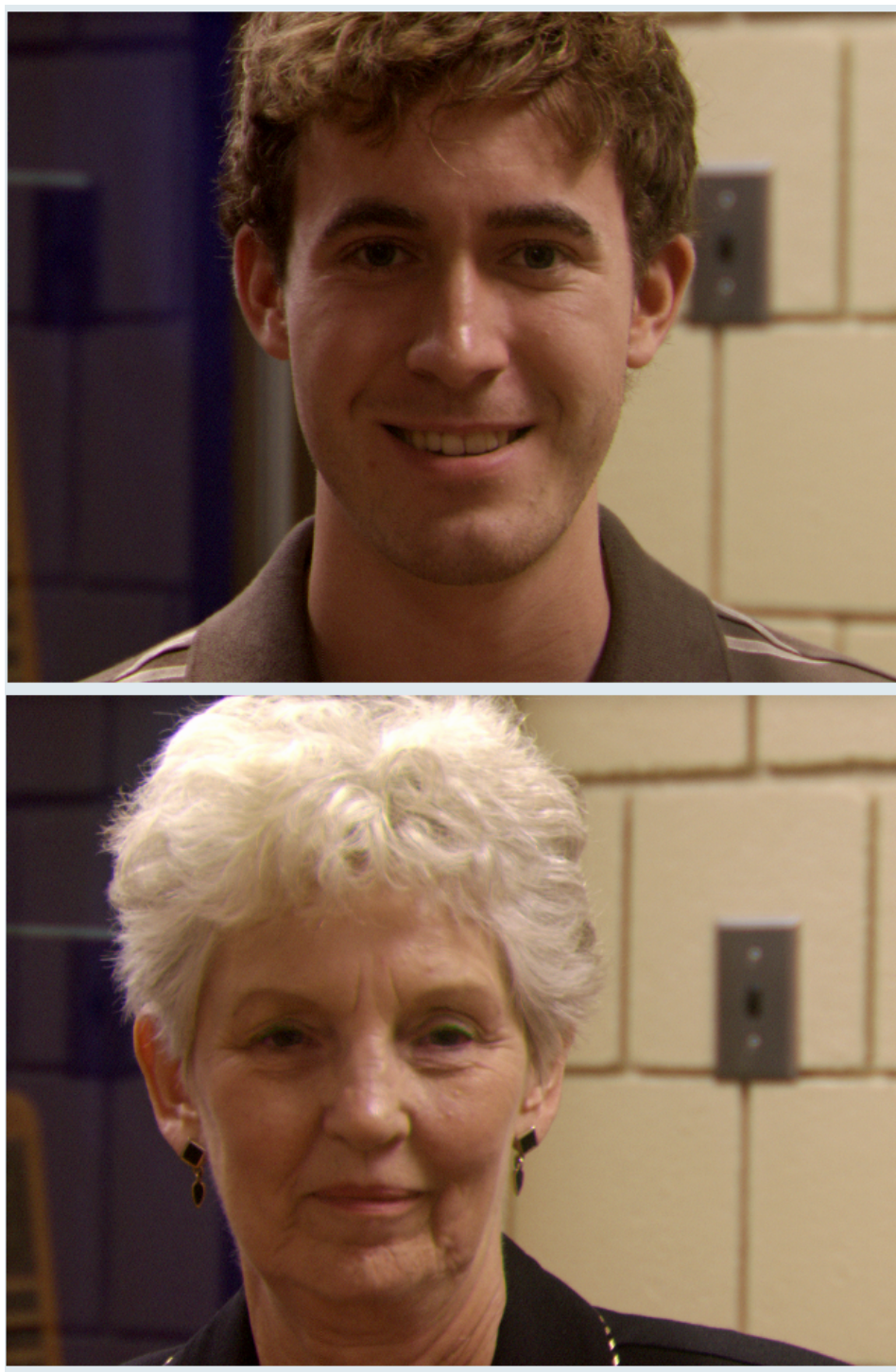


Figure 57: The two images that automatic detection struggles with (because of the hair)

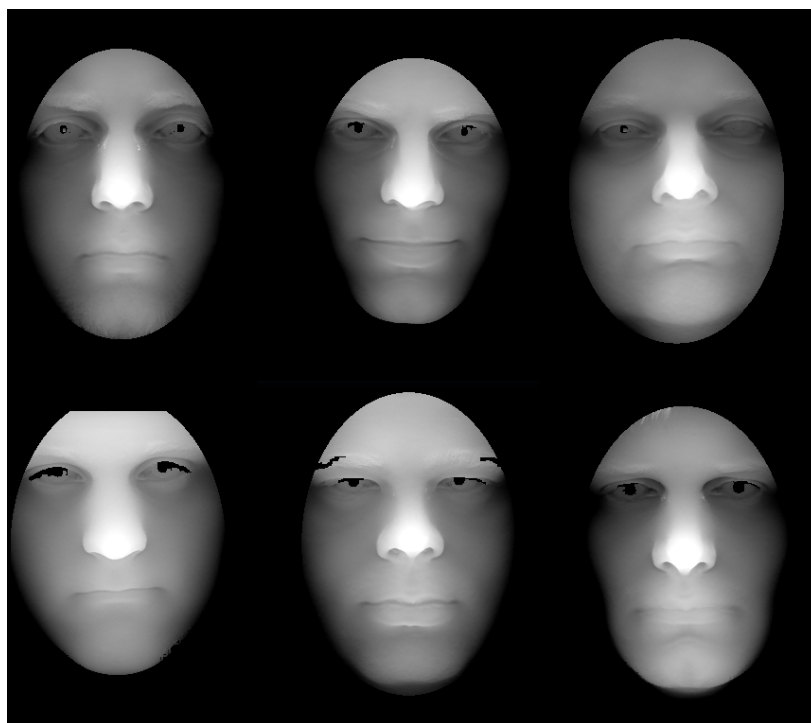


Figure 58: The first six neutral images taken from the set and cropped by the algorithm correctly

7.10.1 Residue Adjustments

The set is reasonably well handled following the removal of two images from almost 200 images in total, leaving a group of 86 distinct people. Shown in Figure 59 and Figure 60 are two masks that were tested on the group. Shown are just the first few images, not a bunch of cherry-picked examples. Among the 4 binary masks that are used (with different thresholds for depth as well), the latter works better and it is aided by cropping that more or less normalises the region under consideration, making it easier to sample and subsequently compare. Worth noting are the difference around the mouth, which reveal some teeth.

By profiling people's expression residues and then testing to see if these profiles – be they based on a model or not – can be used to detect the identity of the person, we can reason about the approach and compare pertinent, exchangeable parts, swapping them and assessing the effect on overall performance. First, something more basic like sum-of-squared-differences will be tested as a differentiator (for match/target).

7.10.2 ROC Curves

At this stage we are able to form many kinds of benchmarks (ROC curves) on some of the data sets. We already get some numbers, but to get good numbers and organise them in ROC curves we need to finalise the protocols of dividing up the sets. In order to get ROC curves to be tested ASAP only

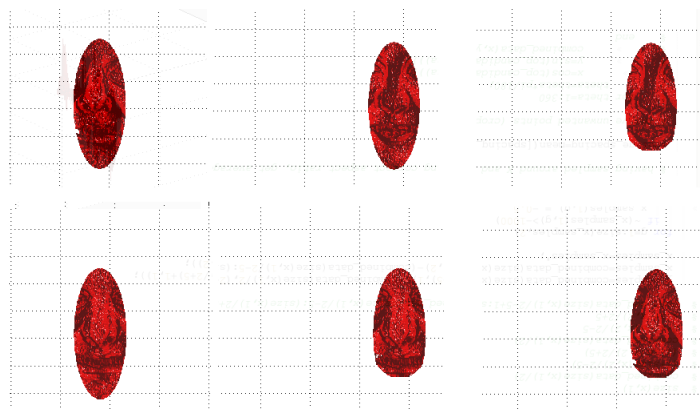


Figure 59: The first 6 images in the set with a narrow mask used to extract and attain a neutral-to-non-neutral residue

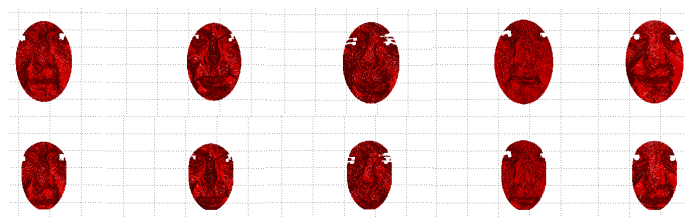


Figure 60: Same as the previous figure, but with only 5 images. The top row shows the effect of using a broader mask and the bottom part shows the effect of applying a fixed mask and thresholds to make the data more trivially comparable.

X data was used, which is not terribly useful as X contains little signal in general (low entropy, too). All the preliminary results will therefore be more like a proof of concept.

The careful arrangement of sets will be necessary to ensure that many tests without too much overlap or repetition can be enrolled and used as our "standard". The set of 86 distinct people should be partitioned sensibly. In order to test this and show the results are reasonable for a test set of

just 13 faces, Figure 61 displays the ROC curves acquired based on mean of differences (one of several similarity measures). We will need better ones, preferably with comparators too (overlaid curves for human judgment).

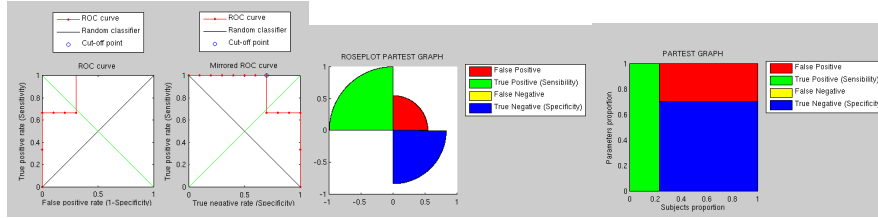


Figure 61: ROC curves plotted for just 13 tests done on the FRGC datasets with expressions isolated

It is usually worth checking if other groups or even people who work in the same lab have pre-partitioned and classified data (as per individuals). That would save the researcher the hassle of doing it manually. Just picking out expression took nearly 5 hours. The larger the dataset, the smoother the ROC curves will be, obviously.

Classification by hand takes a while, but it is crucial for results. The work done for NIST should have it for the training. The training partition contains nothing with expression variation, however, so we classify the image already isolated for the task of expression removal, comparing an approach that does not annual expression against a similar one that does.

Basic experiments were soon followed. In this very preliminary test we are dealing with a rather difficult set, using different acquisition conditions and different expressions from many people. We focus on dealing with just rigid registration (GIP latest ICP implementation) and simple metrics. ROC

curves are plotted in accordance with the data gathered from 50 examples (see Figure 62).

Next, we intend to improve the results with more cunning registration, annulment of facial expressions (e.g. the EDM approach), and most importantly improved algorithms for masking and aligning image parts, then measuring more meaningful properties in them.

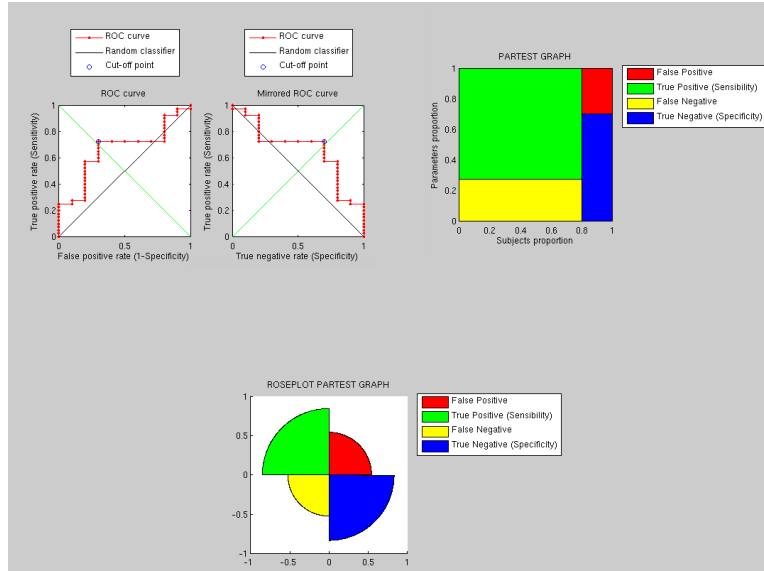


Figure 62: Preliminary test where several images (not complete set) are used to get a rough idea of what the ROC curves will look like

With a much larger sample set which includes all the neutral-to-non-neutral pairings I ran the same experiment, this time using an older ICP, which uses PCA, to plot the ROC curves (see Figure 63). ICP is only used for translation in this case. There is plenty of room for improvement and it should not be hard to get that improvement shortly. This has been an exercise in just

testing the foundations of the framework, which now streamlines a lot better.

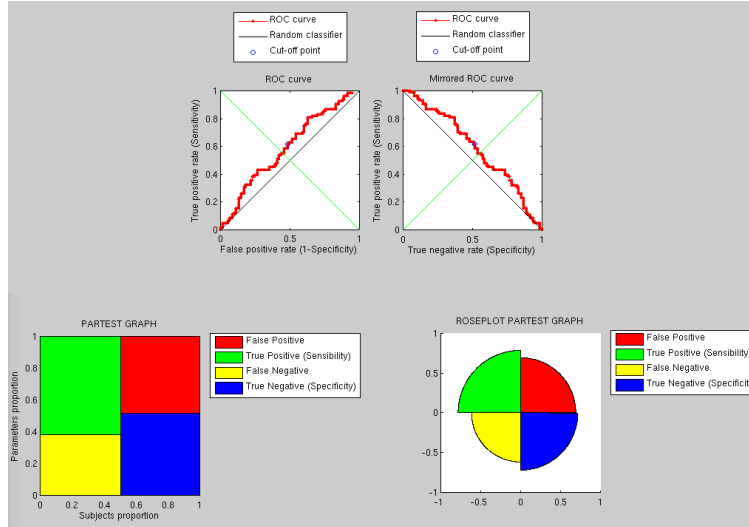


Figure 63: A somewhat larger test on non-neutral sets, where ICP based on PCA is used for alignment, then mean of residuals get used as a similarity measure

In Figure 64 is the same type of curve for a method which was made more robust to noise and sensitive to differences.

Comparisons have so far involved just the X/horizontal axis data (see Figure 65 for X, Y, and Z data overlaid), which was not particularly useful for telling people apart. It was intended to test and explore some new code. A median-based method with squared differences taken into account is now put in place and it uses actual depth (Z alone used as signal/data) to perform tests on neutral and non-neutral images, as before. The results are, as expected, far better than before. Figure 66 shows the 5 first matches that are correct and Figure 67 shows the first 12 that are not correct (belonging to different

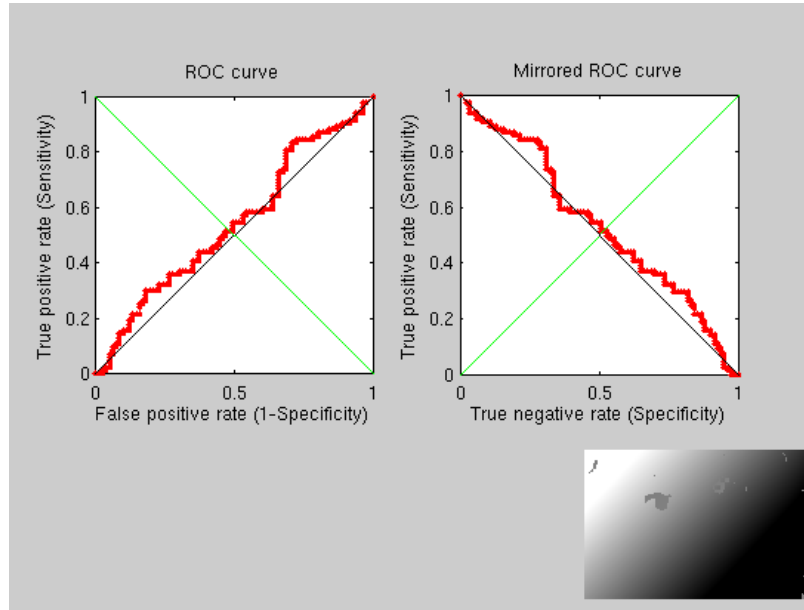


Figure 64: The same type of comparison with the same type of set (as in Figure 63) but with a more cunning similarity measure and an example of the X data at the bottom right

people). Figure 68 shows the classification of those 17 images, which are simply the first ones in the test set (no selection bias). The small scale of this experiment is intended to help track, on an image-by-image basis, what it going on. Larger experiments will follow.

Next, model-based approach will be incorporated and then benchmarked against others, notably counterparts that do not take advantage of statistical expression annulment.

Figure 69 shows the same ROC curve extended to account for a lot more image pairs (for which there is no accompanying matrix representing the contribution of each, as before). Comparative curves should be trivial to

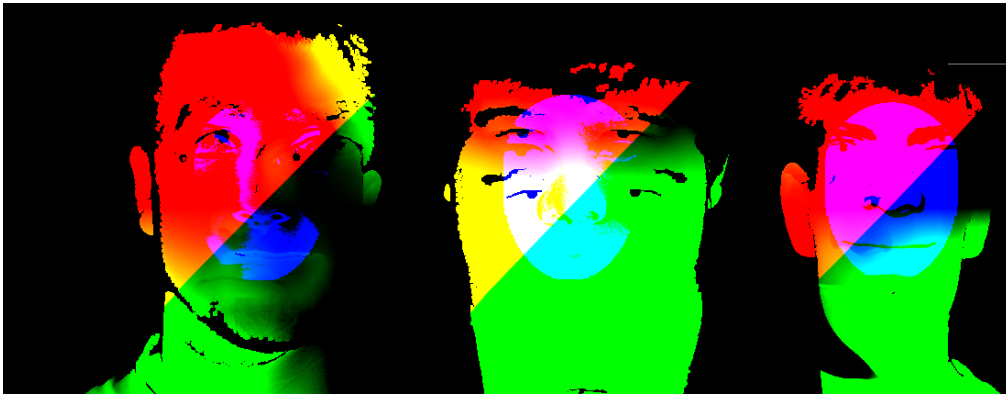


Figure 65: A combined view of X, Y, Z, the image before ICP, after ICP, and the reference image

produce.

7.10.3 Initial ROC-based Benchmarks

Results have changed somewhat after making big changes to the code, mostly in order to improve performance and also address some errors. Bugs were introduced as part of these changes, leading to a slow debugging process and some basic assessment stages that helped guide development. It's cleaner now and it contains more modes of exploration.

Reasons for lower performance than what is possible include a need for improvement in location, addressing for example the almost problematic pair in (see Figure 70) To test performance in a quick way, half the set (first half) was used to yield a ROC curve, or two as shown in Figure 72. Shown with diamonds as markers are the older results and the matrix of many images (Figure 71) shows the type of masks being used to to classify unseen

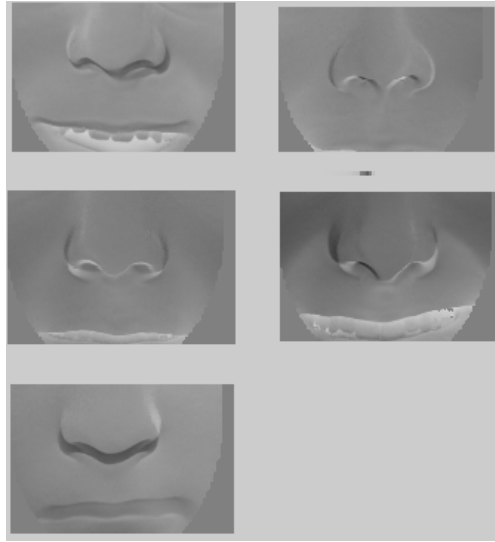


Figure 66: Difference images of the first 5 pairs taken from the same people non-neutral images (hardest task).

The next comparisons will be more interesting as they will involve different strategies. The aim is to measure expressions-resistant properties using eigenvectors or geodesic distances. The harder the test set, the more profound the performance advantage will seem.

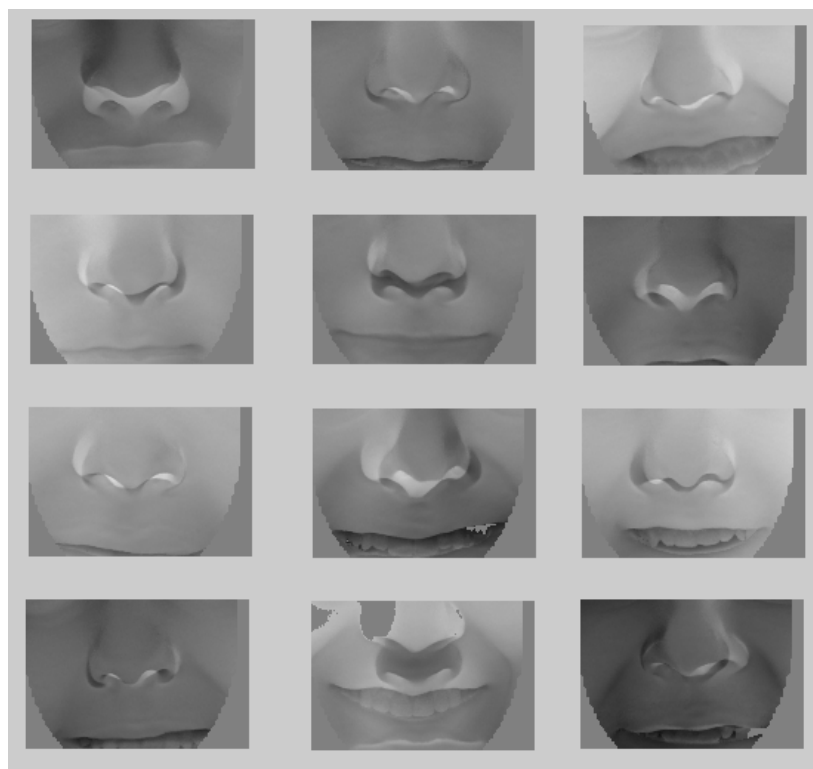


Figure 67: Difference images of the first 12 pairs taken from different people

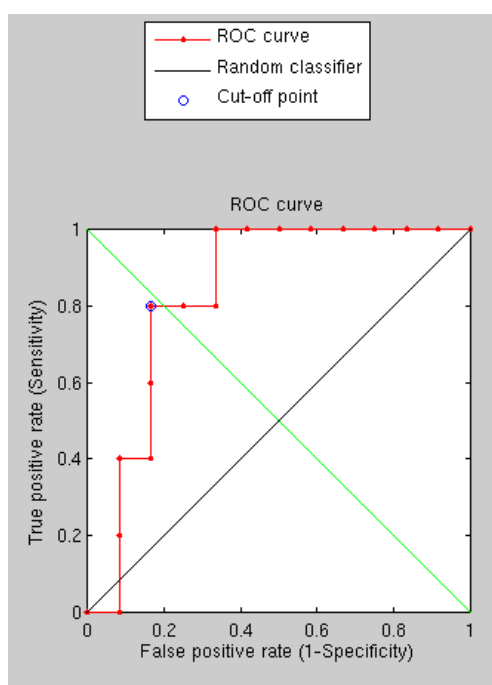


Figure 68: ROC curve of the 17 images from figures 66 and 67

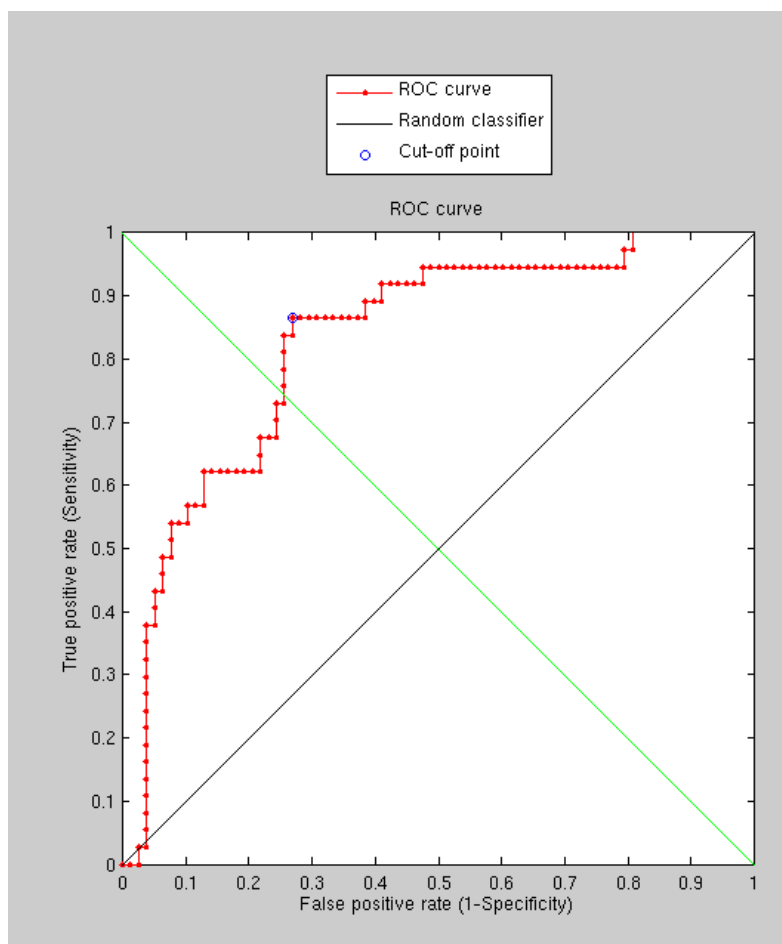


Figure 69: The curve showing the performance for 83 pairs from false matches and 37 from true matches



Figure 70: Images `~/NIST/FRGC-2.0-dist/nd1/Fall2003range/04557d337.abs` and `~/NIST/FRGC-2.0-dist/nd1/Fall2003range/04557d339.abs`, where there is some detection difficulty

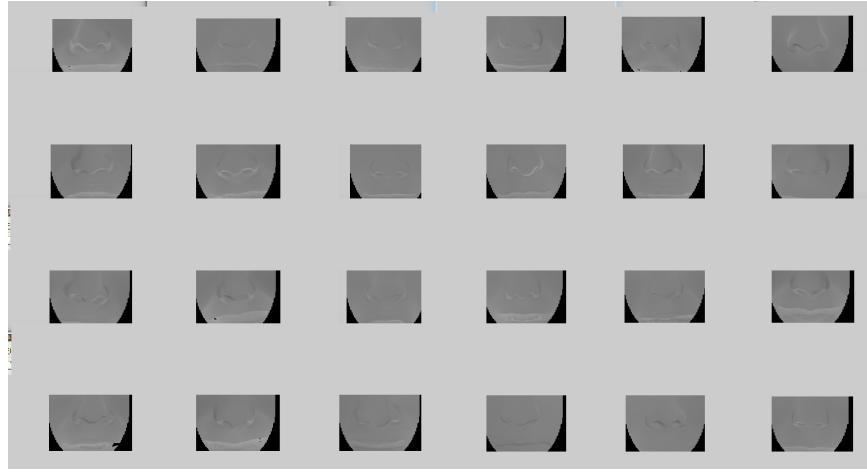


Figure 71: Example face-to-face comparisons

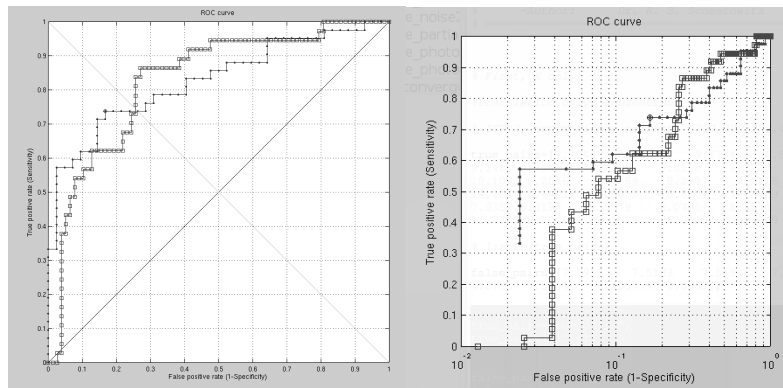


Figure 72: The ROC curves comparison on the left as linear scale and on the right log-scaled

With a broader facial range of view (bigger face-imposed mask, display of residues and partial image selection), smoothing significantly increased, the use of GIP's geometric ICP, and after bug removal (ICP totally disabled for testing purposes as well), median of quadratic differences was replaced by average of quadratic differences, we have rerun some experiments (the results

can be seen in Figure 73) and spent 3 hours (in vain) trying to build a model from the whole set. When it came to PCA, the program just took over 4 GB of RAM (including swap) and never completed the operation. It hanged for 6 hours, so this needed to be aborted. The GIP dataset comprising smiles from one person (young female) could be used instead, however the image dimensions and the nature of the images is slightly different there. Treating these two sets interchangeably would not be so trivial. For this set where all the pairs comprise one neutral and one non-neutral, the absolute differences are not so meaningful, as expected. But the removal of expression very much depends on the quality of the model and the recipe for building it counts a lot. It seems as though MATLAB exceeds some memory thresholds even with 166 images where the points are densely sampled. This necessitates a redesign. For testing purposes we will start down-sampling the images by sampling at equally spaced points on a grid. This can speed up experiments and when everything works satisfactorily, every component in the pipeline can be scaled up again, maybe even applied in a multi-resolution-type approach, as done with Active Appearance Models (AAMs) for performance gains.

7.10.4 Downsampled Images for PCA

The images were downsampled by a factor of 10 along each dimension, lowering by two orders of magnitude the Z axis data that gets sampled by PCA based on a grid. This ought to keep the models more manageable for the purpose of algorithm/performance testing. Interestingly enough, downsam-

pling hardly affected the ability to recognise faces. As Figure 74 shows, the classification remained almost the same, even though the images were tiny (see Figure 74 at the top left).

The differences are very minor since the current performances are not within the expected (state of the art) results. Much better results could be expected (if, for example, at 0.01 false positive rate we get about 0.6 true positive). With GMDS, for example Bar Shalem got about 0.9++ for sampled NIST database.

It would be surprising if our current recognition rate was high because the points compared are sampled around the lower part of the face and all of them are neutral-to-non-neutral (or vice versa), so the mouths move. There is nothing to annul this at present. Next, we a model of fine-scale image equivalents and will have some ROC curves. We deliberately chose the most difficult tests (not working in the high 90s/99th percentile).

7.10.5 Model-based approach

Using the same data and preprocessing as before (for the sake of a sound comparison), we have applied a PCA-based approach to get the following results, which are clearly by far superior. The variation incurred by expression is detected by PCA in the sense that it is not seen as a new type of variation.

The performance, as shown in Figure 75, is therefore greatly improved and there is room for further improvements as this implementation uses tiny

images to save time and it does not use the sophisticated approaches partly implemented by now. We need to explore and compare breeds and variants of the same methodology, which is easier to get to grips with when the matrices are of scale that can be viewed and understood by a human (breakdown of the modes of variation is shown at the top-right corner). Caveats can summarised as follows:

- ▷ the statistical model is built from a small set of images and therefore it cannot capture much of the variation
- ▷ we are dealing with the most difficult subset from the NIST/FRGC-provided dataset, which helps tell apart poor method from good ones (without looking at fractions)
- ▷ the granularity of the images is low (for testing purposes)
- ▷ the approach is simplistic as it is intended to be exploratory

After some exploration around separation, a mean in the measure was replaced by median of quadratic changes (in the modes of variation). The cutoff point was subjected to exacerbation though (see Figure 76). To really improve performance we must address the real caveats as well.

As a more novel experiment, we could use an approach for expression classification (possibly with the GIP dataset assigned for training).

We are still improving performance, despite all the caveats that remain inherent. Improving performance by exploring different similarity measure helped

yield Figure 77 and Figure 78.

With the GIP data given us from the lab, it ought to be possible to perform expression classification based on a set of expression models. It should, in principle, be easy to build a model for each expression and then test our ability to classify an unseen picture/3-D image for the expression embodied and shown by it. In order to distinguish our work from that of UWA (which I firmly believe took some shortcuts), a classification benchmark¹⁶ would be worth pursuing. FRGC 2.0 data can be used for multi-person validation of the approach, preceding other experiments in a publishable paper. The face recognition problem seem to be a crowded space and the EDM approach is good for automatically tackling expression variations, via variation decomposition. Alternatively, performance on par with whatever is in the literature can be pursued, only with the goal of showing an EDM-based approach to be inferior to another. The image from Figure 79 shows handling of unseen images by an expression model.

In general, once we get a good enough recognition rate (say within range of the Mian *et al.*) for the FRGC data, we'll have to introduce a more generalised way of compensating for expressions. One would say GMDS, but again, other G-PCA approaches are possible.

In its present state, with data mostly consisting expression variation, pro-

¹⁶To grossly define a classification benchmark in this context, it ought to be possible to model different 'families' of variation (such as anger, fear, etc.) and then classify an unseen image based on model fit. It would seem quite novel and it probably ought to work.

gram performance depends overwhelmingly on the ability to recognise and eliminate/blur out/cancel the expressions contribution. Model description length can be used to determine how much of the variation is due to expression change. Empirically, so far there are signs that it is working, however there might be other explanations for it. With noses superimposed and faces generally facing the camera, ICP does not play a major role. It's all about the handling of expression changes. Suffice to say, by just detecting rigid areas one could calculate a lot of attributes that identify an individual. Then there is texture, which can further validate it although we do not use texture at all (so far).

It would be nice to draw a link between GMDS and G-PCA. In fact, there is a nice way to link the two theoretically.

There have been some relevant talks available for viewing recently (over the Internet). So far, PCA seems to be serving primarily as a similarity measure with respect to entire sets of observations, where a given image gets compared – via residuals – to a set of other images of its kind. Surely there exist better uses for PCA; in the context of this work it gets used as a throwaway tool for measuring similarity, which utilises little of the information conveyed in the learning process.

Manual markup of data or classification/pairing based on common properties (by hand) is extremely time consuming, especially if large sets become part of the protocol. In order to test on pairs excluded from the training process and

then compare them to non-correspondent pairs (with or without expression), an experiment was designed to take 1,000 random pairs and compare them to: 1) unseen pairs with expression (figures below) pairs from similar sets of people (figures below). The results do show the ability to distinguish, but for more impressive results we will need to address existing caveats, which include the number of images building the mode and their size, among other important factors. It will take more time. The University of Houston did not respond to request for such data.

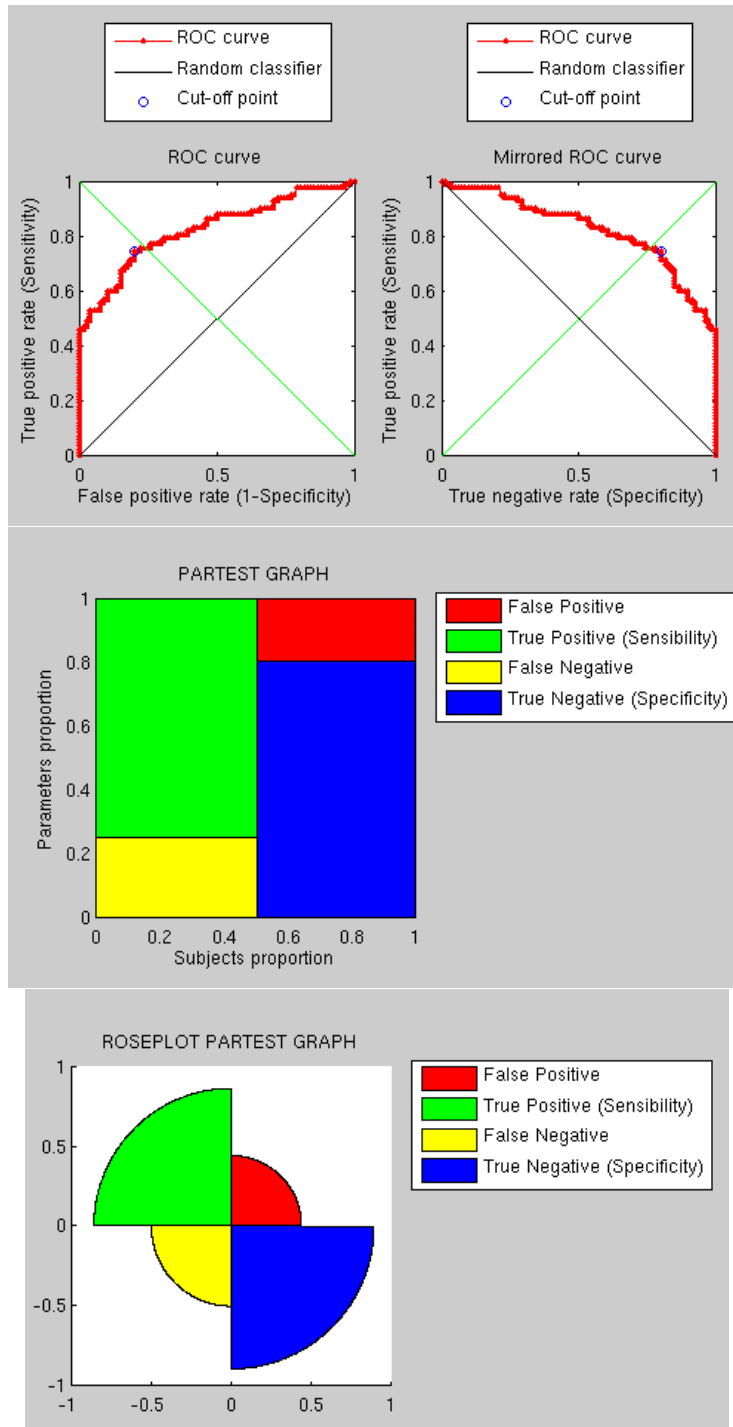


Figure 73: The results in terms of recognition rate after widening the mask and also changing from median to mean

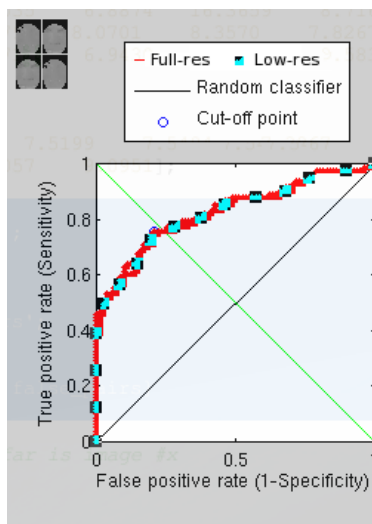


Figure 74: The results from full-resolution image sets and low-resolution equivalents (as seen at the top left)

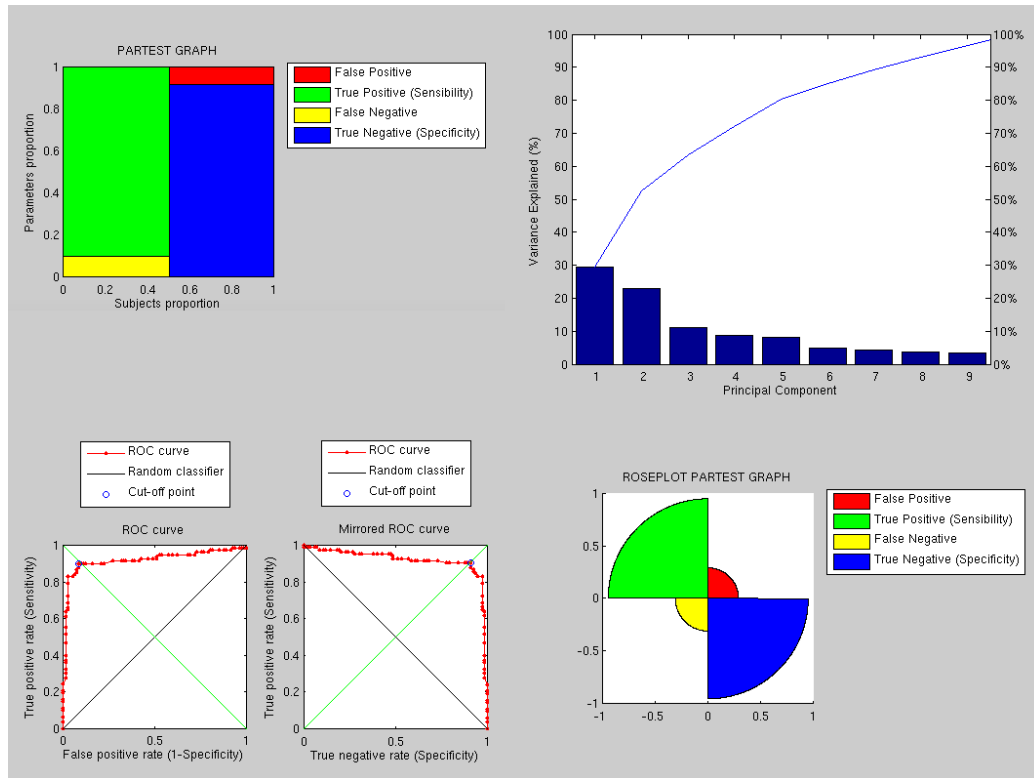


Figure 75: The FP analysis of the results of a model-based approach, with the breakdown of modes shown at the top right

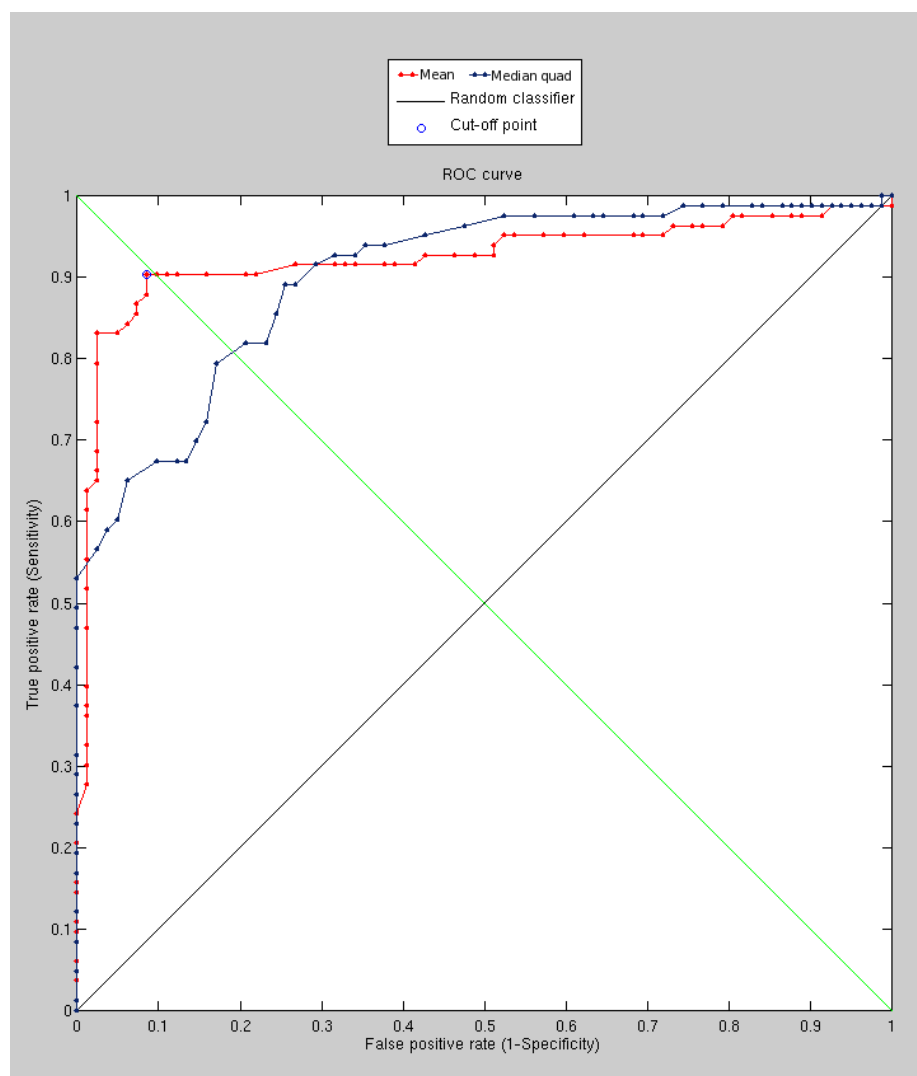


Figure 76: Performance comparison between an approach where the median of squared differences gets compared to mean of model changes

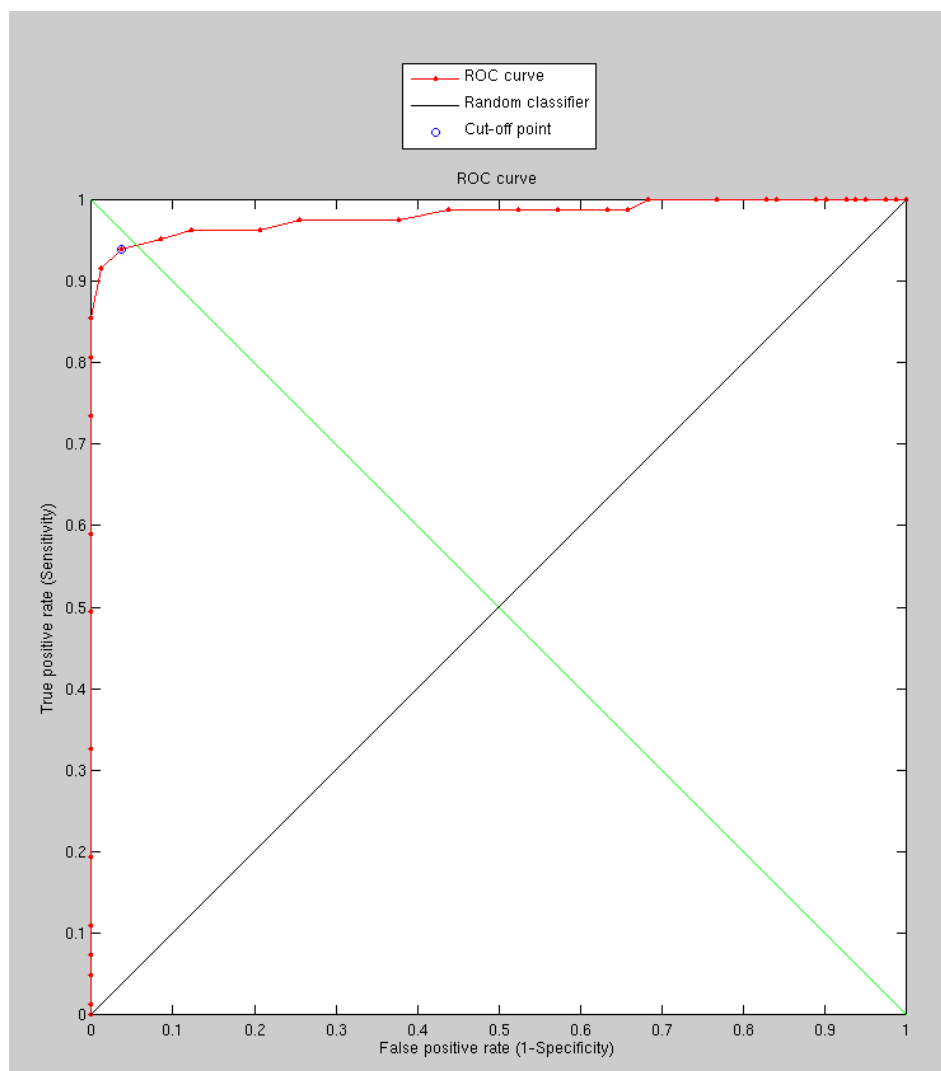


Figure 77: Performance of recognition when the absolute differences are gathered by their median

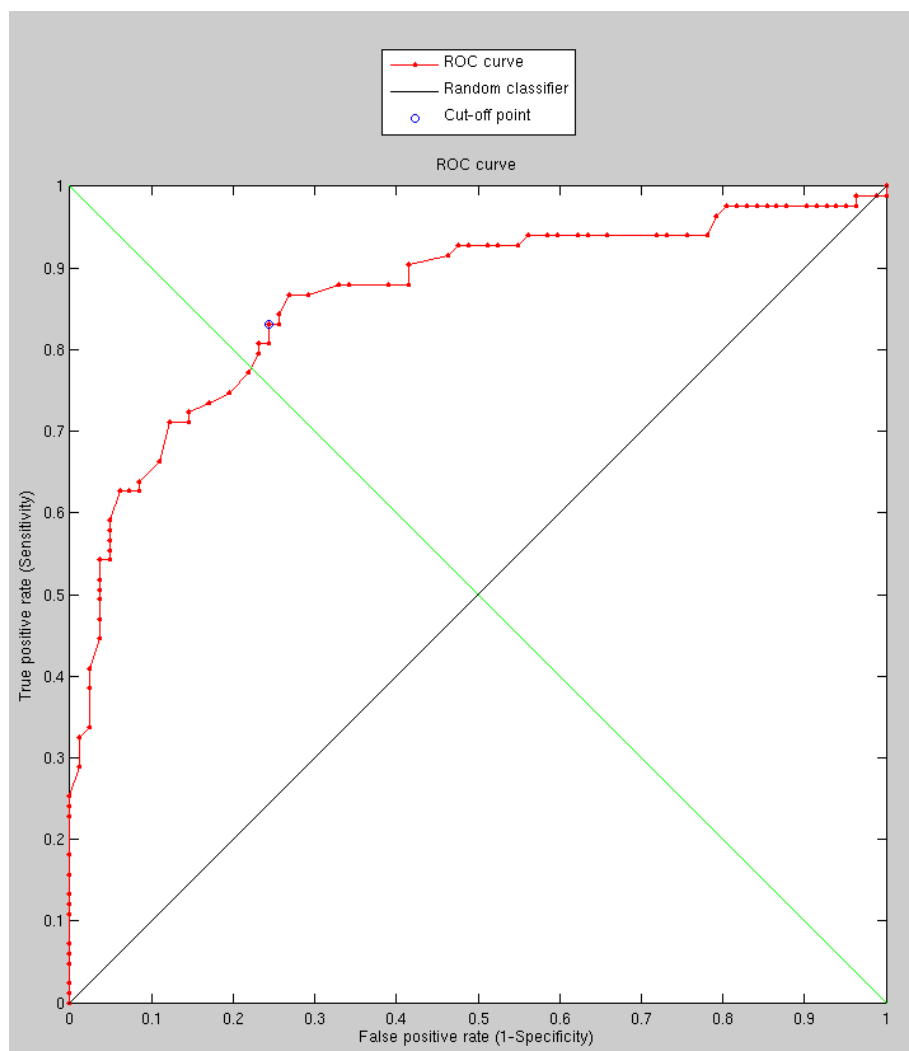


Figure 78: Performance of recognition when the squared differences are gathered by their means

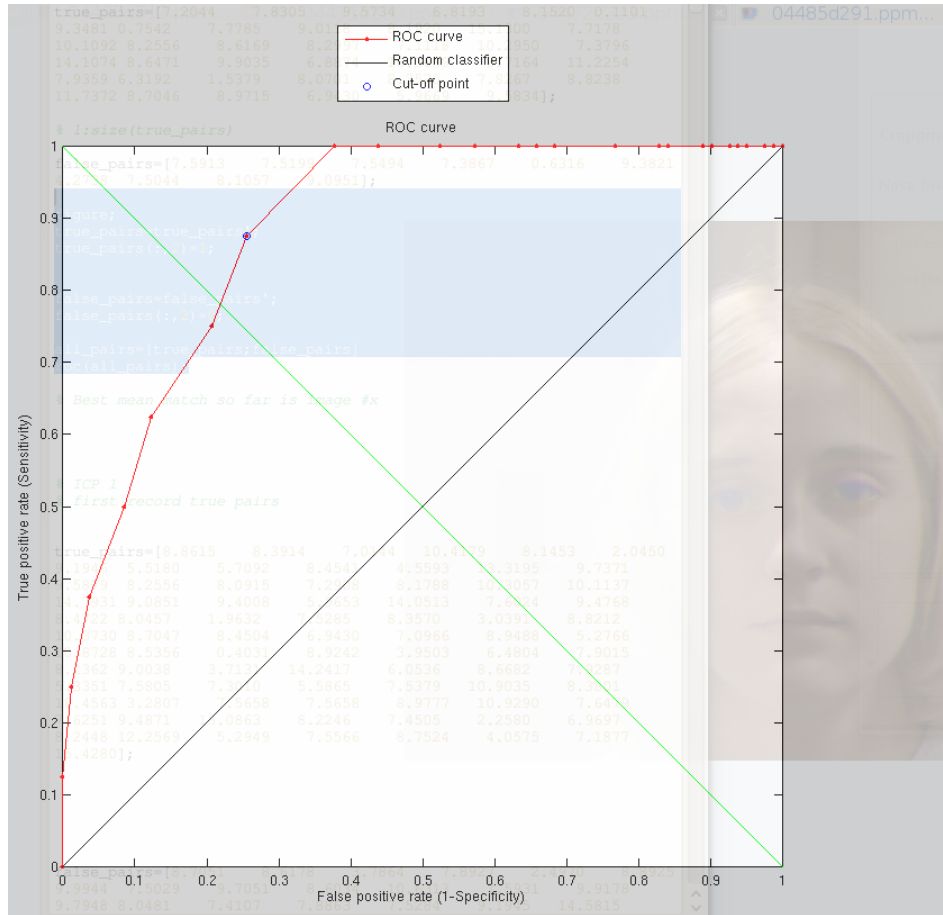


Figure 79: Degraded performance when the compared face pairs are not ones that were used to train the PCA model

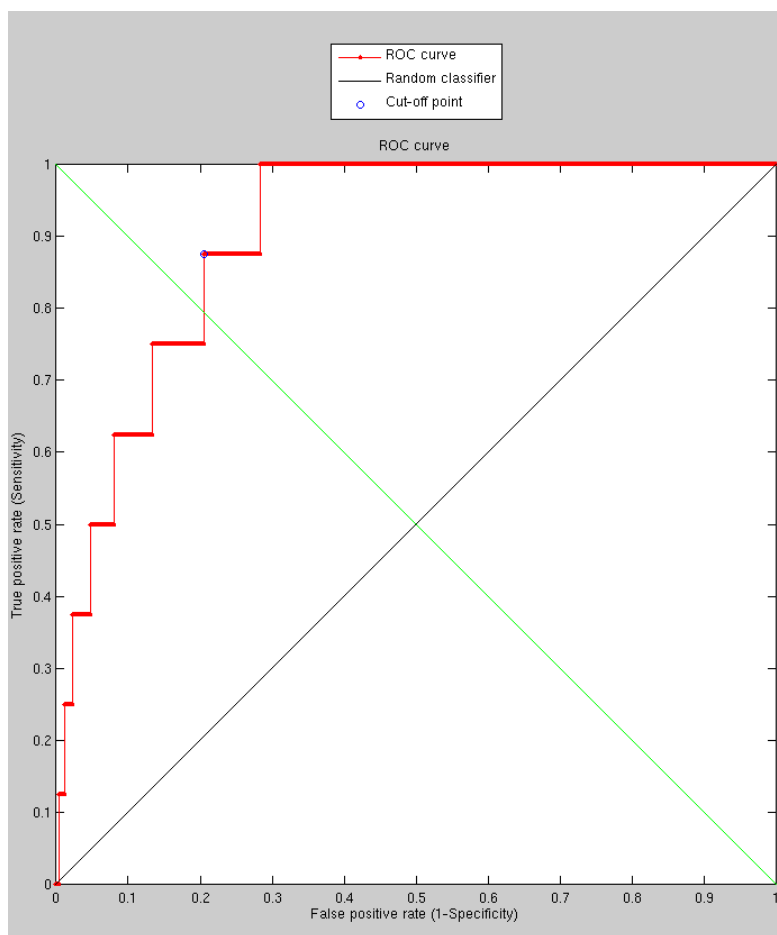


Figure 80: Comparison assessment with a large set of false pairs. The results of random pairs versus unseen pairs with expression differences.

Addressing the caveats in turn, we can start compensating for set sizes by manually selecting more of them, then building models of a greater scale (high level of granularity). What would further complicate the process is an inclusion of a wide range of separate expressions, without separation between them. The next experiments will elevate the level of difficulty by basically modeling the variation within many existing pairs, first without special ex-

pressions and later with all sorts of unknown expressions. The goal then it to show detection rates with or without expressions, either in the training set or the partition of targets. Experiments will take longer to design (requires manual organisation per individual) and also to run. The favouring of arduous tasks helps distinguish between good methods from lesser effective ones, especially at edge cases. If run on simpler sets, the results will improve considerably.

It has become evident that by doubling the sampling density, the models are now a bit more detailed and they incorporate more pertinent bits of information. To merely test the surface, the harder sets (from the "fall" semester) had been taken and 10 individuals were selected from there. Using 100 images (i.e. 50 residuals) in total we build a model and we set aside images of the same 10 individuals – those which will be containing 10 residues. These are separate from the training set. Checking model match for these 10 images and then comparing that with random pairs we get the following ROC curve which Figure 88 depicts.

Selected from the set to be used as correct pairs are just the first 10 people, do there is no cherry-picking. The random selection of the rest is truly random, too. The reason for the nature of this experiment is that it is extensible in the sense that we can carry on selecting people sequentially from the set (a lot of manual work required) and set the number of false pairs to be anything we like, as the random selection leaves $N \times N$ possible pairing for the $N \simeq 4,000$ images that we have at our disposal.

The next step would involve introducing more images into the experiment. If that good use of time turns out to give equally promising results, then other experiments can be designed with similar images too.

Granularity level, where sample point separation for PCA is 6x6, 8x8 and 10x10 (10 pixels/voxels apart), was next tested for better understanding of the space being explored, and subsequently for insight into how to set parameters. The next experiment explores the impact of granularity level in model-building on the overall performance. In order to make the experiments more defensible, the training set is changed from a size that can be described as minimalist (100 images) to 170, which requires further manual work. The model is being built from this set and performance then tested as before. The number of targets has been doubled. The experiments are still extensible in the sense that they can be rerun with a larger set.

The process involves building a model (limitations on the size of the training set notwithstanding), then doing assessment work on the target set with correct matches and another target set with incorrect matches, repeating for each model granularity level. These experiments take many hours to perform and so far it can be shown that by sampling more and more points performance is actually degraded rather than improved. This is not so counter-intuitive because by oversampling we lose sight or focus of large-scale structures and start comparing almost-to-be-treated-as-textural changes on the surface, including teeth for example. This probably needs a more careful look for better comprehension. Many of the caveats remains, particularly

set sizes and the nature of chosen datasets. Figure 83 shows the results and Figure 84 helps validate the consistency among the cases.

Smoothing was considered but hadn't been implemented as it was assumed that with enough points the impact would be limited. Points were knowingly sampled at fixed intervals without compensating for neighboring points in the vicinity/locality.

With smoothed surface sampling we get slightly different results, based on a couple of experiments I ran; the results of the smaller one are shown in the image. As the image in Figure85 shows, if we smooth before sampling, then the results are actually degraded (accompanying model decomposition in Figure 86). Maybe the smoothing was just excessive, or maybe the experiments were not large enough to inspire confidence (the error bars would be large if they were plotted). We could run the same experiments on unseen faces with expressions in all pairs in order to get smoother ROC curves, but it is really performance which needs to be addressed first (getting into the ballpark of 90% recognition rate or better for difficult sets/edge cases).

The next experiment explores a new model-based approach that I am implementing as the current approach leaves much to be desired and it also requires far bigger sets for training of the model. The group of people whom we are emulating used thousands of images for training and these images were not part of the FRGC set. Currently we use a poor training set which is also very small, so the models are of poor quality, just like the targets. Nevertheless,

this enables ideas to be explored, at least in the proof of concept stage, and it generally works.

It is reasonable to assume the smoothing most probably occurred already at the scanner level, and additional smoothing may damage the information rather than just act as an anti-alias filter.

7.10.6 ICP Revisited

In order to improve the underlying framework which aligns the training set – without taking any or much account of the structure of the face (e.g. parts to weigh more heavily) – exploration of the ICP process is undertaken again. There is a lot of work that can be done on improving it. The first random pairs (random examples) are shown as the residual, based on ICP with translation and rotation in all 3 axes (see Figure 87).

7.10.7 New Similarity Measure

In order to stride forward, another improvement is being explored. Currently, results where similarity is derived from the determinant of the eigenvalues of the covariance matrix seem promising (Figure 88). But the experiment was probably too small. It shows training on 170 images with just 22 images being targets. As before, the sets are generally hard and they are picked with expression variation.

Having spent many hours exploring other objective functions (or similarity measures with rigid transformation), the one which tended to work better was applied to a somewhat larger set and with the exception of few images that need to be looked at carefully, recognition in hard cases is basically improved, even with a coarse model. This one experiment samples 8 points apart and uses no smoothing. The next logical step would be to look at the cause for incorrect matching and also test to see the effect of rotation, translation, smoothing, etc. Literature on the subject also suggests how Lambda might be tweaked to account differently for eigenvalues. Results from the experiment are shown in Figure 89.

Putting the simple experiment in perspective, Figure 90 shows what happens when δ is varied in the sense that it is increased. As expected, this weakens the measure because it reduces the impact of zeroes but also weakens the signal. To succinctly explain the point of this measure, it is inspired by Kotcheff's work in the late nineties. It is quite simple to implement and it relies on an implicit similarity measure, which is an approximation of the quality of a model. This model is an aggregate model of known face residuals and a newly-introduced one (the probe). A correct match is one that results in high similarity -and builds a good model, characterised by concision. This observation was exploited to create a similarity measure that is data-agnostic and generalisable.

Similarity is computed indirectly in this case. The algorithm does so by calculating the model, namely by looking at the covariance matrix of that

model. To efficiently evaluate model complexity, $\sum_{i=1}^n \log(\lambda_i + \delta)$ is obtained where $\lambda_{1 < i < n}$ are the n eigenvalues of the covariance matrix whose magnitudes are the greatest and δ is a small constant (around 0.1) which adds weight to each eigenvalue. This approximates

$$\det(\mathbf{M} + \delta) \equiv \prod_{i=1}^n (\lambda_i + \delta) \propto \sum_{i=1}^n \log(\lambda_i + \delta) \equiv \log(\det(\mathbf{M} + \delta)) \quad (6)$$

where \mathbf{M} is the model's covariance matrix under consideration and δ is a constant which would quite importantly ensure nothing gets multiplied by 0 or a summation stuck too close to 0. This whole term is an approximation of similarity between images.

In order to test performance for much smaller values of δ there is a need to limit how many of $\lambda_{1 < i < n}$ to remove (those of least magnitude). This will be the next step. Later on, large sets can build better model with data which is easier to deal with and yields better results.

7.10.8 Effects of Lambda Changes

The image collection that is combined in Figure 91 shows 4 ROC curves. This comparison shows how varying the number of eigenvalues (organised in descending order) affects the results. The weakness of this experiment is that

it treats random examples that are not the same and given the size of the sets used for plotting, there is plenty of room for dependency on the stochastically-chosen set. That having been argued, it does not appear as though there is remarkable merit in limiting the number of greatest eigenvalues. The next experiment will look at how changing the value of delta affects performance, where the number of eigenvalues will be set high for obvious reasons (it is negligible when only high eigenvalues are accounted for).

While the choice of δ may heavily depend on the number n in $\lambda_{1 < i < n}$ (the further we go down the list of eigenvalues, the smaller their value is and the more impact δ has), Figure 98 shows the effect of altering the value of δ on overall recognition performance (still just a small set with facial expressions varying).

Based on the above results, we are from obtaining the published results by Mian *et al.* The idea we had in mind is to obtain similar results by re-implementing their exact methods and then introduce modifications using either R-PCA or GMDS, and investigate if and how we improve. The recognition rates (including expressions) are at a completely different scale than those reported even for early NIST/FRGC tests.

The scale of the experiments is vastly different because UWA trains a model on thousands of instances (some proprietary ones), which require one to sort pairs. In comparison, we build a model with just a couple of hundreds of examples and in order to speed things up we rescale the images.

Additionally, we could train on images without expressions and then apply the algorithm to the whole NIST/FRGC set, which comprises a vast number of neutrals. The main caveat is the need to prepare more data. It should be possible to invest some hours expanding the size of the training sets and auditing the results. There are many ways to make the recognition problem easier, e.g. by selecting particular types of images rather than edge cases we currently deal with. From a general point of view, performance can be improved later by designing an experiment to also include easier cases.

7.10.9 Debugging ICP

It is worth clarifying that the better performance shown before was achieved by applying the algorithm to a different set which was too easy to deal with. Further improvements are still needed to avoid the rare occasions of mislocation of the face (edge cases) and also ICP stepping out of line. It is only in the interests of speed that we still deal with coarse images such as the one in Figure 93. It impedes performance improvements but makes tweaking/debugging considerably simpler, even if it's an interim phase.

Whilst introducing various improvements, a side effect was the emergence of some bugs, an annoying one of which affects ICP and leads to some failures that are difficult to explain by regression.

I found some bugs, but identifying the main culprit is still an elusive task which affects all 4 families of ICP currently in use. Figures 94, 95, and 96

can help some light on the debugging process.

Following further regressions, the bug which some previous changes had introduced along the way was found as per Figure 98 and then removed (resolved by reverting back to correct code), leading to the sorts of image differences pre- and post-ICP that are seen in Figure 97. The new distribution is shown in Figure 99, but in order to start showing competitive performance many hours are being spent going through the thousands of images – including those which are easy to handle – and sorting them for intra-subject sets that are necessary for model training and later for easier assessment (we mostly death with difficult cases so far). Rather than train a model using just dozens of people with various facial expressions we can use many hundreds of them with and without expressions (mostly with none), then show high performance as before. This has required a massive time investment so far, but it is likely to pay off. Two universities which appear to have data of this kind had been contacted months ago, but this engagement was unfruitful. Organising the reminder of the images accurately can take many hours. It is also cumulative in the sense that faces already sorted can be merged into the newly-organised sets.

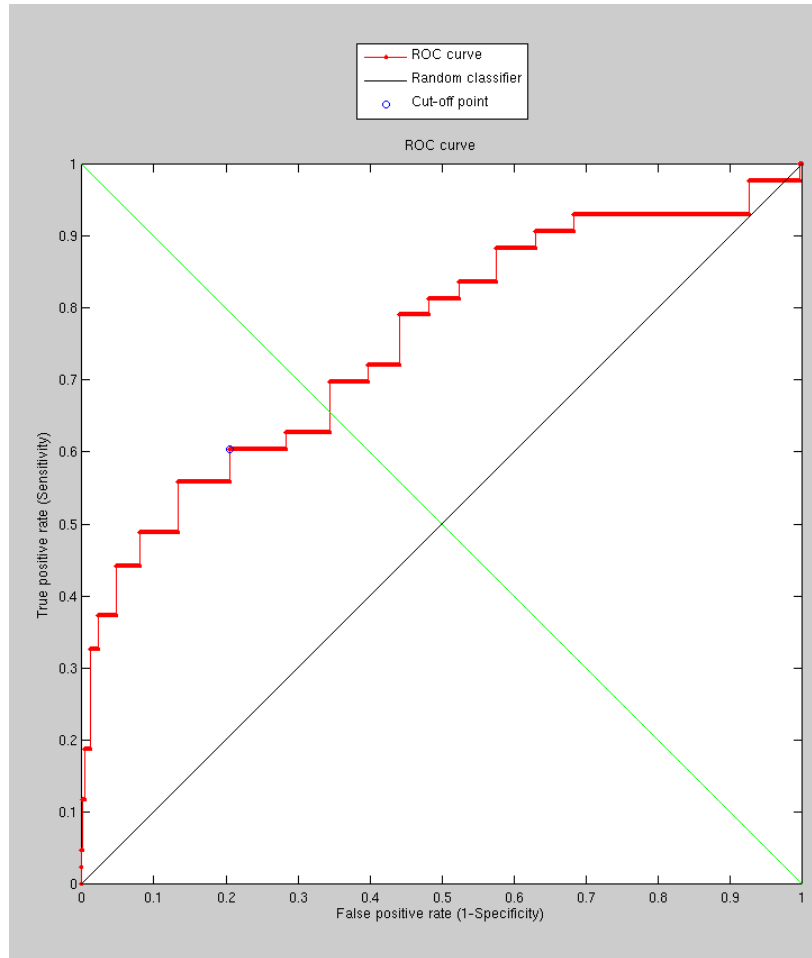


Figure 81: The results of matching random pairs from different people and from similar people, with and without expression (based on expression models)

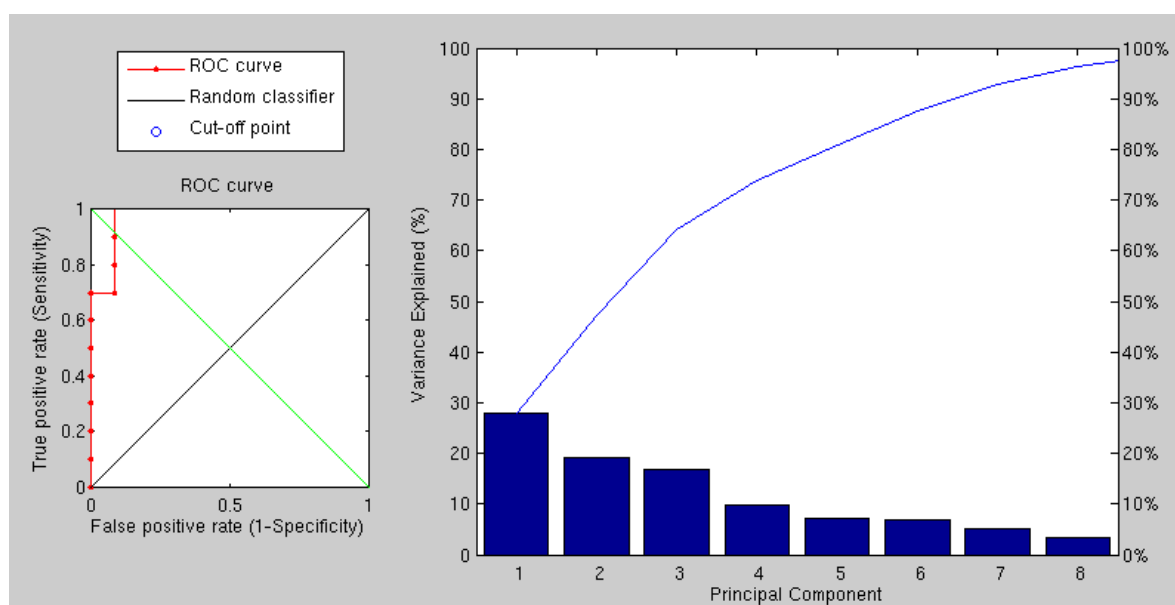


Figure 82: The results of comparing correct pairs to random (and false) pairs using the model-based approach. The right hand side shows the breakdown of model modes.

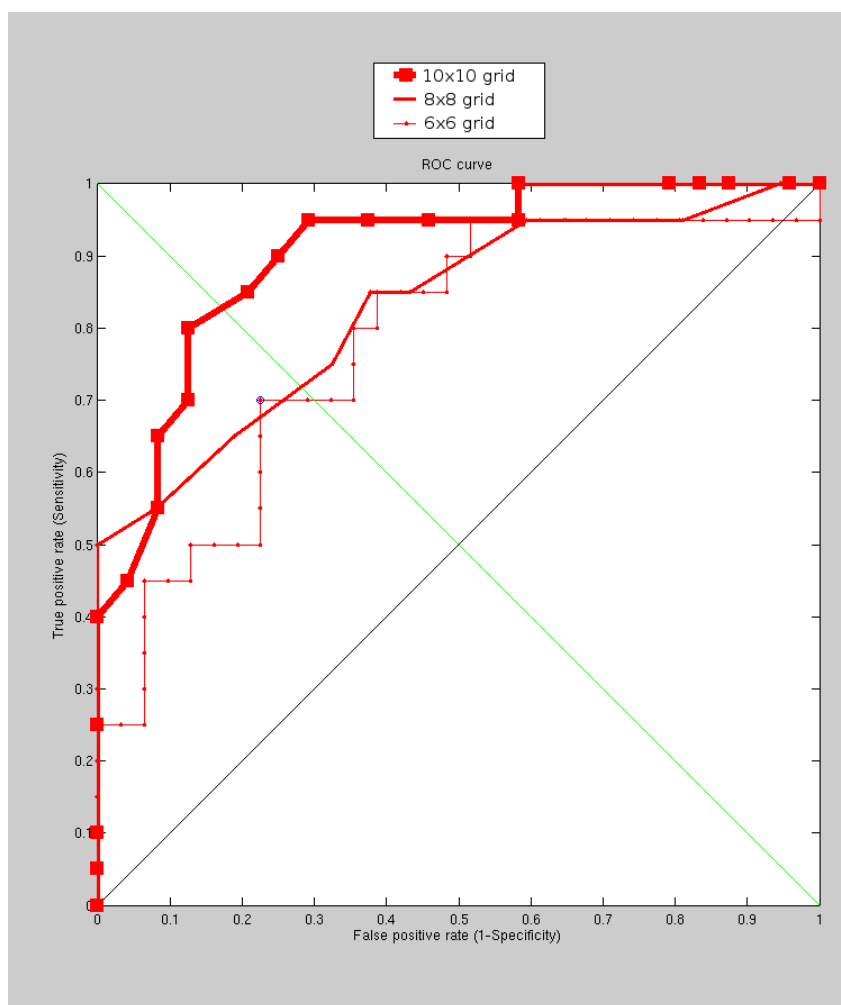


Figure 83: Performance measured on relatively small sets, empirically showing that coarser grids yield better recognition performance

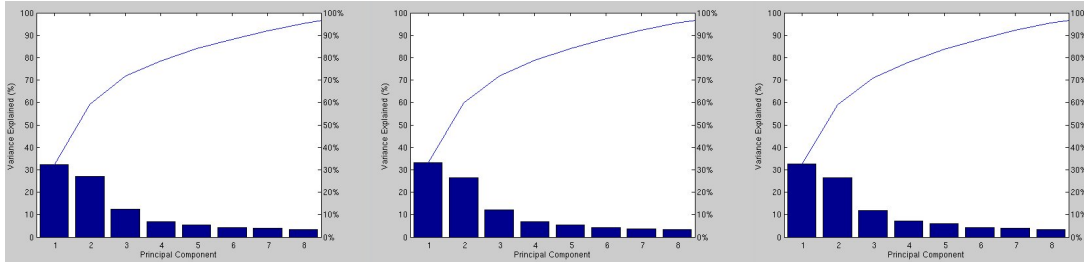


Figure 84: Decomposition of the different modes of variation for the three cases, namely granularity levels 6x6, 8x8, and 10x10, respectively (10 pixels/voxels apart) demonstrating that despite the changes in resolution the model modes have a similar distribution and are probably inherently similar, as expected

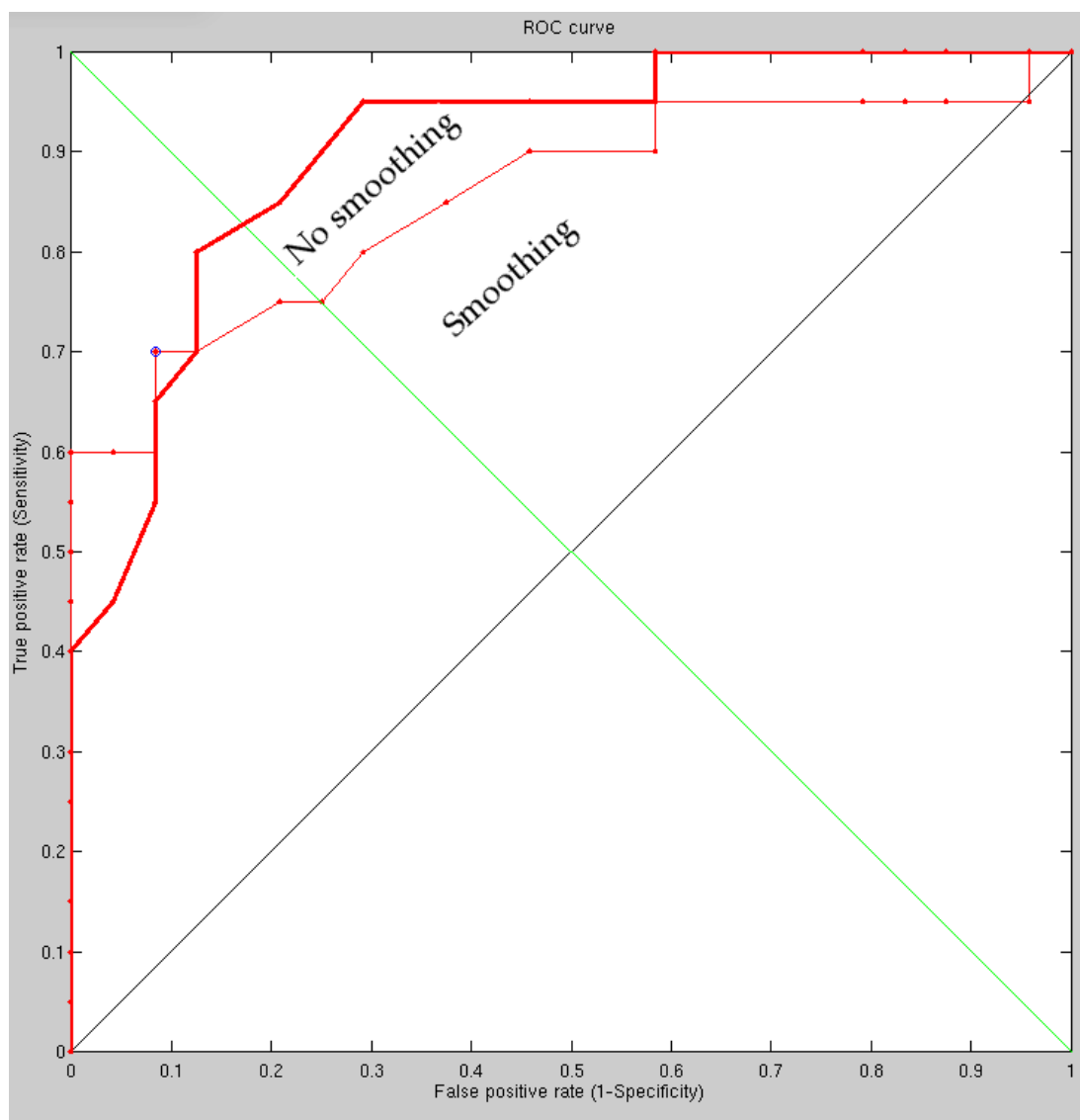


Figure 85: Comparison between the performance of the method with smoothing applied before sampling and without any smoothing at all (which gives similar performance)

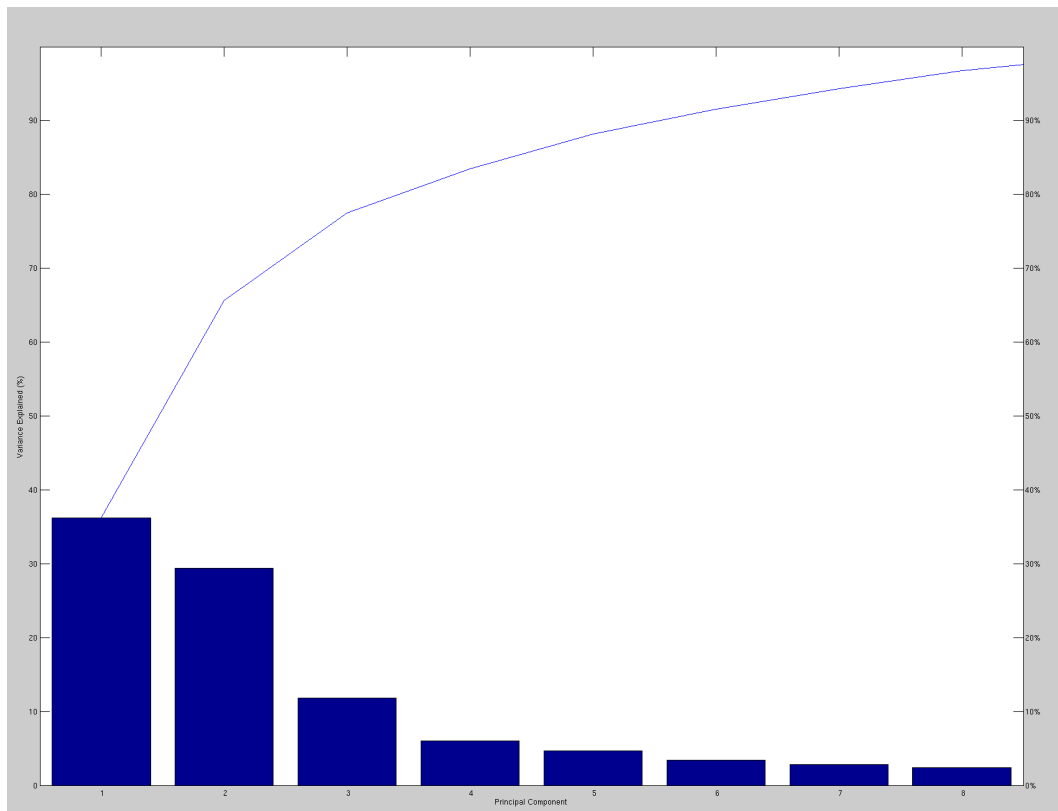


Figure 86: The decomposition of the model as a chart corresponding to Figure 84 on the right, this time with smoothing on

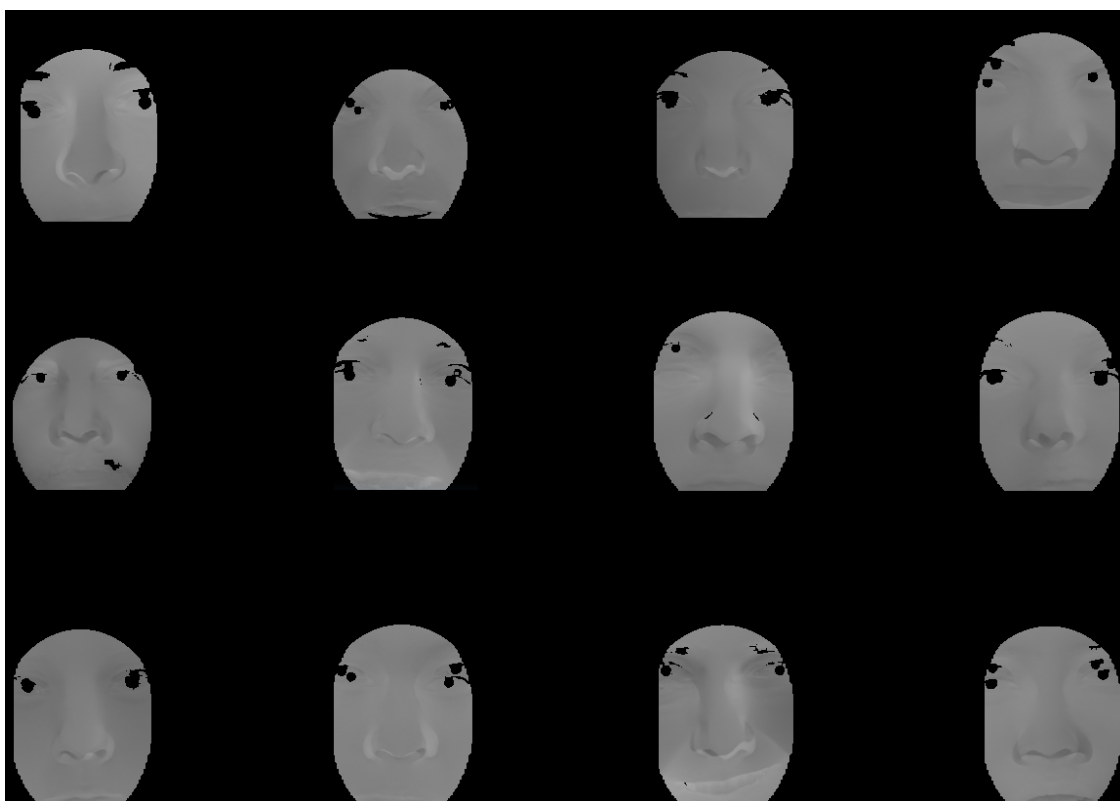


Figure 87: The first few face residues following alignment with ICP (sample points being around the forehead, nose, and eyes)

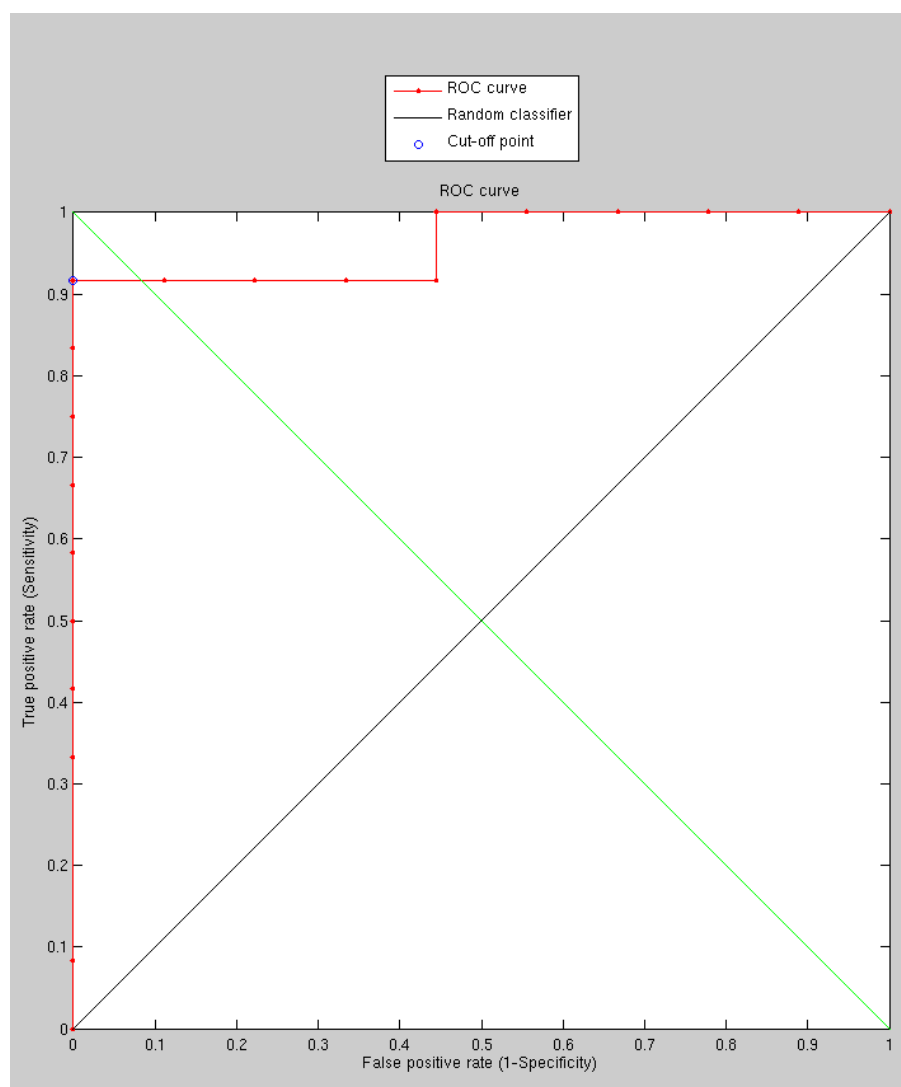


Figure 88: The results of measuring the similarity by determinant of the eigenvalues of the covariance matrix and engaging in a recognition task

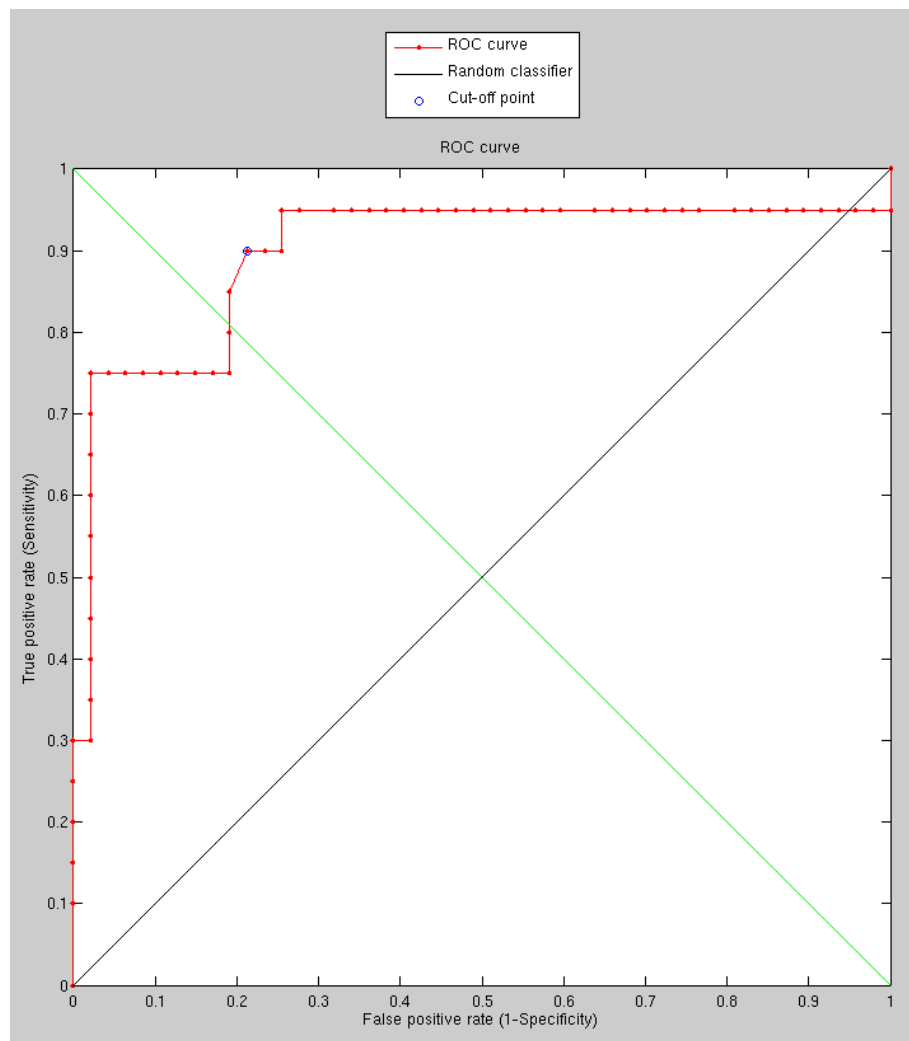


Figure 89: Results of an experiment where the determinant is again being explored, this time with a larger set

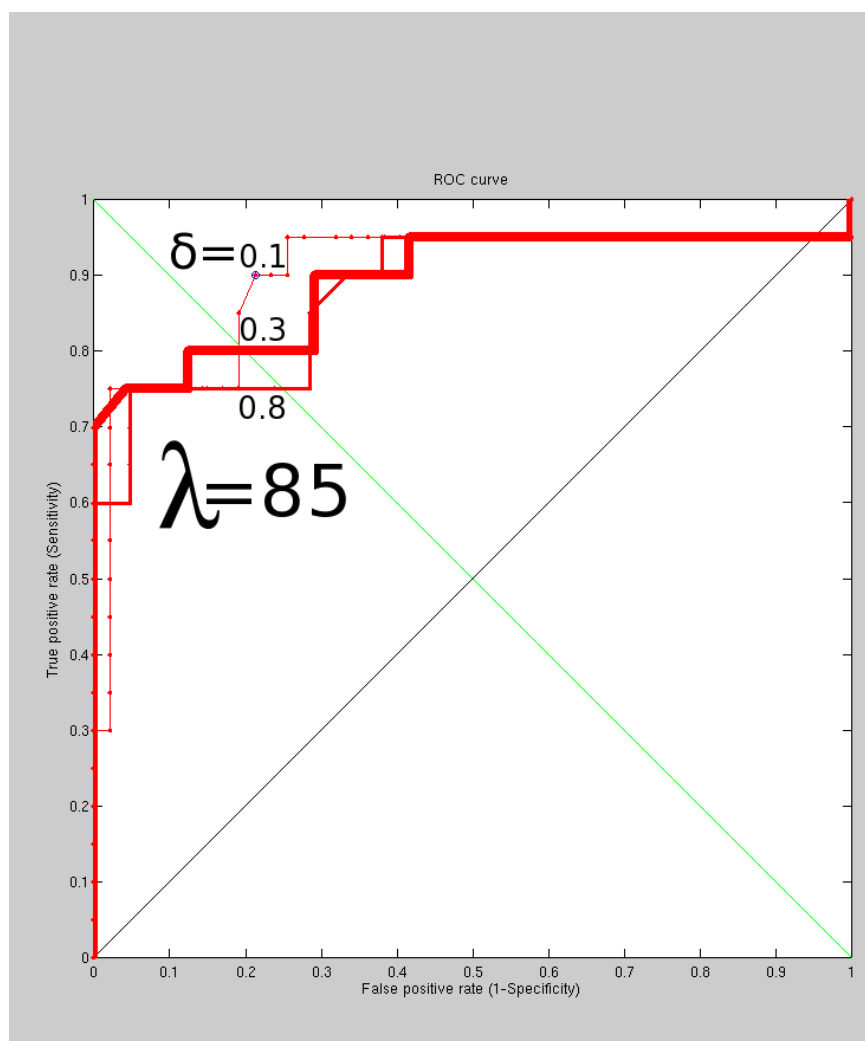


Figure 90: Results of an experiment where the determinant is again being explored with a comparison of the curves for 3 values of δ

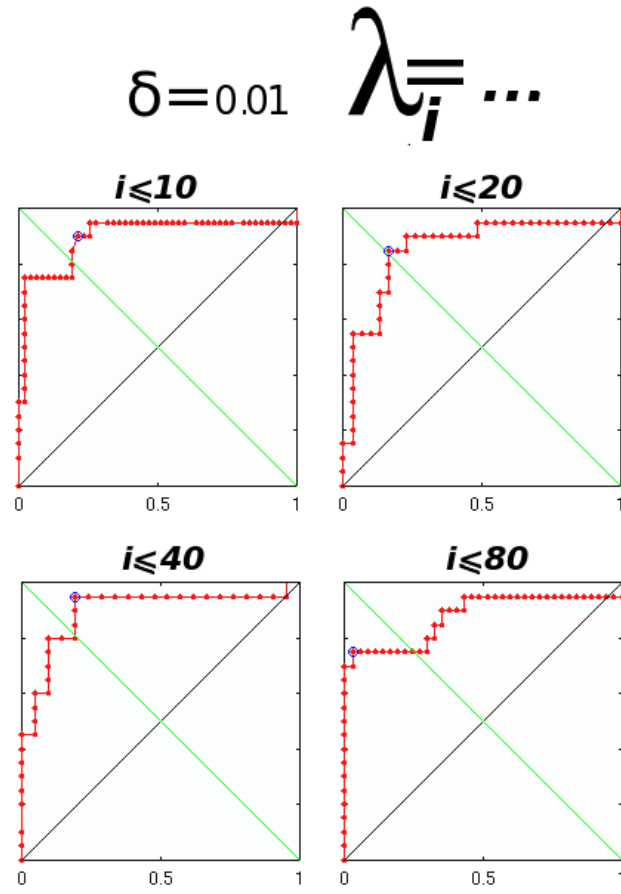


Figure 91: The result – in terms of performance – of varying n in $\lambda_1 < i < n$

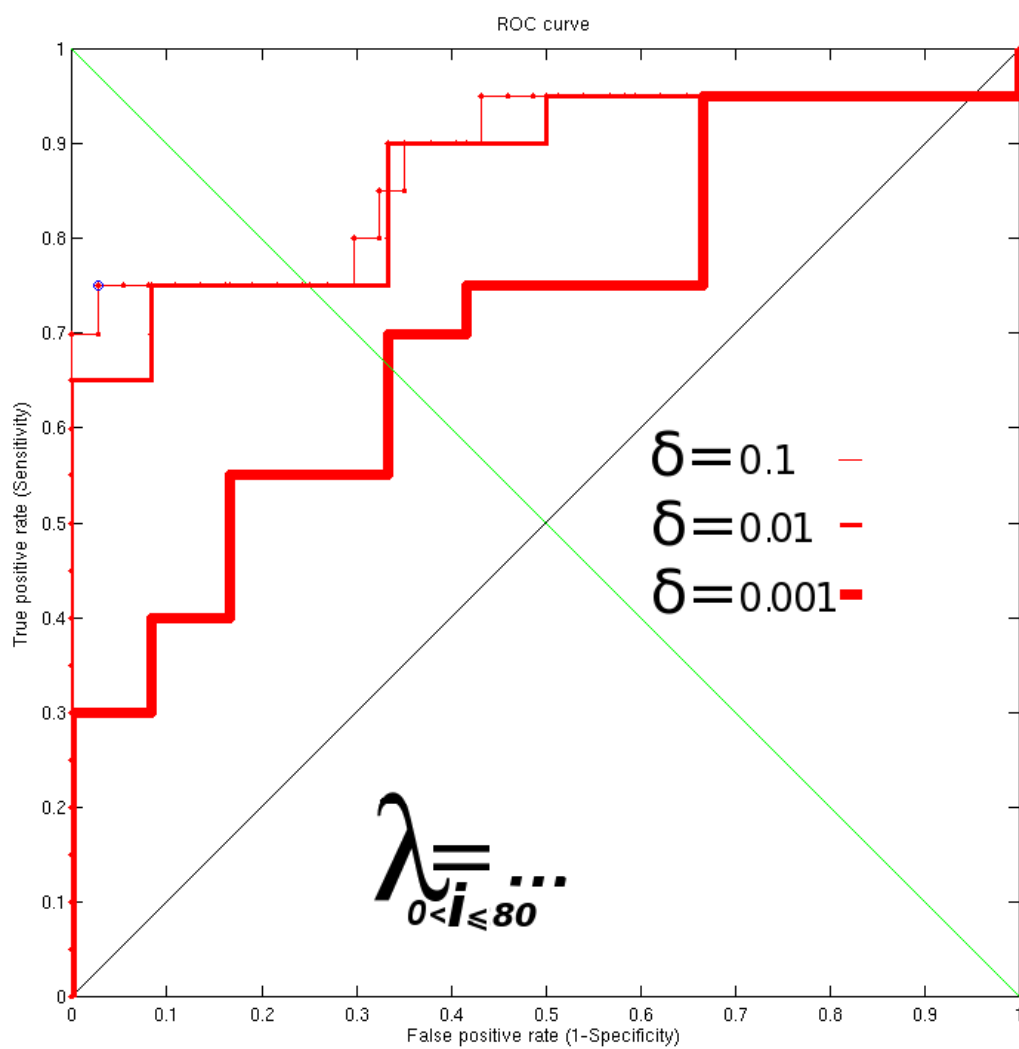


Figure 92: The effect of changing the value of δ on the overall recognition performance

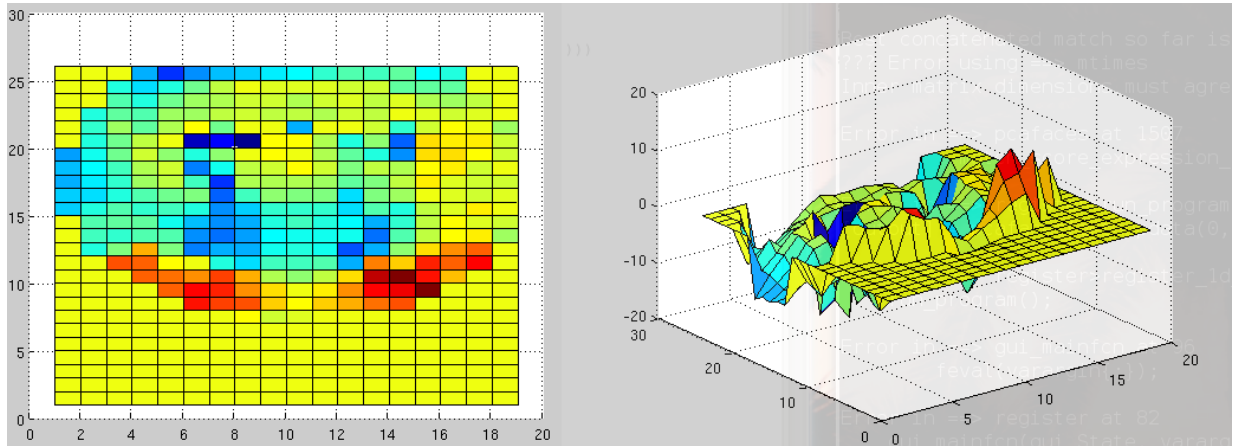


Figure 93: An 8x8 separation between points in the image (shown from two angles), with downsampling done for debugging purposes

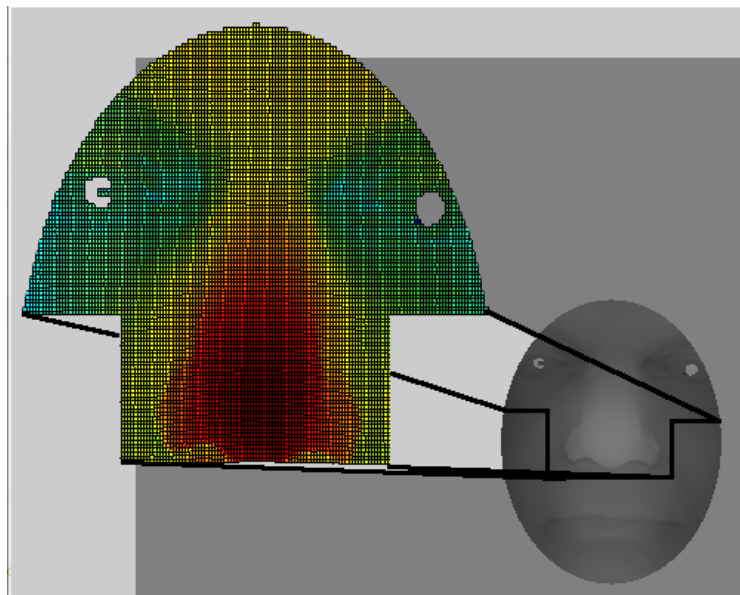


Figure 94: A slice or subset of the data being used for ICP (on the left) and the masked face from which it is extracted (right)

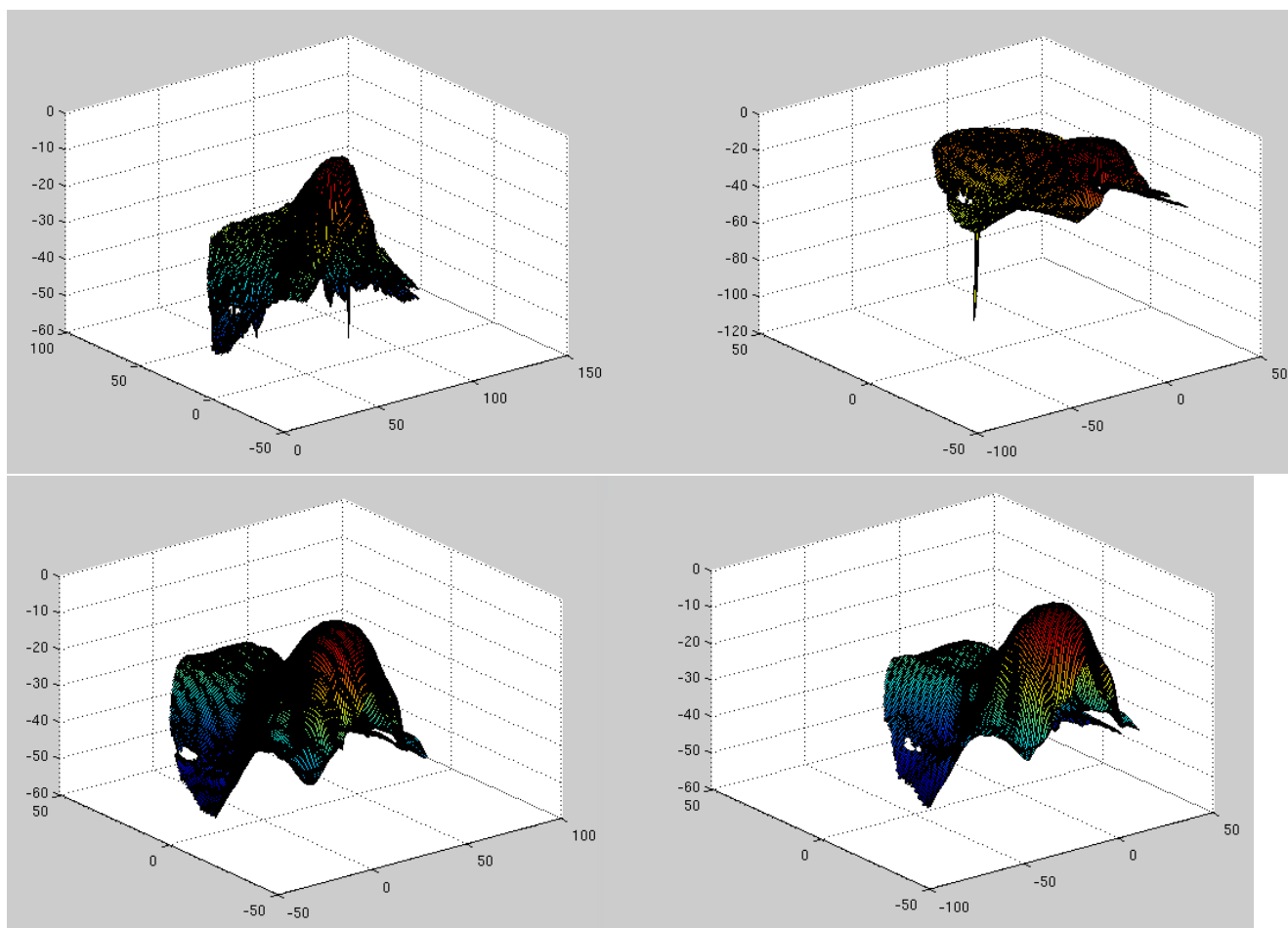


Figure 95: **Top:** Two images taken from the same individual being compared when there is insufficient compensation for noise. **Bottom:** another set of such images but where smoothing is applied to reduce noise-imposed anomalies

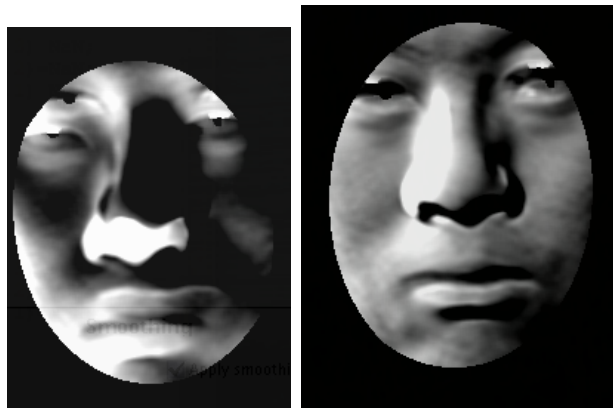


Figure 96: On the left: the result of poor or buggy ICP (difference); on the right, an image is shown of the type of image we expect to have and also get when ICP performs well

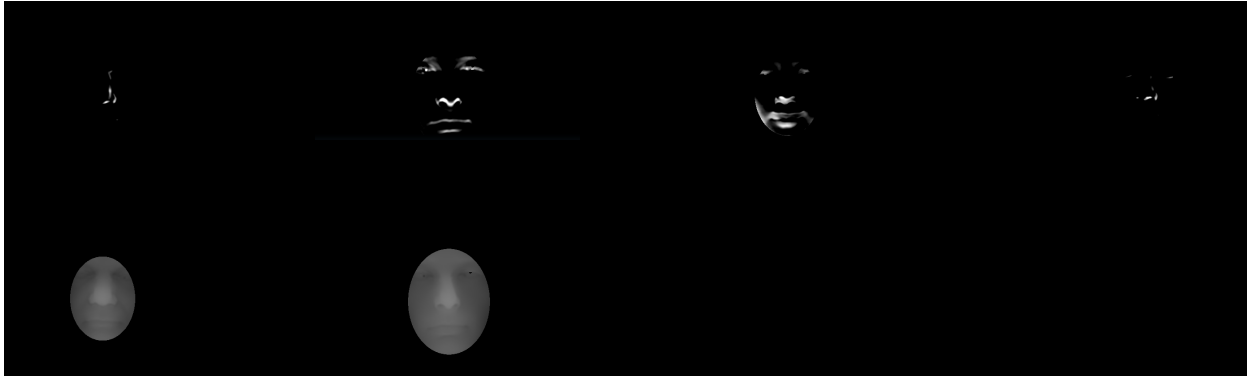


Figure 97: Difference between the first 4 images before and after ICP (rotation and translation), with two of the first reference images shown at the bottom just for a sense of what the images at the top are derived from

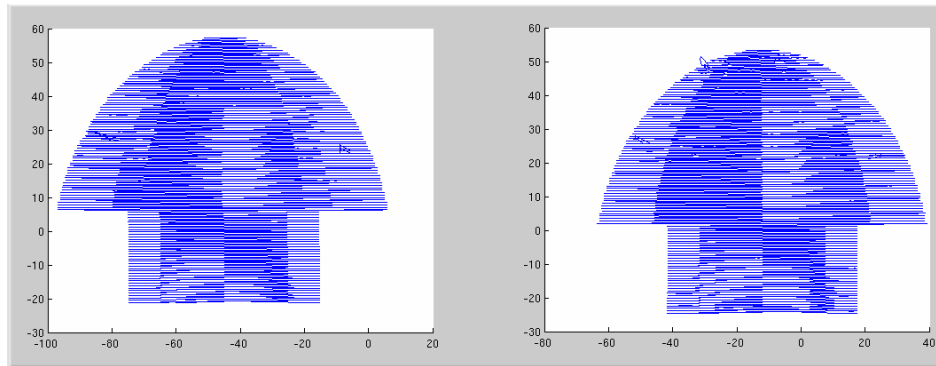


Figure 98: The effect of the bug demonstrated by showing misalignment on the X axis (and to a lesser degree in Y too).

Preparation of many images for experiments with superior results was an important next step.

Not-so-considerable improvements are arrived at by taking a Spring Semester set and building an ICP-free model from it (not complete, about 250 pairs) with sampling separation of 8 so as to avoid running out of memory at the

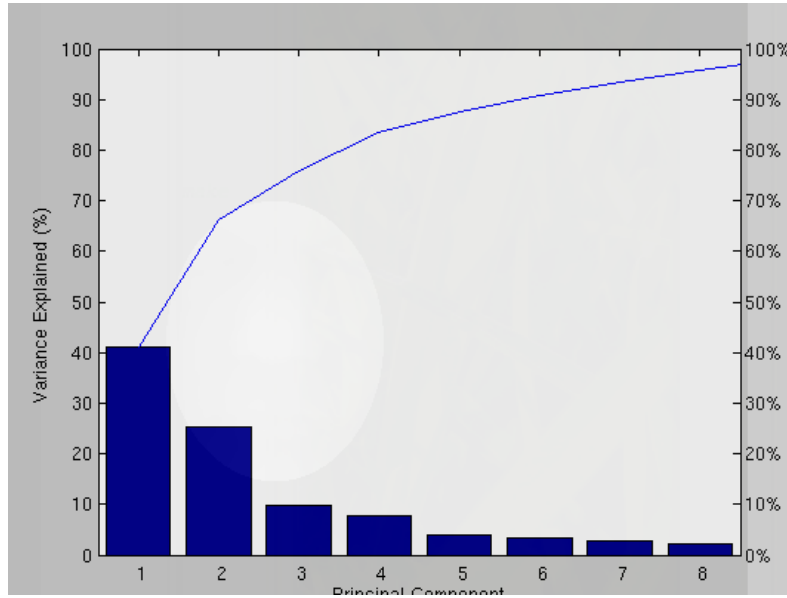


Figure 99: The new distribution of modes following the bugfix

PCA stage (dense sampling makes the model unhelpfully vast). This was tested on a separate Spring Semester set of real pairs versus random pairs from NIST/FRGC, where the examples tested on are unseen, i.e. none got used to train the model (if some of the probes are used for training, performance comes near 98% because the model is familiar with the probe). With a lot more data at hand it should be possible to produce much smoother curves. There is still debugging and fine-tuning around ICP (as shown in Figure 100 and Figure 101) with 4 different implementations that give different results. Clearly these have a lot impact on the results provided they work correctly. In many of the experiments so far ICP rotation gets switched off. This enables the modeling of rotation although, ideally, we should try to remove head rotation also in the probes. To put it differently, the model al-

ready incorporates rotation as part of the variation, whereas aligning around the centre of the face can (and probably should) be assured.

What makes this while process enormously time consuming is the adequate division into sets, which makes the reference arbitrary and the process rather autonomous. The goal is not to cheat with statistics by biasing the results with a training set not belonging to the targets; it seems to be what some others are doing in order to prepare the matcher for particular observations. In any event, much bigger sets (with almost 1000 images to cycle through) are now generally available for the next experiments, which will compare ICP algorithms and yield results with less human intervention. The computational server had been under a lot less load recently, so getting results like those shown in figures 102 and 103 takes about 4 hours.

My flight arrives at Israel next month (booked now). Looking forwards to it!
Cold and rainy here...

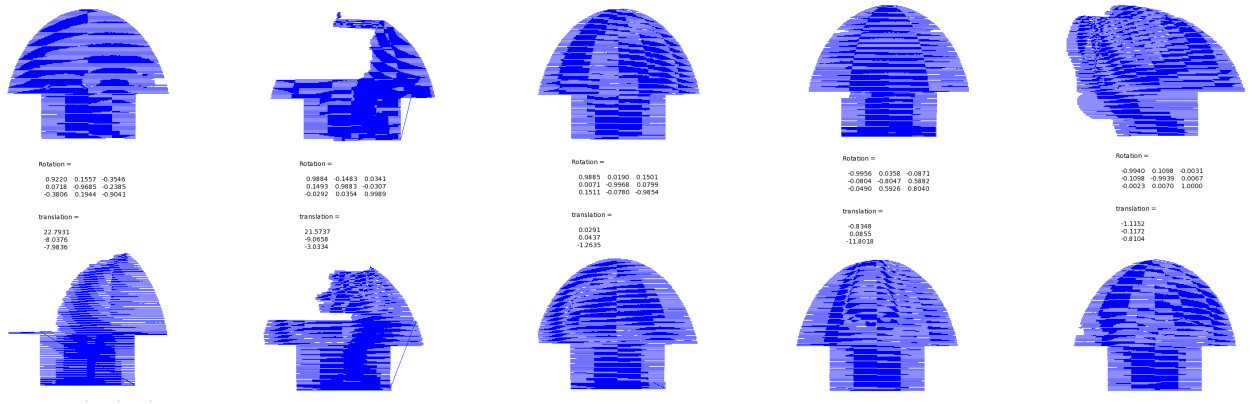


Figure 100: The effect of perturbing the points on ICP

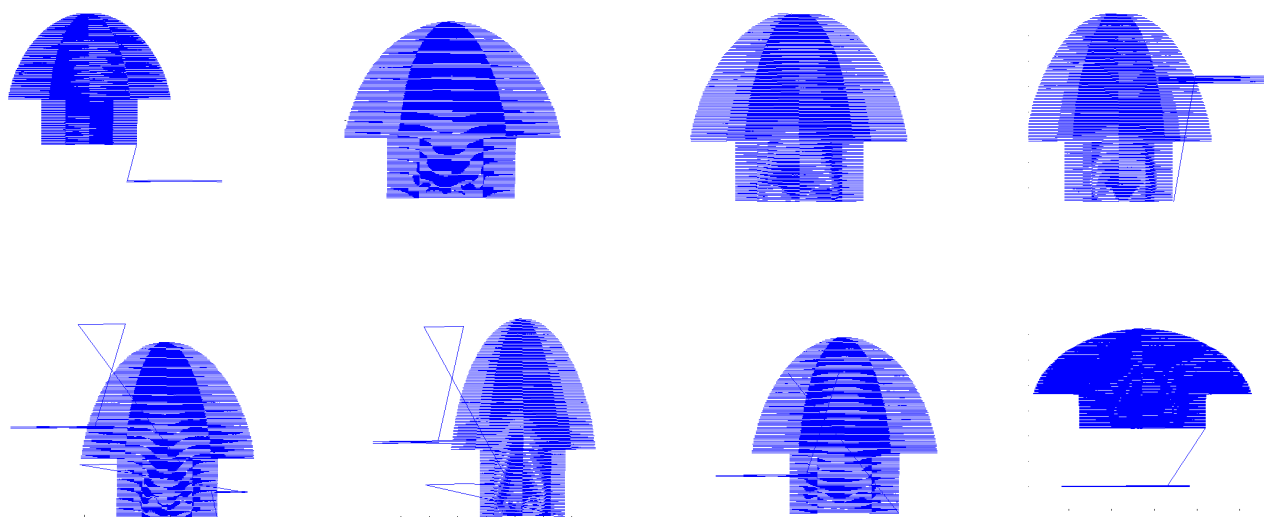


Figure 101: The effect of noise on ICP studied by aligning images 1-5 at the top to images 6-10 at the bottom

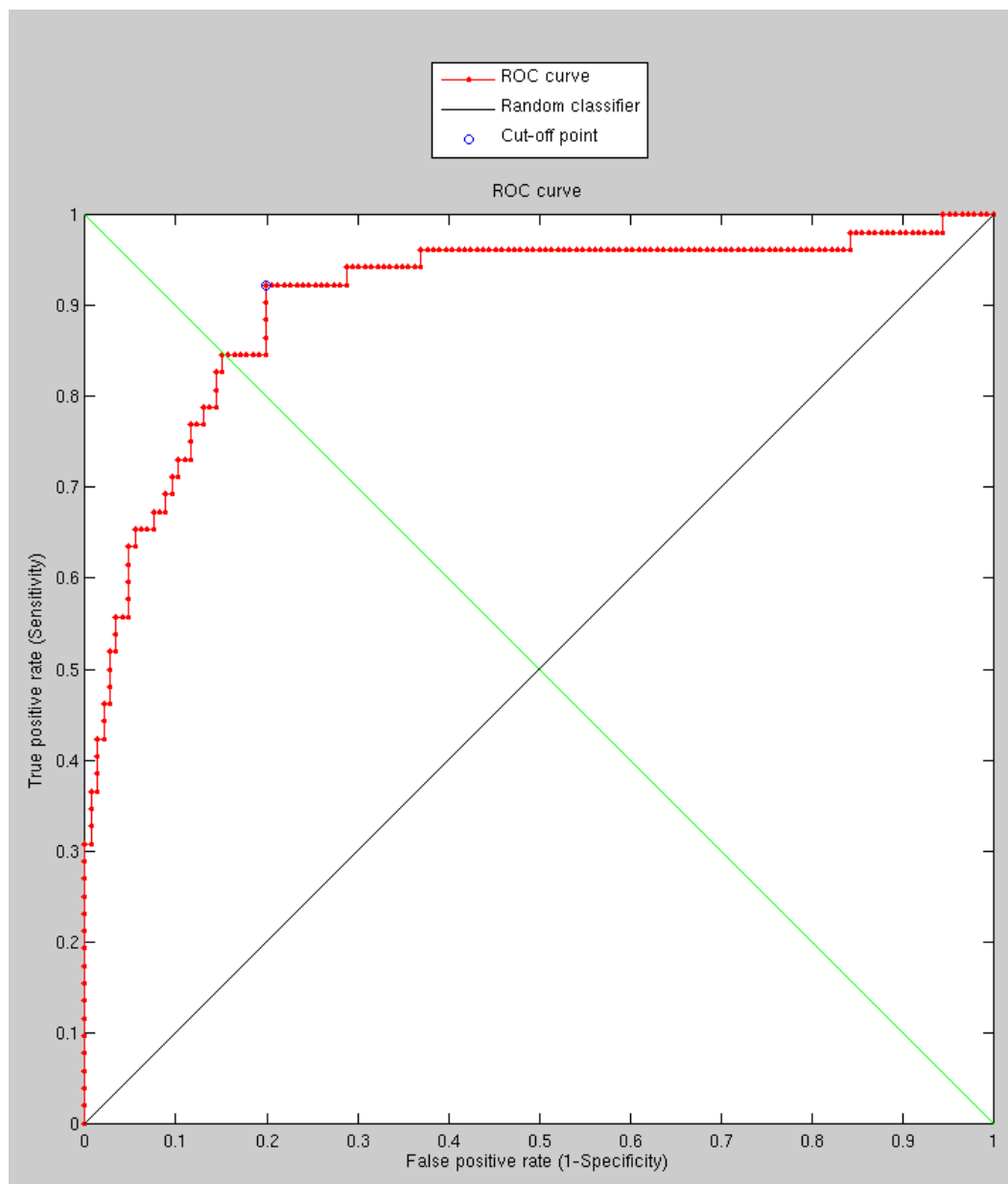


Figure 102: The purely median-based performance on the Spring Semester set, without ICP

7.10.10 Translation Explored

Encouraging results come about with the changes to the ICP routines, adding to what previously was tested at another level, namely model and residual. With improved smoothing and with GIP's v1 of ICP (older) the results are improved significantly. Visual examples with and without translation (but with GIP's v2 of ICP) are shown...

Improving this further with rotation and a model that annuls translation and rotation (currently it does not) should be trivial, with additional refinements incurred by use of larger models and simpler sets (the ones currently used are full of different expressions).

ICP experiments were run until papers were explored again. One experiment in the pipeline strives to emulate ICP as described in Mian's earlier papers, on which he based his Ph.D. (he did it at the same time as myself).

Upon completion of a larger experiment it turned out that it was designed incorrectly because ICP – with Mian-style translation¹⁷ – did not work correctly and therefore the model built was improper. It did occur, however, that despite this fluke there is decent capability within the new model to detect pairs (maybe an accidental discovery worth exploring in the future). The results are shown in Figure 109.

Our newly-combined set of Spring and Fall Semester builds a model with

¹⁷Mian-style objective function needed for comparisons, with rotation/clipping addressed as part of him transformation needs fixing.

translation, where the ICP algorithm is as the same as UWA's. The test sets are also of Spring and Fall Semester, but they have no overlap with respect to the training set and they still have many expressions in them. To distinguish between expression differences (intra-personal) and inter-personal differences we use a coarse model which is fast to build and match to.

Unintended arrival at experiments where a model-based approach by far outperforms the median-based are probably not of high priority at this stage.

To begin the exploration of robust PCA as proposed by Yi Ma (of MSR-China, a copy of his book (350+ pages) was obtained, hopefully with concise summaries too.

7.10.11 Multi-feature PCA

A multi-feature *PCA approach is being embraced and a suitable algorithm is being put into the same framework as before. For testing and debugging purposes, X and Y derivative images are being calculated (estimating depth differences in the face, from a frontal perspective). See figures 110 and 111 for visual examples.

A closer look at the SVN repository revealed nothing relevant that can intuitively give MDS-esque matrices, but that too will be added. The foundations must be laid down and debugged first.

For each of the two surfaces, S and Q, the steepness of points along the Z axis can indicate the degree of curvature and irregularity, although distances

are absolute and unless measured with a signed value, they will not convey information about directions. To measure this more properly we may need to travel on/near the surface and perhaps even interpolate to measure those distances more properly, namely in a way that preserves invariance properties. This effect will be studied shortly. The imminent goal is for PCA to be applied to the GMDS-esque geodesics matrix, which is a concise representation or coding of a face, mostly invariant to pixel-wise difference and motion of parts in connected tissue.

As a first stage, we take the X and Y derivatives (gradient) and consider these as implicit shape descriptors. To be more precise, we use derivative images with smoothing of radius 6 to have a sense of direction to be used as an identifier, not necessarily expecting it to be a valuable discriminant. This image is being smoothed because of the sparse sampling on a grid (8x8 points apart, which make up about 150 dimensions).

One could argue that the equally sampled set of curvatures provides insight into the spatial information in a way that is hardly affected by length of nose relative to the face, for example. Using a fusion of both might also be worthwhile, e.g. a combined PCA model of depth and curvature and/or geodesic/Euclidean distances.

So, we first come to grips with an experiment dealing difference or residual of derivatives (initially along Y only), essentially by building a model of these. The test set is still a hard one which is not sanitised from hard cases, but it

is merely used for comparative purposes here. The PCA is also not as robust as it could be, especially not to outliers.

Partial matching of faces is basically facilitated by these methods as omission of points is possible, although it makes the observations' length inconsistent (unknown position along one dimension or more). Throughout the preliminary tests (Figure 117 and Figure 116) the program mistakenly treated the X derivative of the Y derivative as though it was the X derivative (compare figures 112 and 113), but the matter of fact is that although this approach works poorly (no fine-tuning attempted and minimal post-processing *a la* Figure 114), it does help test the ground and lay the foundations for some new ROC curves in a pipeline that supports multi-feature PCA support, e.g. Euclidean distances fused with derivatives, depth, and geodesic distances as measurable attributes for characterising a surface. It would also be worth revising the PCA we use.

It is still implemented further so as to support two distinct features of different scale. Currently it is limited to two, but should be extensible enough in the code to support more with minor tweaks. There are also ways to get vastly superior performance, it just takes a lot longer to set up. The results here are to be treated as results from toy experiments (with bugs and unreasonable magnitudes).

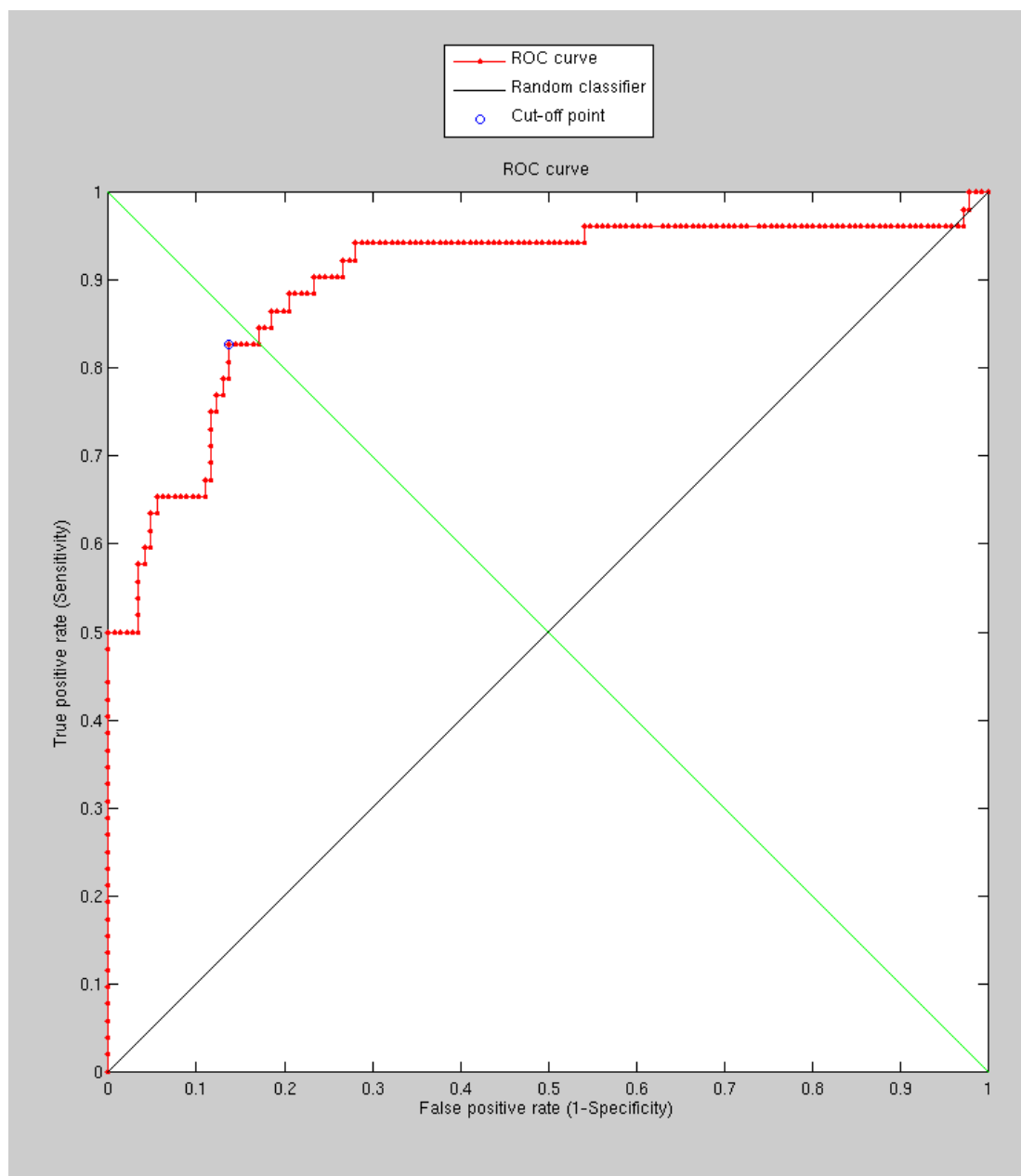


Figure 103: The purely model-based (determinant) performance on the Spring Semester set, without ICP

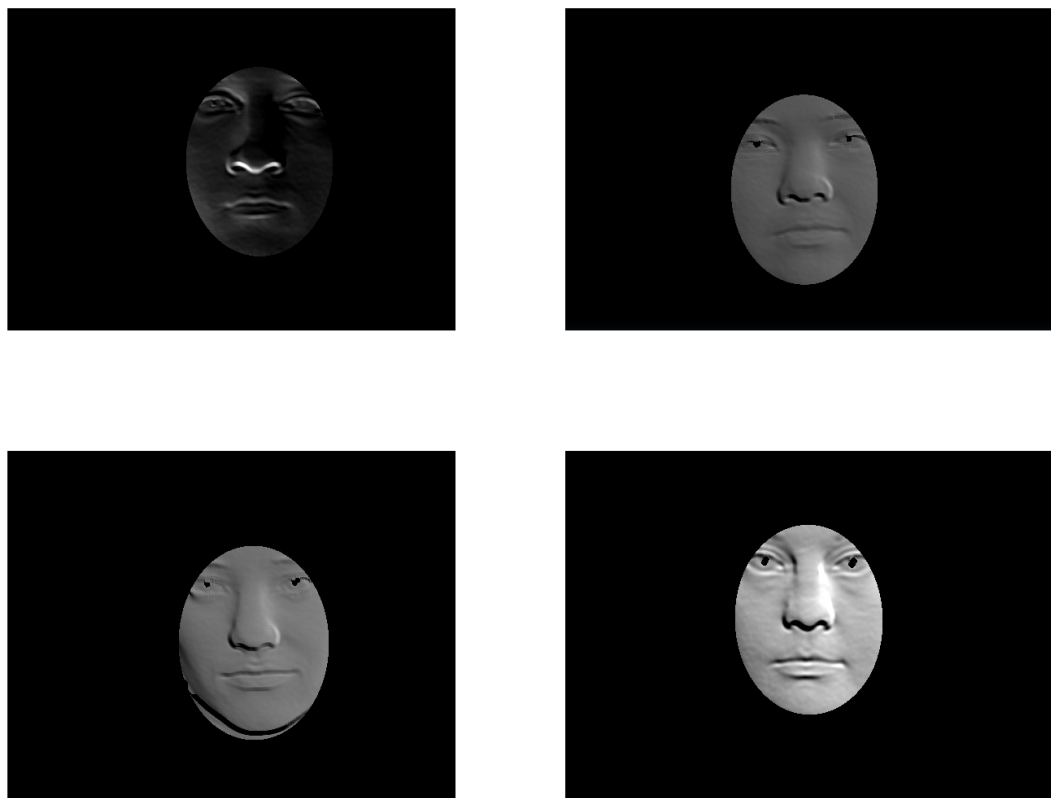


Figure 104: Example differences between an image before and after translation (in all three dimensions)

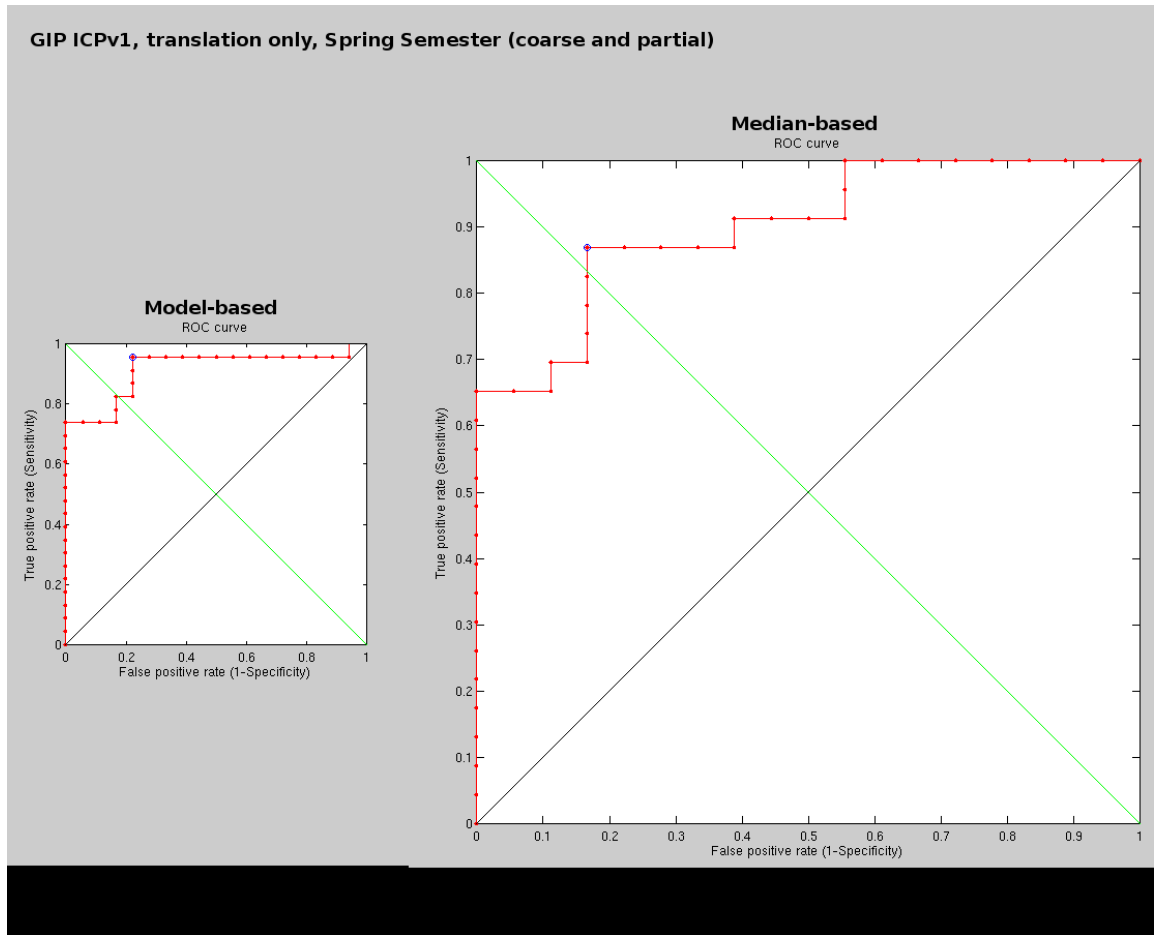


Figure 105: On the left: the results from a test run (first 20 images) using the determinant-based objective function. The model was not constructed with translation, whereas matching did. On the right: the same but with a median-based similarity measure.

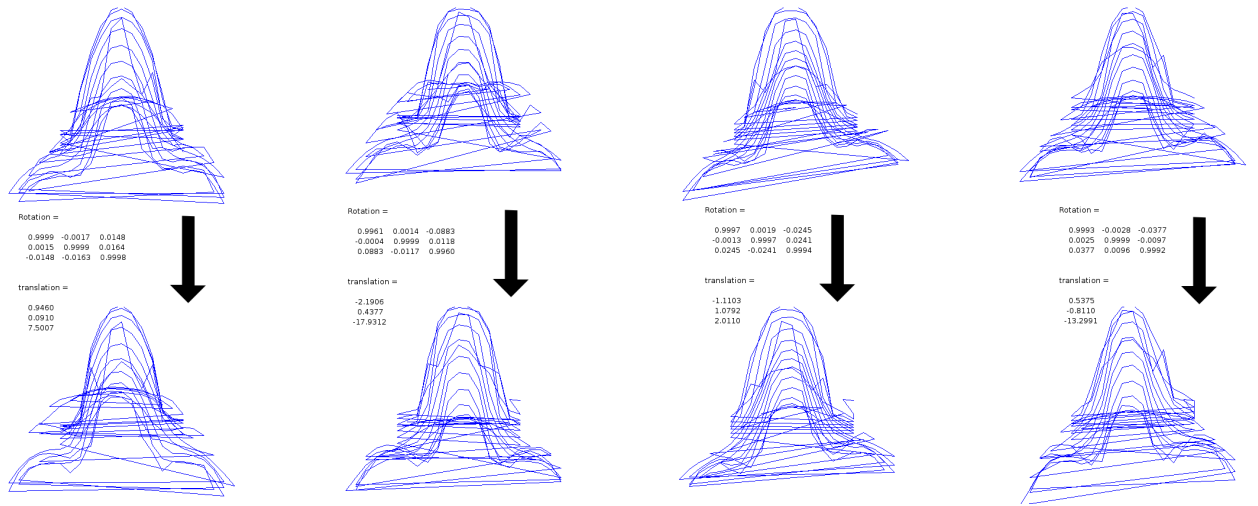


Figure 106: A top-to-bottom view of one's face (the rigid part) with corresponding translation and rotation

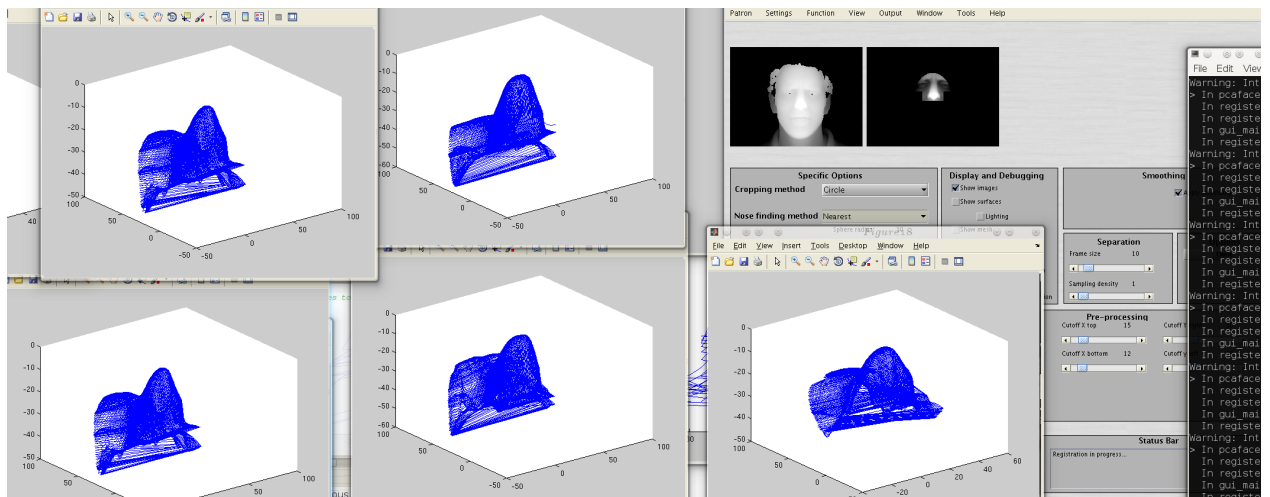


Figure 107: A look at some of the tweaking and debugging process of ICP, where the angle shown is pointing from underneath the nose, going towards the top

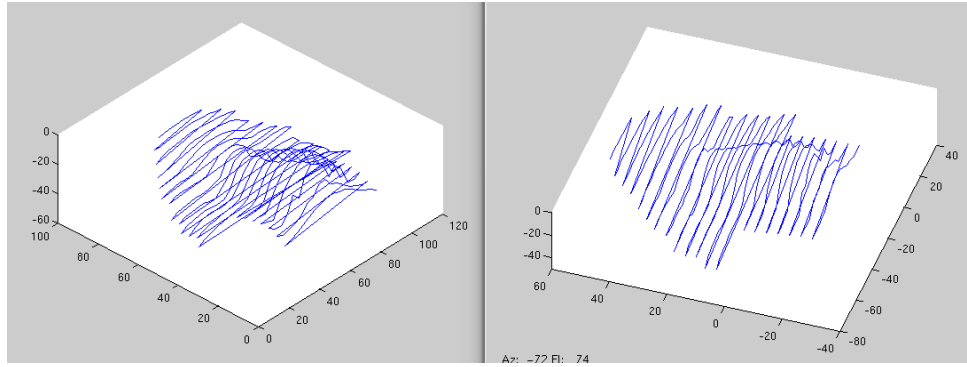


Figure 108: An example of the first image pair, visualised separately as stripes as coarse as the image sampling rate (for the model)

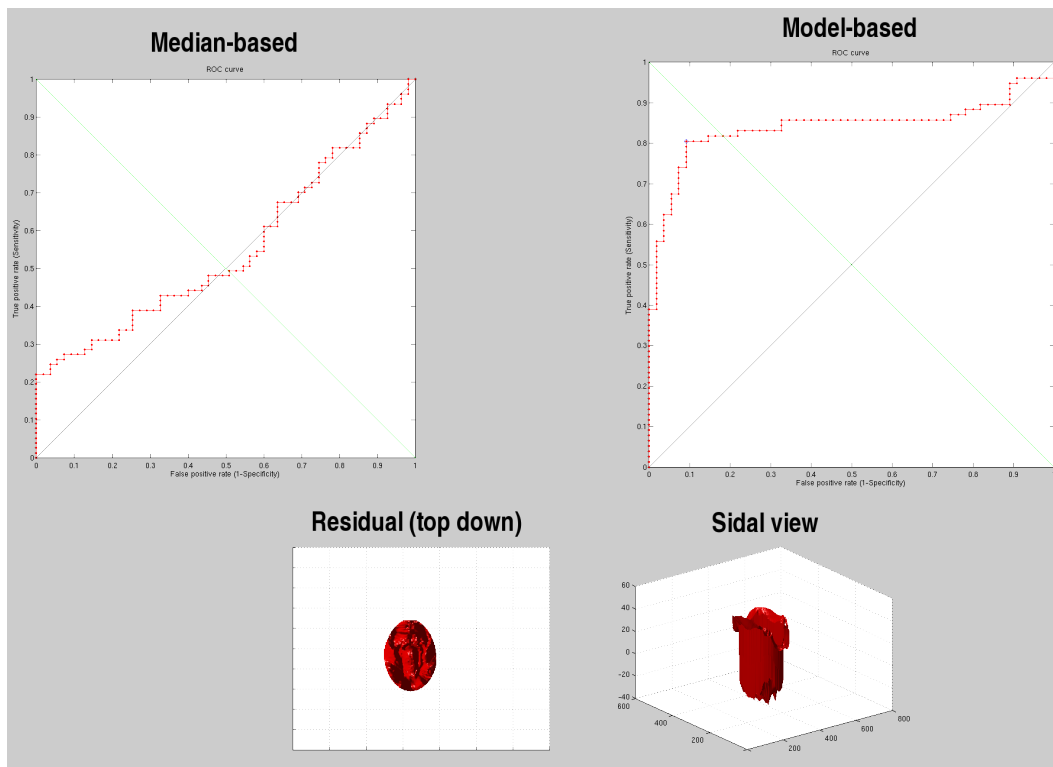


Figure 109: The results of a mis-constructed experiment where ICP did not work correctly and nonetheless, the model-based approach did not fail so miserably

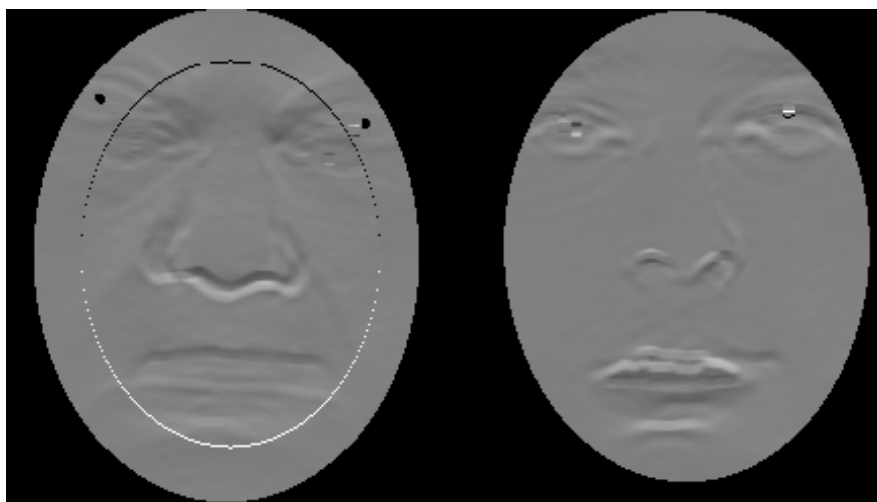


Figure 110: Aligned and misaligned derivative difference (Y only)

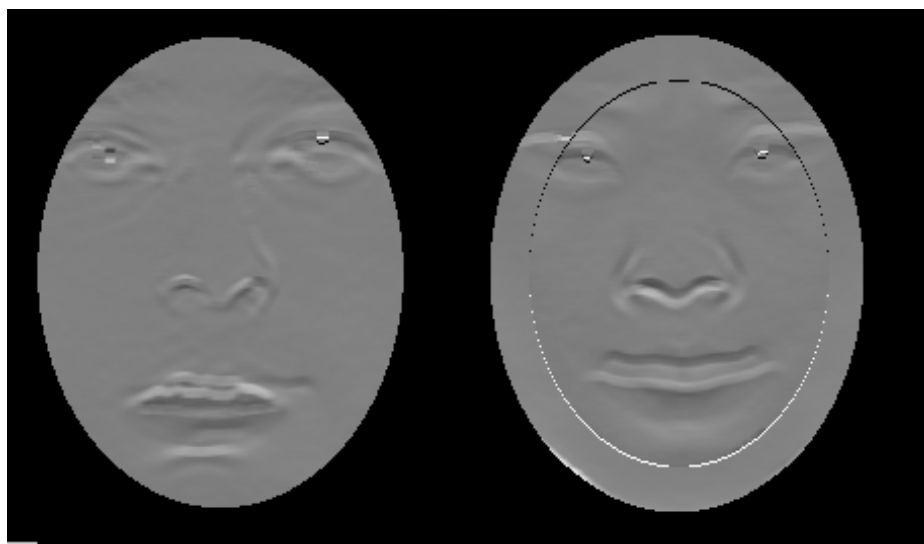


Figure 111: Multi-feature experimental data (y-only derivative)

With previous bugs removed, derivative-based descriptors were used with plain PCA to get the performance shown in the ROC curve (Figure 117). There is some certain correlation between the smoothed derivatives and the Euclidean distanced between points placed on a fixed grid in both surfaces, but there are far better measures that find meaningful correspondence (e.g. areas of high curvature) and measure the distance along the surface on inside the volume.

Taking a similar approach and applying it with robust PCA and multidimensional scaling (MDS) distance matrices, the early steps can involve stress reduction, as seen in the example in Figure 118.

The faces have partial similarity and very dense resolution. We can sample them 10 points apart (as shown in Figure 119), then smooth and triangulate.



Figure 112: Y derivative (left) and X derivatives (right)

By applying these to faces and then building a table of distances (optionally with stressed minimised) these faces can be put in a frame of reference within which they can be compared, e.g. using a variant of PCA.

We need to select a sort of tessellation for triangles that define distances, e.g. for barycentric triangulation of generalised distance maps. Then, finding canonical forms for each pair of faces and matching those forms (or measuring their isometric properties) may help provide ordered measure/s for



Figure 113: A couple of faces with the Y derivative on the left and the X derivative of the Y derivative (result of a bug) on the right

PCA. It's non-trivial where faces do not have geometric correspondences. Experiments were done on some test data where the triangulation is dense and pre-supplied. For partial matching where the number of corresponding points is unknown, ordering becomes tricky. It should probably be safe enough to just sample in areas of interest inside the faces, probably where it is abundantly clear data will always exist, i.e. not near edges of the face; rather, near the centre, the eye, the mouth, and so on.

The picture in Figure 121 could be shown in the form of an animation, characterising the optimisation of point distances and relocations (compare

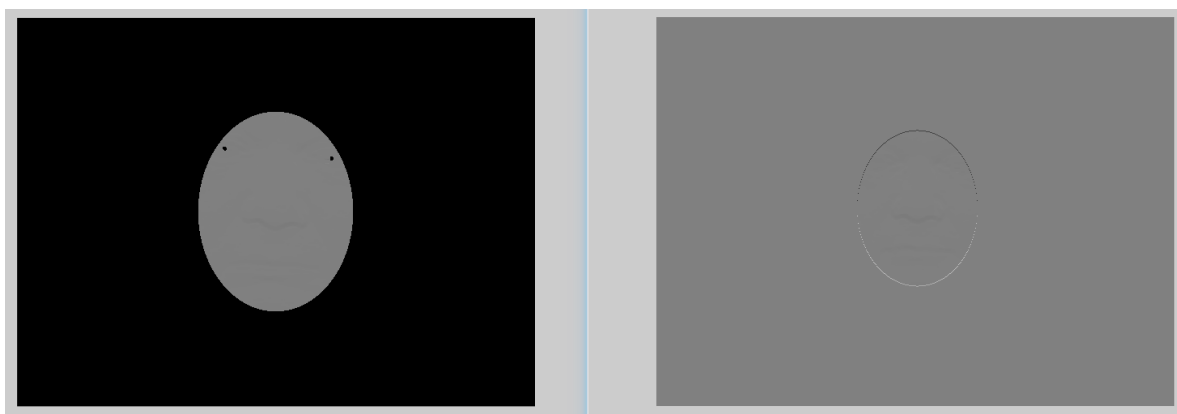


Figure 114: An example Y derivative image before (left) and after (right) signal enhancement

to random in Figure 120). Instead, a curve leading toward convergence is shown along with the starting point and ending point where triangulation is very poor. Ideally, nearby points ought to be connected to neighbours and it is likely that a wide variety of algorithms exist for achieving it. Any preference may bias the results.

7.10.12 Multidimensional Scaling - Animated Example

As a demonstration of canonical forms and stress reduction complemented/guided by multidimensional scaling, we've created an image, as in Figure 122, which shows the process applied to each image in the Face Recognition Grand Challenge (FRGC) 2.0 set – albeit Fall Semester only in this case – in turn, in order to approach a more mutually-isometric and pose-agnostic state where distances are tied to inherent surface details (curvature, size, etc.) and the static image shows the original image too (added at the top). To use this

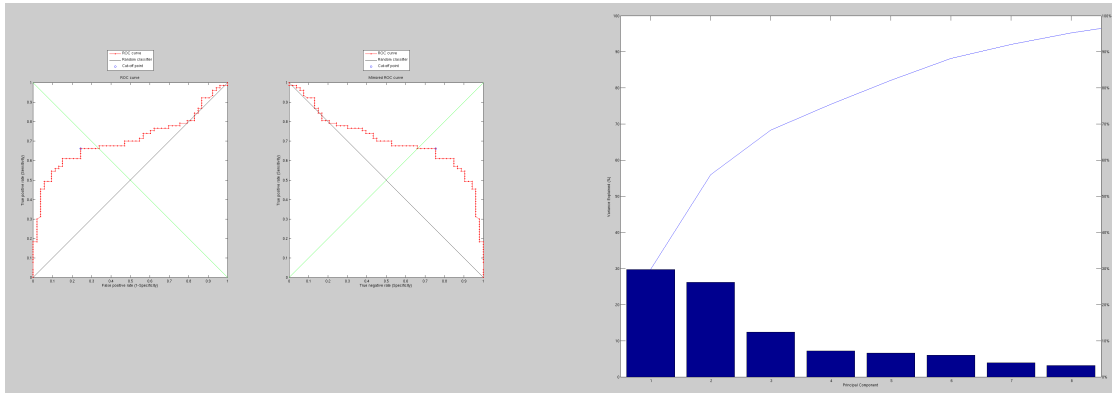


Figure 115: The result of a very crude experiment on Fall Semester datasets, which build a PCA model of derivative differences and then perform recognition tasks on unseen faces. ROC curves are shown on the left, the composition of the model is abstracted on the right.

within an objective function it will need to be clearer how points are selected consistently and where correspondences can autonomously be chosen to improve overall performance. The triangulation in this case is Delaunay-based although 3 methods have been implemented and they offer room for further experimental work. The factors affecting performance may be the PCA component, the triangulation, the placement of points, the optimisation of lengths, the pre-processing (ICP for instance), and few minor technicalities less worthy of consideration. Each one of these represents one parameter among many but feasibility tests – those exploring whether the overall framework is effective in the first place (distances as an encoded signature resistant to expressions) – must come first. Based on a preliminary look, this ought to serve as a reasonable discriminant, but many of the pertinent parts of the framework may need tweaking based on trials and errors.

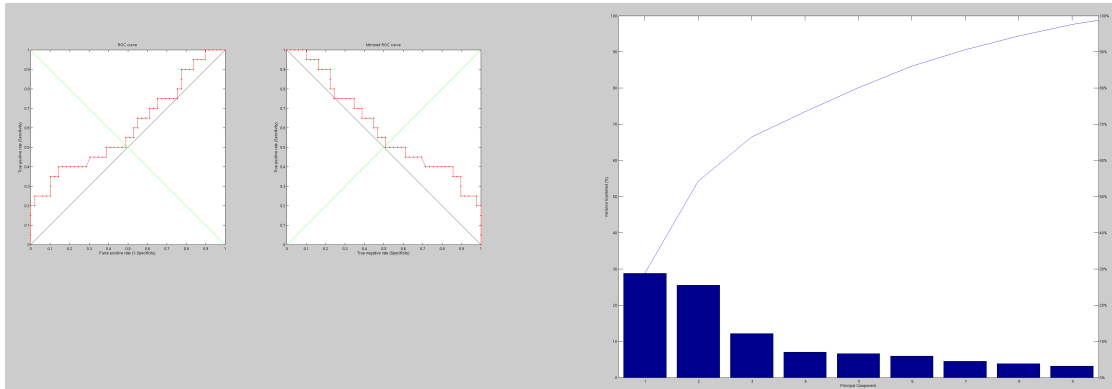


Figure 116: The result of a buggy code creeping into experiment as in 130 (incorrect values were sampled). ROC curves are shown on the Left, the composition of the model is abstracted on the right.

There are some fundamental problems in interpreting how MDS should work. Indeed, one needs to start from the same point (say tip of the nose) and clip a geodesic circle. Then, start from, say the lower or upper point, though it could be better to break symmetry, and use furthest point sampling to effectively sample the surface. It should also lock onto significant features, as those are usually a locally furthest distance from the rest.

Another option would be to map GMDS a given model onto the given shape with key points marked. Or even better, use a flexible mask, like people do in face recognition of images.

After a great deal of additional programming, a framework for multifeature PCA that include multi-dimensional scaling was put in place, along with a point selection mechanism which relies not on a grid but on a set of points in space, which can be selected in various ways depending on positions of

importance highlighted in groups of images consistently. This way, canonical forms get compared, annulling some of the variation which is otherwise difficult to identify and take account of. ROC curves are still unimpressive and the algorithm needs further refinement, with or without making up a hybrid of signals (e.g. distances and derivatives/depths/other). Figure 123 provides insight into the process employed at this stage.

MDS is a preparatory step towards GMDS complexity. It's properly integrated and it has some options for point-selection and other paradigms that exploit a low- but multi-dimensional scaling stage (Figure 124 shows one attempt at aligning images by expanding them to a common frame of reference). GMDS is also put in place for potential work on speed, but this does not get treated the main aspects to pursue, although it definitely improves familiarity with the code, not just the general approach – a generic assembly of methods glued together. In order to get GMDS working within a short period of time, some runtime issues will have to be overcome. To become fluent when it comes to the methods and also the corresponding code may take some time.

While in the process of adding GMDS to the experimental framework which combines it with (G)PCA there are some issues – perhaps easily solvable – making use of the existing fast marching code. In one implementation, the executable is `.mexw32` and in another it's a bunch of `.dll` files. We could find the source files anywhere and this needs to be compiled (unless it is available already) for the Ubuntu 64-bit servers (`system('uname -a')` returns “Linux

```
gipserver 2.6.31-17-generic #54-Ubuntu SMP Thu Dec 10 17:01:44 UTC
2009 x86_64 GNU/Linux").
```

7.10.13 Exploratory GMDS Integration

Code was customised and integrated into the main framework with the aim of putting it in a dimensionality reduction algorithm of another type, alongside signal of nature other than geometric (and geometry-invariant). If done improperly or applied to faces of different people (as the figures below show), it can be demonstrably shown that the resultant correspondence is rather poor. The data dealt with in this case is illustrated in Figure 125. Figure 127 shows this with $N = 50$ and Figure 127 shows the same for $N = 100$. Conversely, as seen in Figure 128, even with $N = 20$ the found correspondence is considerably better *when handling images acquired of the same person*.

Positive pairs/matches are shown in figures 129 and 130, but in the former case (merely the first image in the set) imprecision can be seen, whereas in the latter there is bad data creeping in, leading to serious problems when trying to pipe it into PCA and deal with GMDS as a similarity measure within the larger framework.

By resolving issues associated with fatal exceptions in the pipeline it should be trivial to utilise the generalised MDS, which by far simplifies experiments performed with MDS (still part of the program, at least as an option to be explored or compared to later).

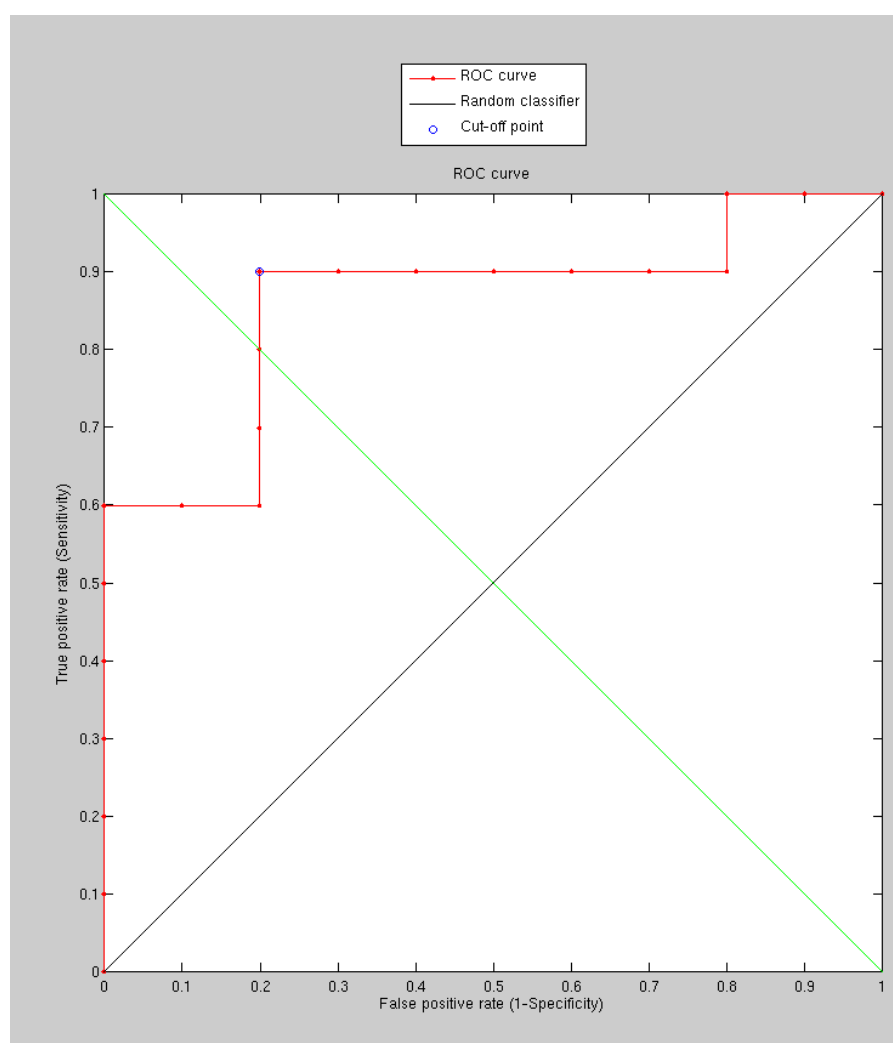


Figure 117: The result of a correct code dealing with an experiment like in 130 but with data from the Fall Semester

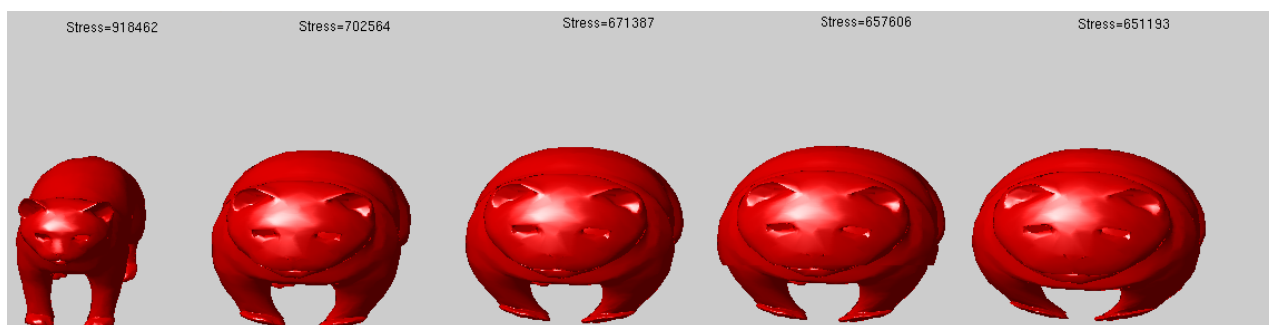


Figure 118: The effect of stress minimisation of the shape of a cat

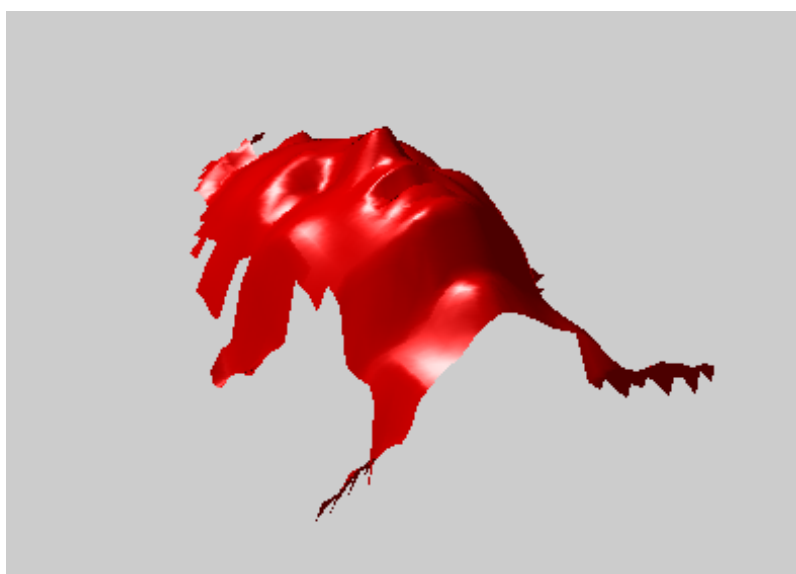


Figure 119: Randomly chosen face sampled 10 point apart along each dimension

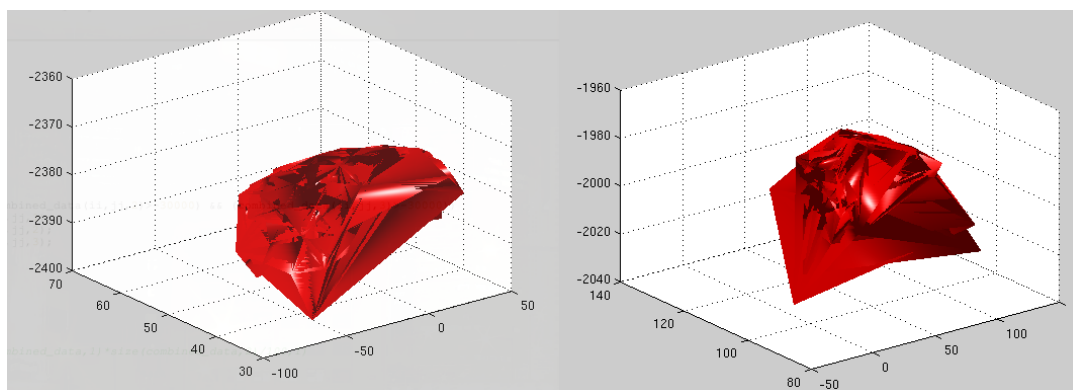


Figure 120: Example of almost randomly selected distances along the shapes

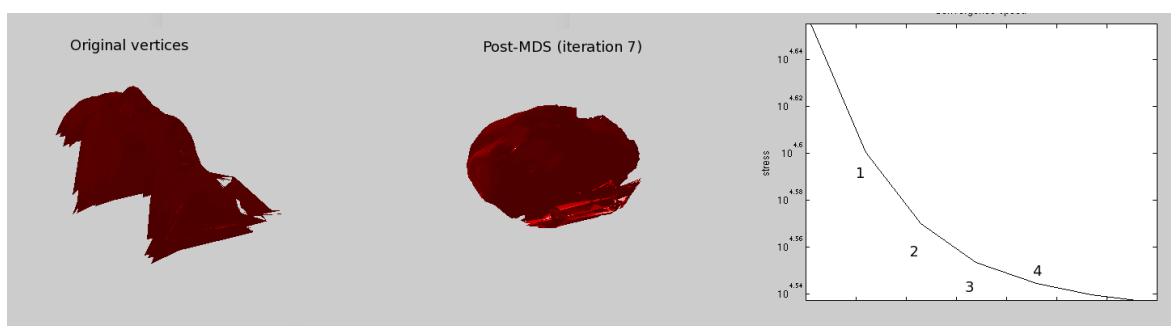


Figure 121: Improved selection of distances (787 vertices) and the effect of MDS reducing the stress

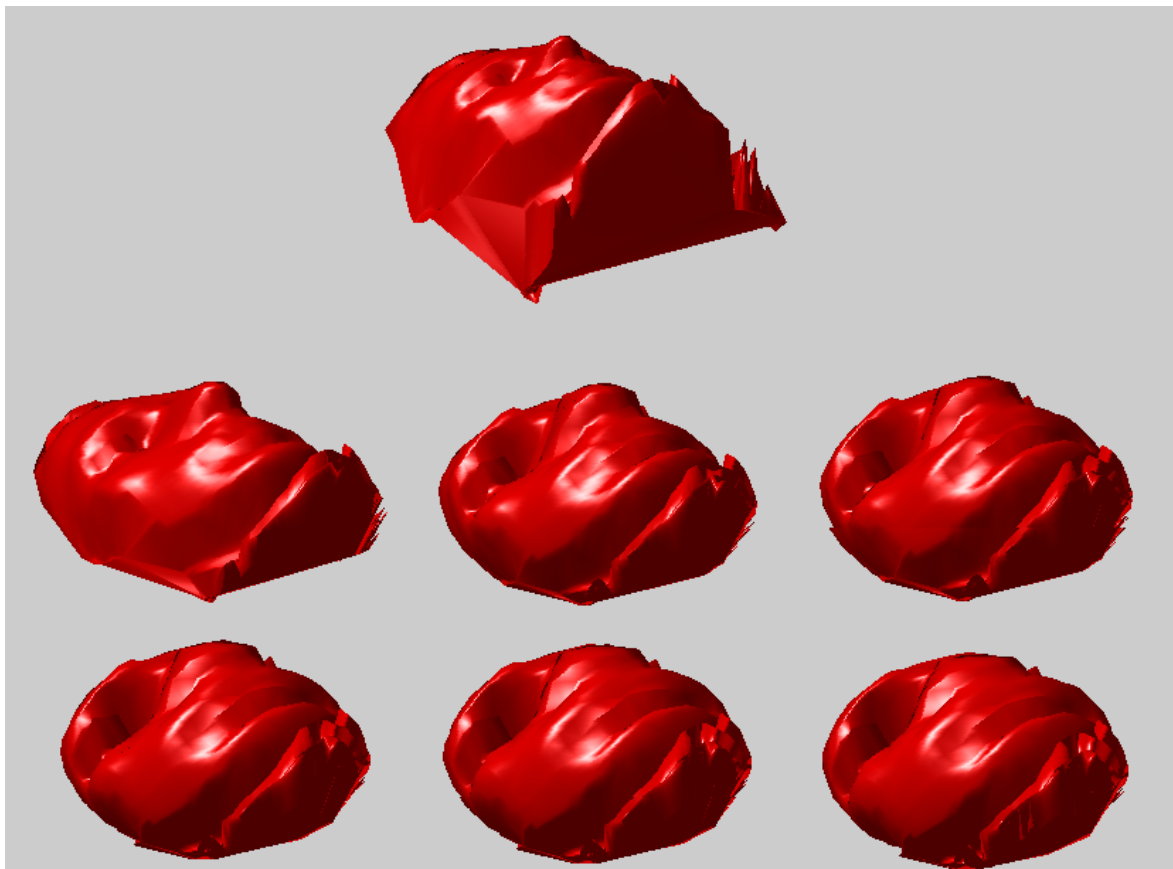


Figure 122: Top: original image. Bottom (from top to bottom, left to right): stress minimisation with MDS, one iteration at a time

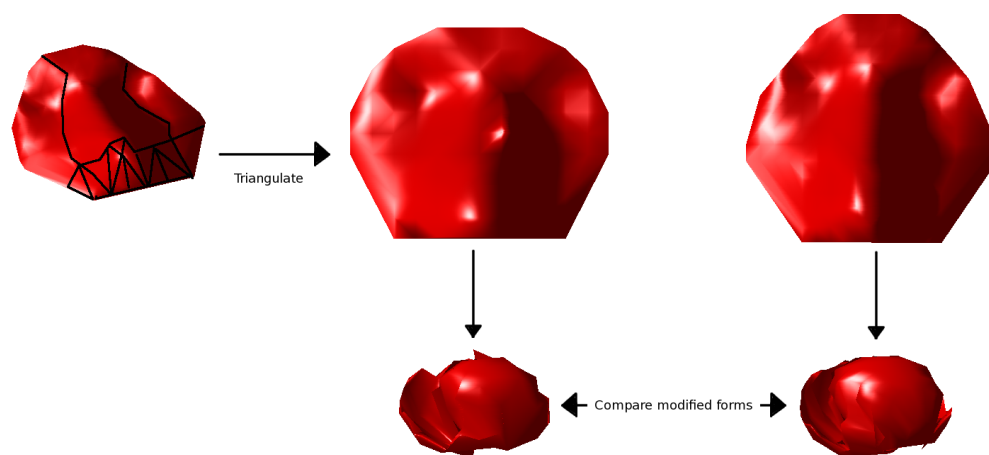


Figure 123: A look at the cruder among ways to perform a comparison between faces

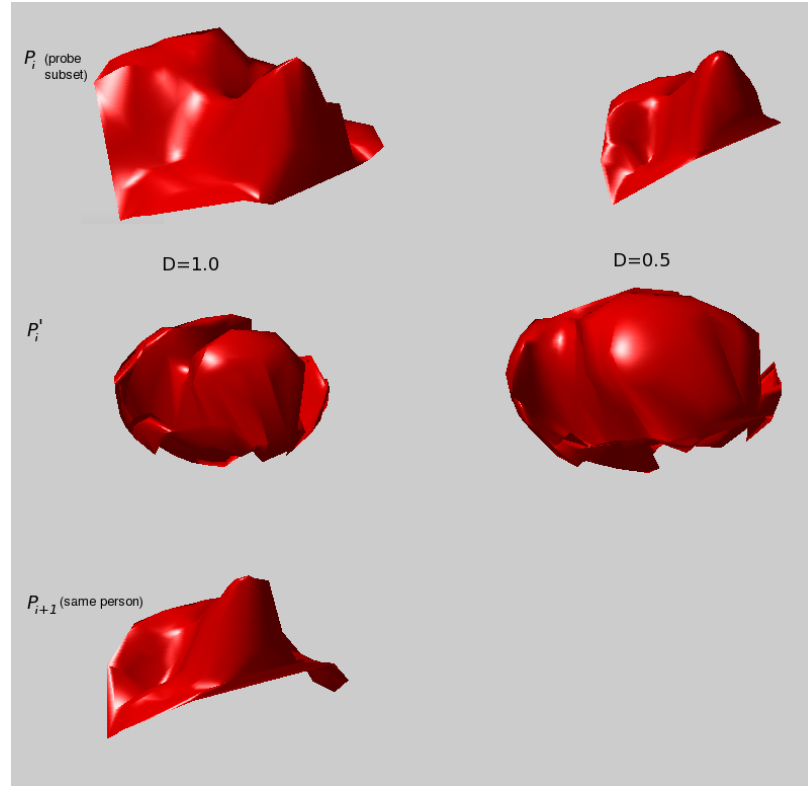


Figure 124: An exploratory look at how applying MDS to face images of the same subject depends on presupplied distances

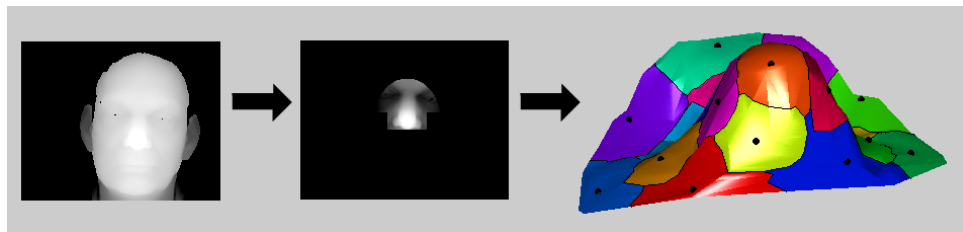


Figure 125: Transformation from 3-D face (left) to a subset of rigid parts and then GMDS handling of the underlying surface (right)



Figure 126: Nose and eye regions from different people (FRGC 2.0) as treated by GMDS ($N = 50$)

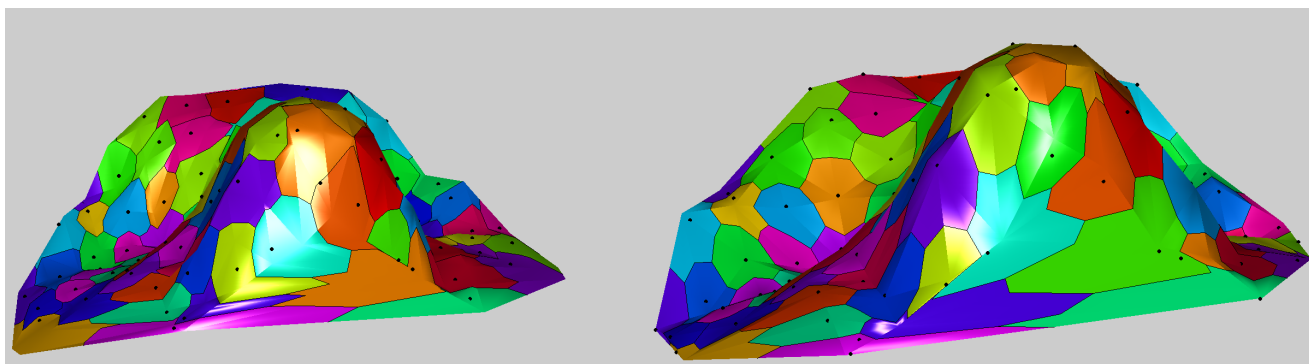


Figure 127: Nose and eye regions from different people (FRGC 2.0) as treated by GMDS when $N = 100$

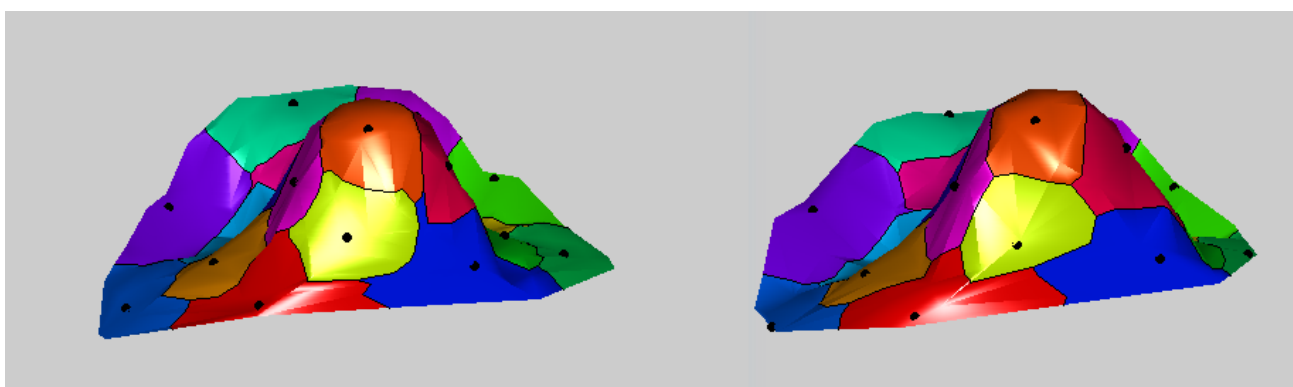


Figure 128: Nose and eye regions of the same person (FRGC 2.0) as treated by GMDS



Figure 129: The first pair in the set of real matches (same person in different poses)

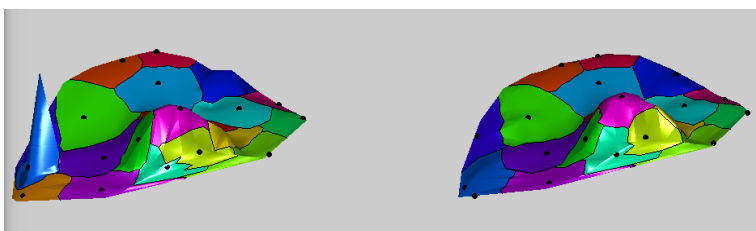


Figure 130: An example of a problematic pair with a false signal spike (left)

Further debugging has facilitated a rather reliable algorithm that is able to assemble GMDS-related metrics (not strictly a metric *per se*) from a large group of images, with or without smoothing and some other parameters that help make the process more robust (e.g. in case of misalignment). While it is possible to derive a similarity measure from raw values without a training process (involving a model), for localised information to bear meaning there ought to be a template or a more high-level abstraction/model that deforms itself to targets or specifies a quality of match. The order of points needs to be consistent with the anatomy and also consistent across examples however, otherwise no consistent markup can be worked on and the discriminant is accordingly weak. Examples of matching between dissimilar faces from different people can be seen in figures 131, 131, 132, and 133.

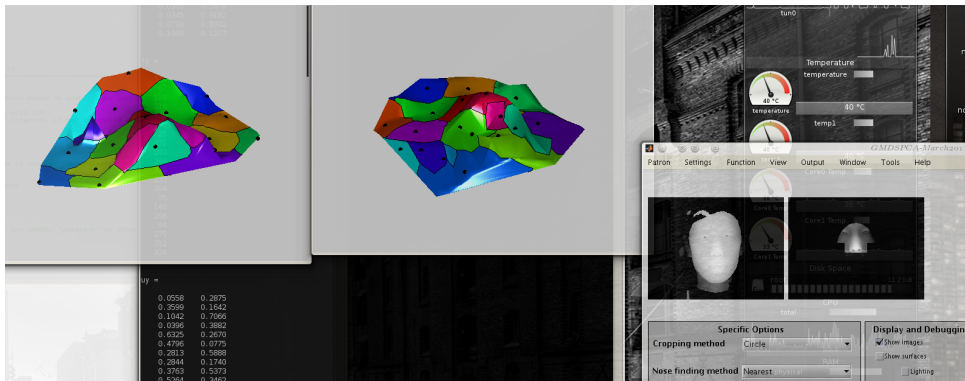


Figure 131: A view of the program's front end (framework wrapper)

Using GMDS, the recognition performance reached at this stage is around 90% (see Figure 135), but there are many improvements left to be made, either in pre-processing or in the suiting of GMDS to the task at hand. The

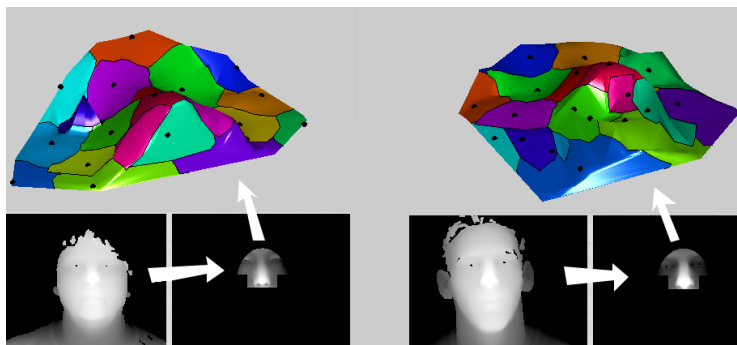


Figure 132: A view of the handling of image pairs and their comparison using GMDS

main barrier was removal of some bugs relating to triangulation, as summarised in very few words in Figure 136, which does not delve into pertinent details as it is uninteresting.

What GMDS does right now is basic and is not yet incorporated with (G)PCA, which would require consistent ordering of points. This is just a set of baseline results to serve as a sanity check.

Regarding (G)MDS versus (G)PCA, it would be reasonable to say that the right mix should probably be some hybrid, where some sort of GMDS is used for alignment (as we do right now) and then PCA for efficient recognition. We are not so sure yet where the line between the two should be, but it is obvious that the truth is there. Figure 137 shows the results from a still-buggy algorithm.

We changed sampling density, changing it from 10x10 to 5x5 grids. Preliminary results on 30 images are as follows:

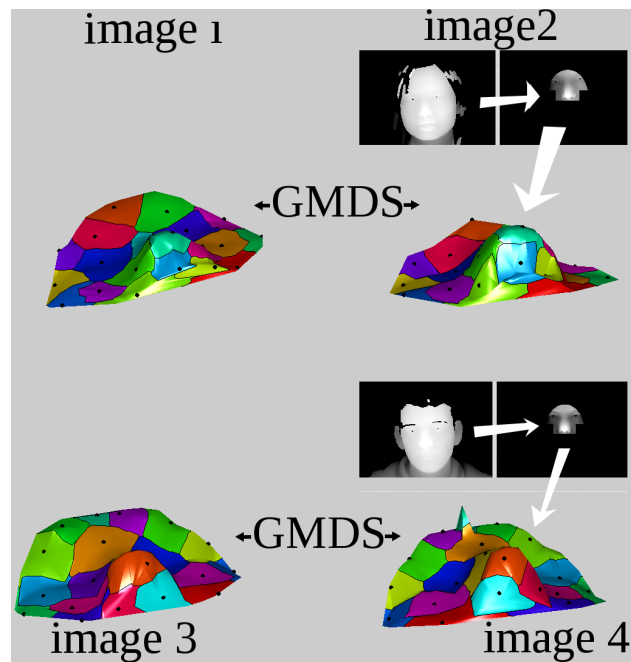


Figure 133: A simple visualisation of the algorithm's processing of images, by numbers

'Predictivity' of negative test (probability that a subject is identical when it is not): 92.9%

95% confidence interval: 79.4% - 100.0%

Negative Likelihood Ratio: 0.1

Accuracy or Potency: 90.0%

Mis-classification Rate: 10.0%

Error odds ratio: 2.1538

Identification odds ratio: 91.0000

As work continues on refinement, it may be possible to find new ways of

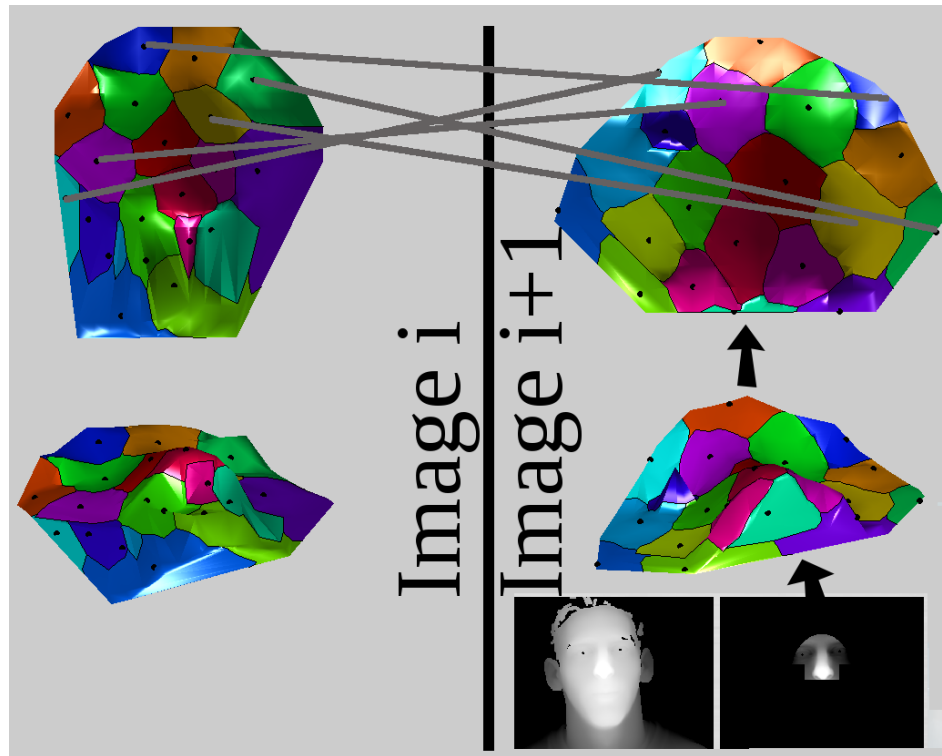


Figure 134: The correspondence problem in GMDS and an abstraction of the data by consideration of a top-down representation

further improving the sampling, e.g. by selecting particular features.

By disabling ICP we can possibly justify the use as GMDS as its replacement, essentially by taking a template image and performing GMDS on it wrt to each image of the current pair. However, ICP should get us a good initialisation for the GMDS phase.

By shrinking the data sampling rate further the recognition performance is further improved to the point where the ROC curve reaches 95%.

Following some further low-level refinements, there is considerably less at-

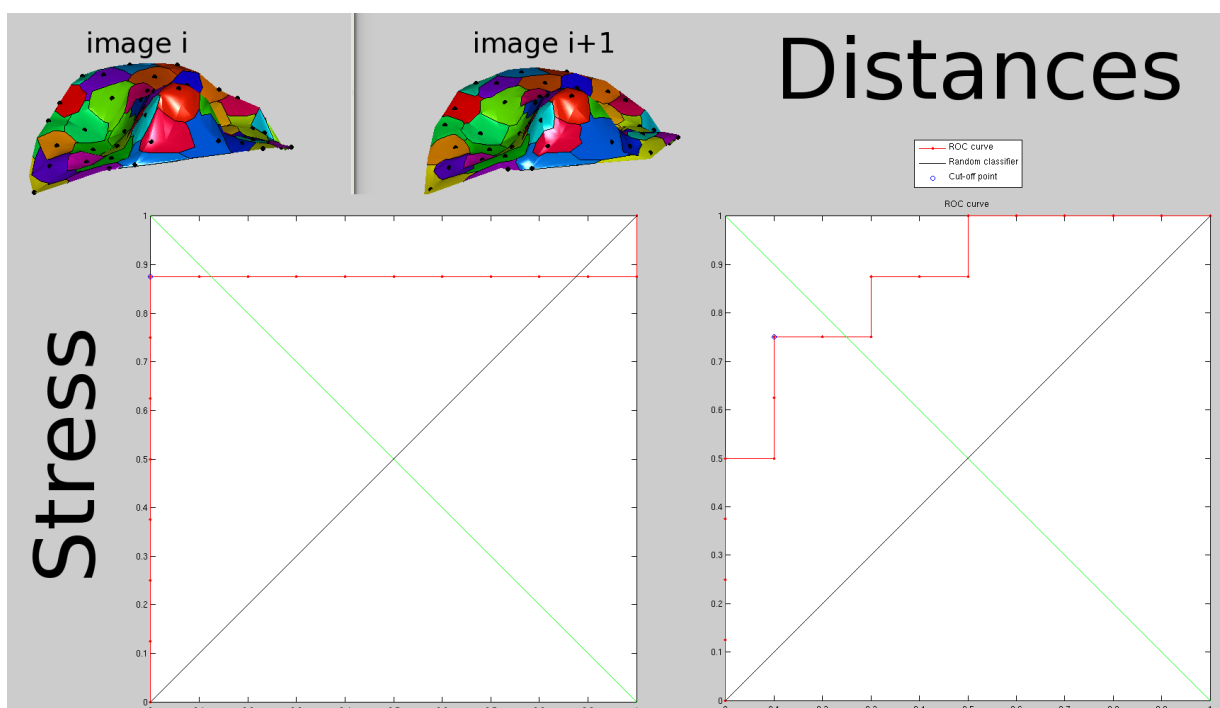


Figure 135: Performance tests on very basic GMDS algorithm applied to rigid face parts

tention paid to minor details around shady areas formerly occupied with voids/holes (a bug with a MATLAB toolbox was also found but not reported after it had wasted hours in vain). This was the result of tedious debugging and tweaking by observation.

This leads to very good detection rates, however nose detection is still short of perfect and provided this can be overcome $\sim 99\%$ of the time¹⁸, matching can exceed 95% detection rate. The FRVT FRGC documents on the Web¹⁹

¹⁸It ought to be a simple problem to fix as it is very clear based on the score whenever bad detection has occurred, the score being an order of magnitude higher than expected.

¹⁹To assess everything more formally, the **Face Recognition Vendor Test** was later on used for reference.

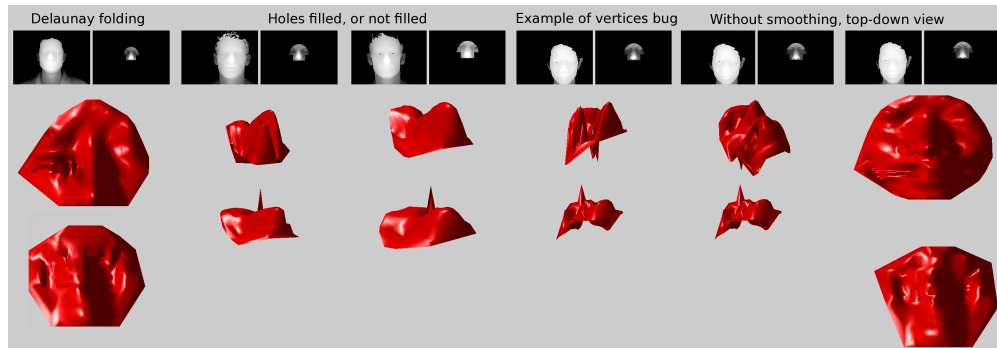


Figure 136: Examples of some of the bugs encountered and overcome while working on GMDS implementation for faces

provide a more formal set of steps to follow, but until the pre-processing stages can be coupled to form a robust enough process, there is no point to adding PCA variants to the pipeline and then performing benchmarks. The pieces are already in place, but it is the failure to accurately and consistently carve out faces (despite hair occlusion) that merits increased attention and effort. In the latest small test involving 30 correct pairs (same person) and 30 incorrect pairs, the only misdetections were due to arbitrary face parts being assumed to be nose, incorrectly. The reasons vary and solution has been found and implemented many times before, encouraging reuse now rather than a reinvention of the wheel. See figures 138 and 139.

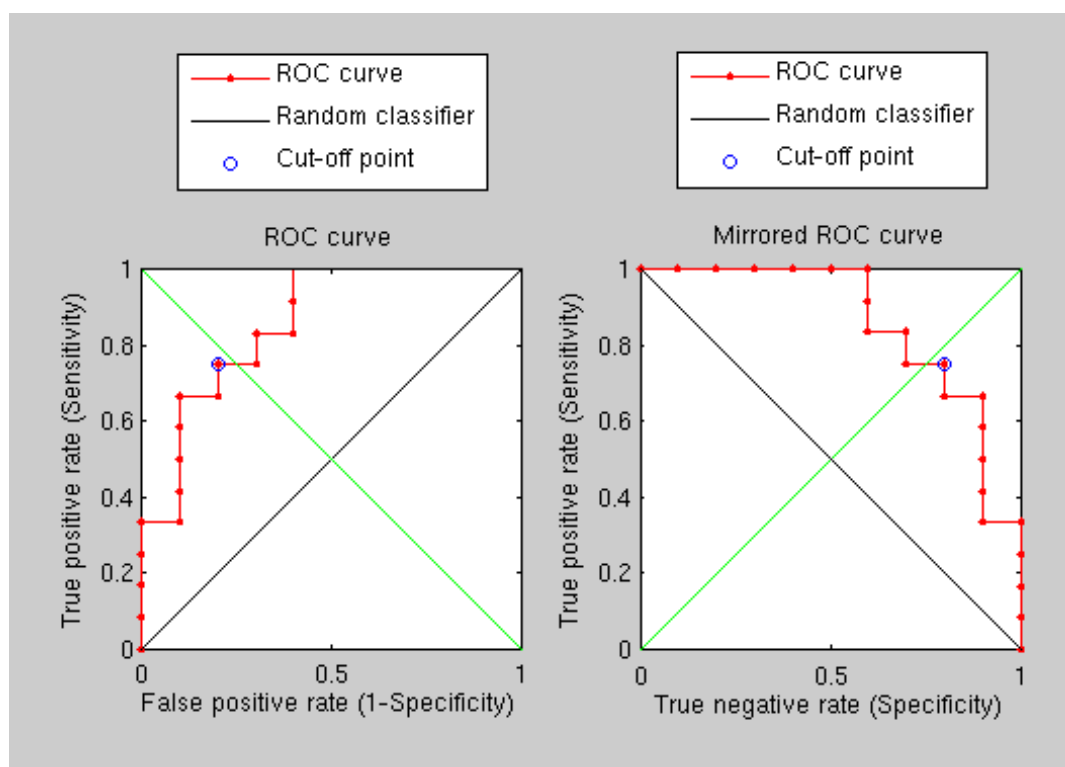


Figure 137: Set of results for 10x10 grid sampling (GMDS)

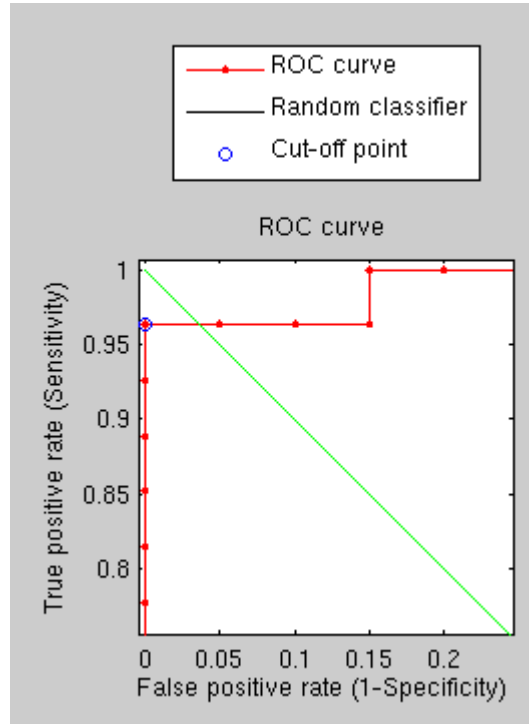


Figure 138: Early performance measures for GMDS more properly done

Following some preliminary overnight experiments, it is possible to show the practicality of a PCA-GMDS hybrid framework, wherein the values on which dimensionality reduction is invoked are the geodesic distances between salient points. The idea is, by studying the variation of distances between analogous facial landmarks – almost as though there are strings between every pair – one can know which ones are expected to vary not across people but within them (intra-person/intrinsic), in which case these variations are very much expected and predictable. The model which is built only from correct pairs (8 pairs in an initial toy example, 76 in the coming tests) is supposed to penalise for variation in areas of the face that do not exhibit much variation

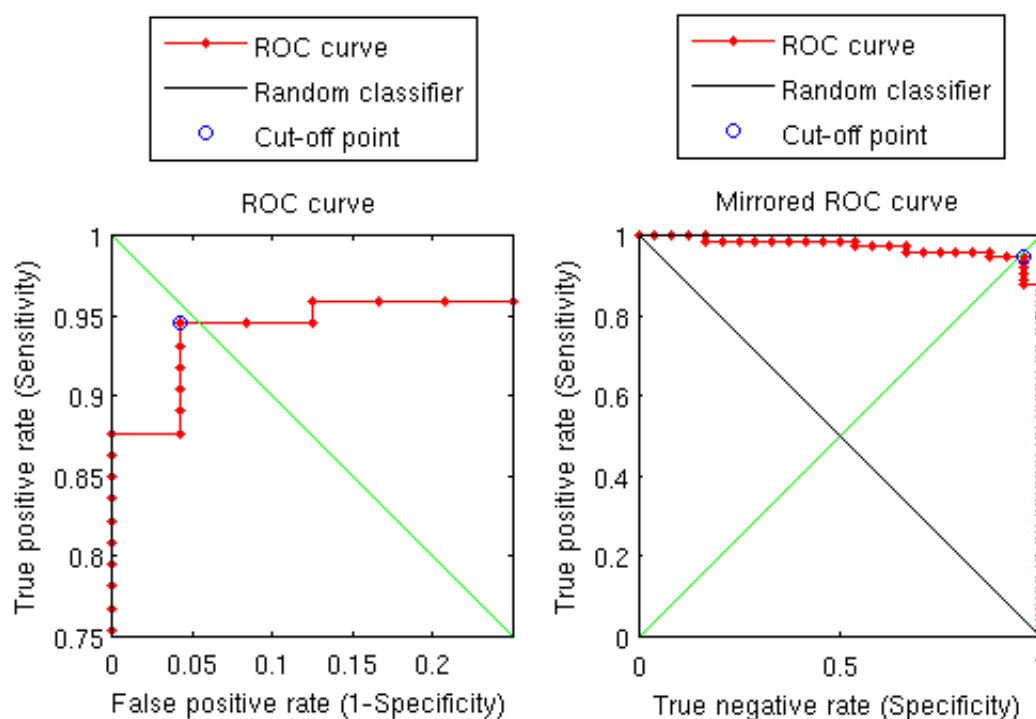


Figure 139: Larger scale examples of early performance measures

in the training phase. Results are shown in figures 140 and 141.

The subsequent steps delved into ways of improving the data and its preparation for classification for an accurate determination of match/no match status. While in principle the method works quite reliably, a lot of room remains both for improvement in the ordering of points and in the quality of the pre-processing, as most of the false positives and false negatives are a result of the latter. Additionally, removal or conversely *proper* sampling of

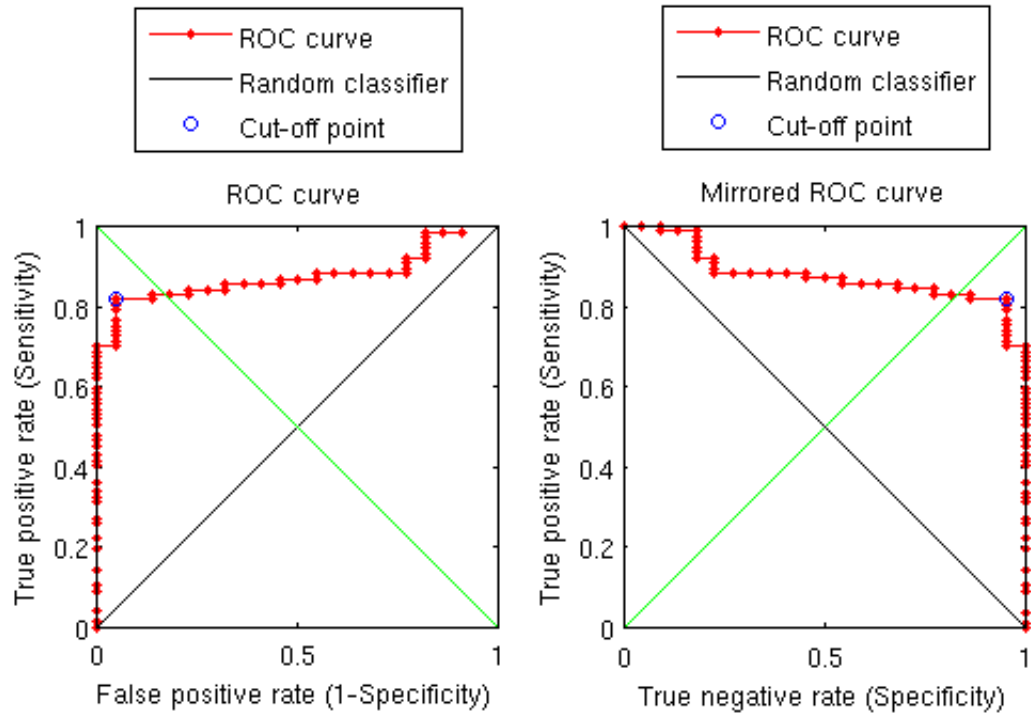


Figure 140: Results from poor PCA model, obtained using GMDS

points around the cheeks should be considered.

The charts in Figure 142 show the distribution of mode weights based on the building of two models, one of 10 people (around 80 pairs), and one of 76 people (around 400 pairs).

We then prepared a short report for a decision to be made regarding how long we give this face recognition project, which could otherwise be morphed to measure distances on a surface where corresponding points can less effectively

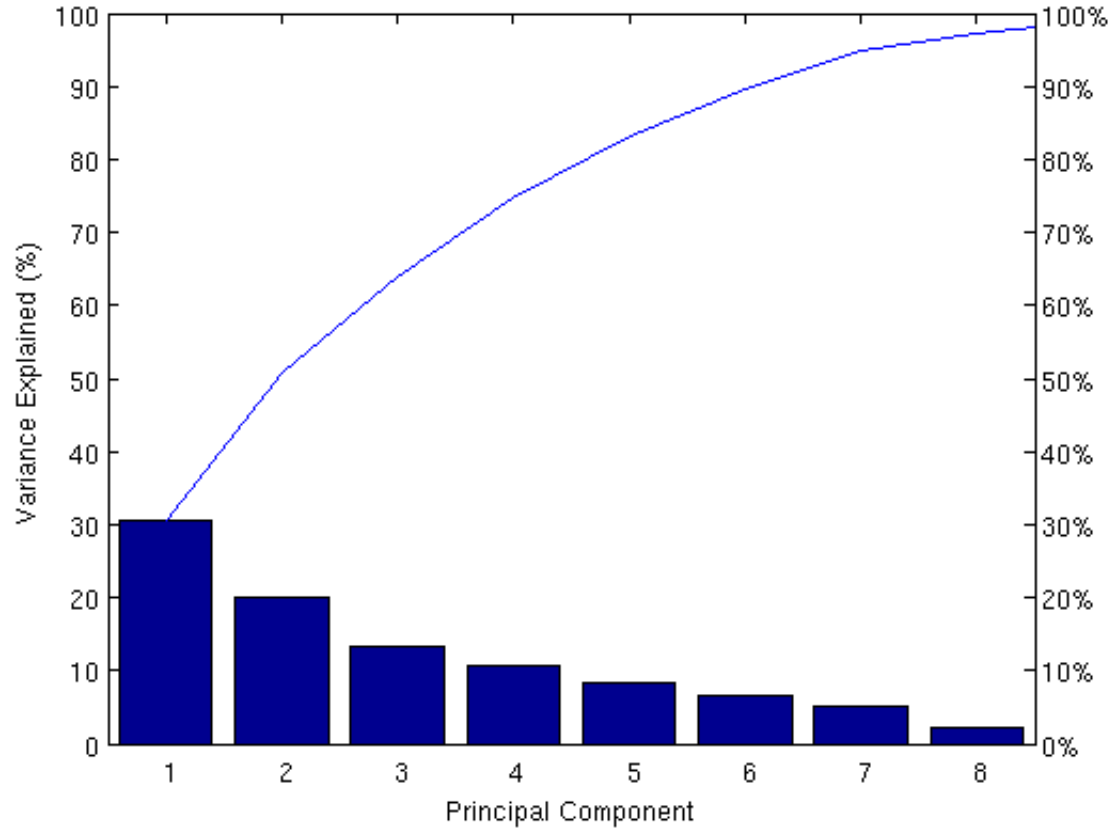


Figure 141: Model modes distribution, corresponding to Figure 140

be identified, e.g. anatomical parts inside the body where there is no easily identified part such as the nose, mouth, and eyes, let alone any photometric data to take advantage of. The strength of GMDS is that it autonomously finds points that are otherwise difficult for humans to mark up.

How the current results compare to the scores reported in FRGC FRVT etc. is still an important question and we can we combine mine with Bar's code for improved performance based on prior work. We can work effectively from

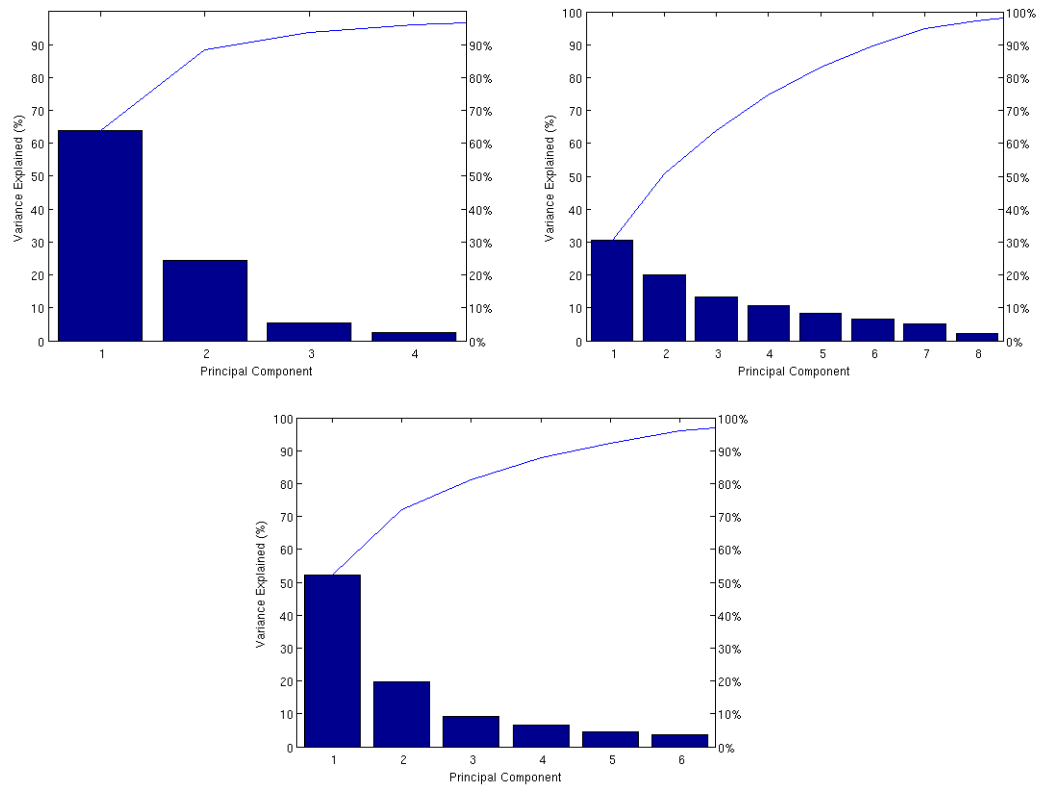


Figure 142: Model modes distributions (of 10 people and 76 people), built with the proper weight, albeit with very heavy and sometimes excessive smoothing

a distance because there are fewer distractions. In general, the bottleneck is pace of work (about 2 hours per day), but the intervals allow for more results to be processed and delivered in-between. Since a lot of the work is done on computational servers anyway, locality has access to informed people as its main advantage. The weakness of work for long periods of time is that time taken for results to arrive must be dedicated to observation or further coding, which would still depend on the observation of results that had not

arrived.

There have been no known attempts to apply GMDS methodology for diagnosis based on deformable atlases (training from patients with atrophies compared to normals). Half a decade ago, Davies, Cootes, and Taylor used reparameterisation on the sphere (Cauchy kernels) in order to classify the 3-D shape (surface, not volumetric) of the hippocampus with the aim is diagnosing disease characteristics of this interesting structure (with known correlation to illnesses), based upon fully automatic training from datasets we may have access to. The work done by Aflalo *et al.* is reminiscent the above, at least from an analytical angle.

The first to use conformal maps for "computational anatomy" is probably Eric Schwartz in the 80s. The more recent examples that immediately crop up come from "[MICCAI 2008 WORKSHOP ON THE COMPUTATIONAL ANATOMY AND PHYSIOLOGY OF THE HIPPOCAMPUS](#)". Xie *et al.* [86], for instance, use shape analysis for Alzheimer's Disease detection.

A. Elad used MDS to map surfaces to spheres. It was around 2002 as far as I recall, but it was definitely not conformal. The mapping to the sphere in Davies' case (his work is still ongoing, but he too only spends only about 50 hours per week on research) is one that warps correspondences onto a sphere (or circle, at least in 2-D) and then applies particular functions to space up the correspondences and make reasonable candidates over which to optimise a groups shape concurrently [27]. The overall goal is to automatically identify

and choose points that represent shapes. My own work extended these ideas to full intensity (texture), seeking points that take both grey-level and spatial values into account at the same time (using a combined shape and appearance mode, or "AAM"). I published papers on the subject over half a decade ago.

If it is true, as claimed by several people whom we spoke to, that face recognition is best handled by carving out few features that never vary in their relative geometry, then GMDS seems a little unnatural as the only absolute points on which to measure distances are easy to identify either by hand or by template (colour can help too). The continuous mapping that depends not on interpolation but on surface characteristics like curvature or distances on surface may be inadequate (an overkill) unless only few fiducial points whose location can be determined accurately get used. This point is worth getting across when GMDS is criticised for utility in face analysis, wherein simpler algorithms can outdo it.

One would completely agree with the observation about GMDS if indeed faces had been rigid. They are not. This is especially valid if you take the face as a whole and just crop out the mouth. Still, cropping only the upper "mushroom" part and considering close to neutral expressions, then, ICP alone could be enough. one would guess that GMDS could enhance it by a small notch, but this may be wrong.

ICP appears to be essential for improved initialisation of GMDS. It is important to be clear about whether we wish to model/sample entire faces with

GMDS or not. The common facial expressions can lead to degradation in the results, but then again, with PCA these ought to be weighted accordingly, e.g. with the expectation of large variation (an already-seen variation, owing to the training set) in particular regions, whereas other regions remain stable, i.e. distances within those regions hardly vary or alternatively vary only along particular dimensions (in hyperspace of M^2 dimensions, where M is the number of points, not in 3-D). I will prepare an experiment which broadens the scope to entire faces. It oughtn't yield good results (on a comparable scale), but at least from an academic/scholarly perspective it ought to validate the inclusion and contrariwise exclusion of particular parts, e.g. those that accommodate mustaches and caused detection problems in previously-run large-scale experiments. Likewise, a Euclidean versus geodesic benchmark (Gaussian fitting for instance) can be produced to provide validation, similarly to the preparatory work from the 2006 BBK paper in IEEE TPAMI. If it can be proven – empirically – that geodesic distances always trump Euclidean equivalents, then at least in the case of 3-D it can be argued that all those leading algorithms (claiming 99.9% accuracy) can be further improved with FMM. Bar Shalem's work partly applied some of the same principles but fell short performance-wise. It is therefore unclear what paths should and should not be explored. By applying GMDS with just 5 points (classically the eye corners and the nose) we might be able to attain good performance but also merely replicate previous attempts by Bar Shalem, thus studying too little. This is why, upon the inquiry about code fusion, I remained a

tad reluctant. To what extent, for example, were the algorithms tested and then refined? Was the newer version of FRGC tested on as well? Since we have got access to code from BBK papers on face recognition (2005), which route would be better explored? How many parts are merely reimplemented. Anastasia has argued that GMDS, as a black box, has not really changed since 2009, so the other building blocks are probably the only candidates for swapping.

Our job is to prove or disprove this issue which involves feasibility. Starting point should be state of the art ROC curves. This is hopefully a reachable goal. If "state of the art" is now an error of 1 in a thousand or thereabouts, then it seems like a monumental task.

ICP could be interpreted as a Gromov-Hausdorff distance when the inter-points distance is Euclidean and points are allowed to move in 3D. It would be interesting if coordinate-wise descent could work as well as ICP (one may doubt it, though using multi-grid it could actually work). So, GMDS could in-fact be used like ICP. Therein lies a possible micro-study which compares the R and T matrices that our 4 (currently) ICP methods output, perhaps rationalising the use of GMDS for alignment. Alternatively, it ought to be possible to compare recognition results with and without ICP as a peripheral/separate part from GMDS.

Regarding the comment about existing GMDS implementation and its age, it is likely that Carmi Grushko introduced some changes to the GMDS, and in

fact he is currently working on further refinements (of the geodesic distance computation).

A different measure to try is using diffusion distances rather than Euclidean or geodesic. One could also consider diffusion on the surface, diffusion "inside" the surface, as well as geodesics in the interior of the face, etc. One distance should provide the best discriminative power among all possible ones. We must check it.

The current experiment deals with the performance reached by adding and removing parts of the face, using binary masks that make very basic sense. In all cases, depth values from X and Y (averaged over each grid) are used to scale the binary mark, such that consistent cropping is assured regardless of distance from the camera's aperture. This is one of the crucial areas of improvement, one of about 6 areas that need further improvement.

It is agreeable that 1/1000 is a challenging goal, but one may strongly feel we could get there, and then just play with building blocks to check which metric gives the best results. The hunch is that geodesics should play a leading role there. Either as dense or sparse matching of surfaces.

How would geodesics deal with eye sockets? The problem is, with the eyes being filled the signal is too noisy and without any filling there is a difference in distance/s which depends on how open the eye is. Euclidean distances do not suffer from this apparent drawback. One solution devised so far is almost excessive smoothing, whereby just the very basic geometry is preserved and

a lot of the rest vanished out of signal. The fine details are unlikely to be present in different acquisition sites/times.

7.10.14 Full-face PCA

Shown in the image grabbed for Figure 143 are the results one gets from applying GMDS with default parameters to datasets comprising a variety of expressions and no constraints on scope, except exclusion of non-frontal face parts, including the neck, hair, ears, etc. As expected all along, the performance takes a noticeable hit. It might be interesting to see what putting/piping the distances through PCA will do to overall performance, at the very least on a relative scale. It might also be interesting to see what performance we get by just looking at the eyes and nose in isolation, perhaps LDAing them having used photometric data for segmentation and then applied GMDS several times. If we had decided to limit the measurement of geodesic distances to only particular segments, this would be simple to implement.

With some new results from overnight experiments, it seems unlikely that adding the cheeks will improve performance much, to say the least (it is too inconsistent there). The current line of work looks at piece-wise GMDS, wherein facial features are taken in isolation to see the discriminative power of each.

Regarding eye-sockets., it is worth thinking about measuring local distances

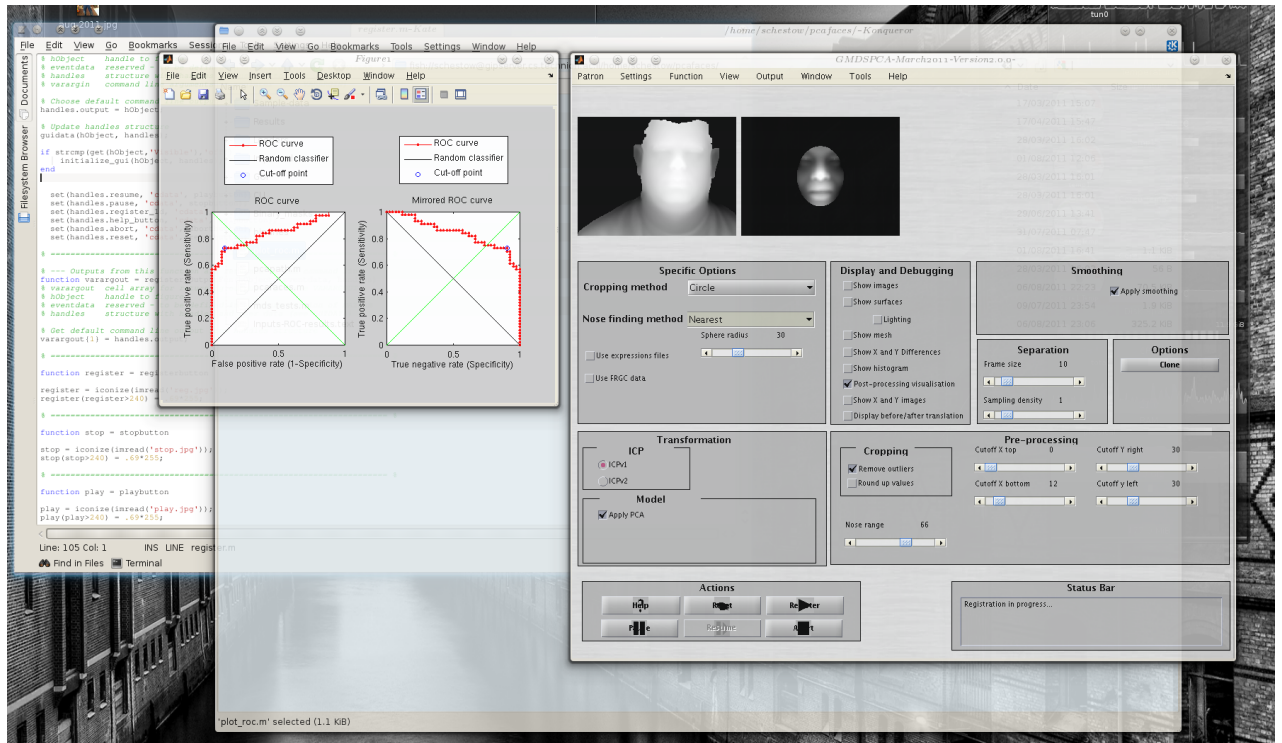


Figure 143: Preliminary results from GMDS-based recognition with full face surface

in a Euclidean fashion and longer ones, the geodesic way. This might be a "best of both worlds" approach, assuming of course that best detectors use the former method and the latter can complement it.

Alternatively, we can just be switching to Euclidean at regions of suspected peculiarities (missing parts) and large depth variations assuming the feature detector could isolate the feature points accurately. This is probably where using texture would be helpful. There must be existing implementations for segmentation of the face based on intensity data alone. The results with all cases really look bad thus far.

7.10.15 GMDS on Smaller Face Parts

Results from additional new experiments are shown in figures 144, 145, 146, 147, 148, and 149.

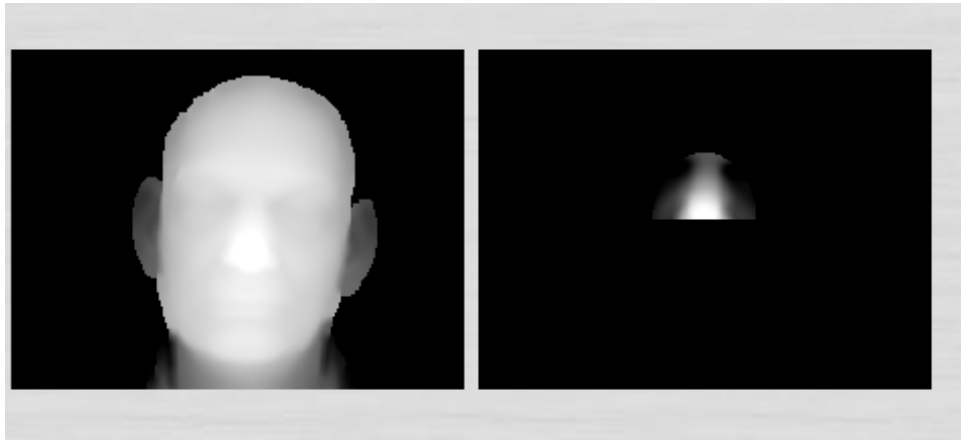


Figure 144: A look at an alternative mask which focuses on the nose and inner eye only

It seems that we have most of the components to have a perfect system, but maybe the MDS implementation is not by the book, as the results are not as one would have expected. We are aware of half a dozen deficiencies and will address each one of them in turn. It is also apparent that we need to take into account special cases that recur. Figures 150, 151, 152, 153, 166, 161, 156, 157, and 158 show the results of some further debugging and gradual tweaking.

Overnight, large experiments were run for 6 hours, flagging quite clearly all the cases that remain problematic and need closer attention as the false recognitions generalise to other examples of their kind. Some mistakes are

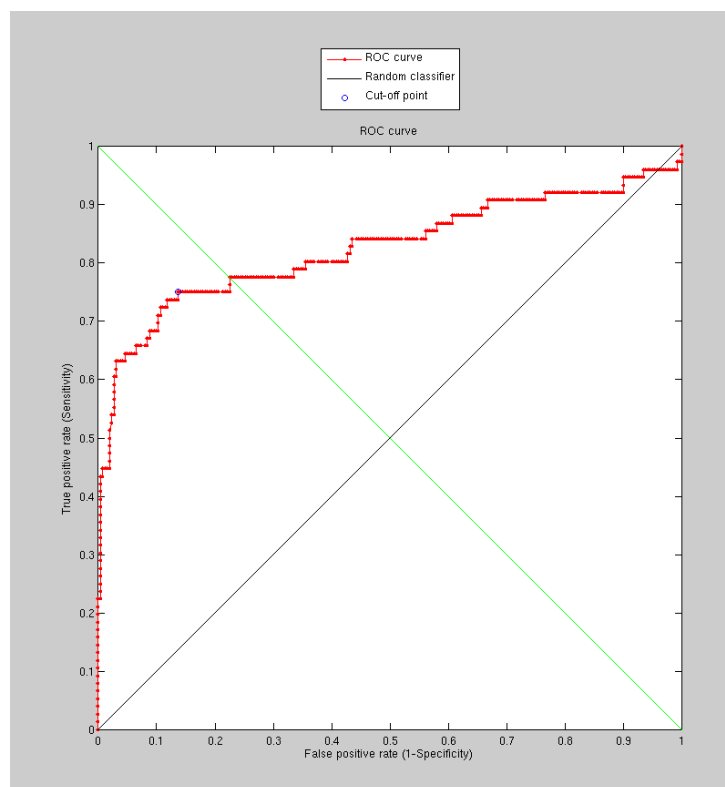


Figure 145: Recognition results based on the mask from 144 with GMDS

caused by bugs in the code, especially in situations like special cases or bad data.

For images with only minor expressions we still hover at over 95% recognition rate. The problematic case are ones where the variation is great (between semesters for example) and there is partial matching in need.

We're working our way up, gradually improving performance by identifying edge cases and addressing them with some more sophisticated and problem-specific code which in turn generalises to more images exhibiting the same

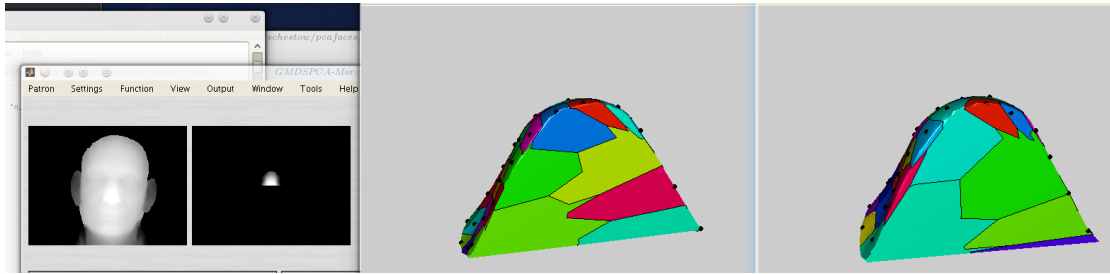


Figure 146: A nose-only mask, which omits areas with potential of facial hair (the examples at the centre and the left are not related)

problem. Some examples of the progress are visualised in figures 159, 191, 161, 162, 163, 164, 165 166, 167, and 168.

7.11 Texas Database

The Texas database is now ready for use (Texas3DFR Database) to get the data. The data must not be used or copied for any usage other than our academic research. This dataset ought to be very suitable for our needs because it is prealigned rigidly, which removes some of the issues encountered so far. It also makes ICP-agnostic comparisons (based on non-rigid recognition alone) easier. Their dataset was used by several US universities (but not many) and it is apparently much higher in terms of its quality. They are still trying to get more groups to use it. We had been waiting for a response regarding access to the new files. It was work in progress. It would be great to see performance attained from pre-aligned data. A lot of the current difficulties are associated with this drawback.

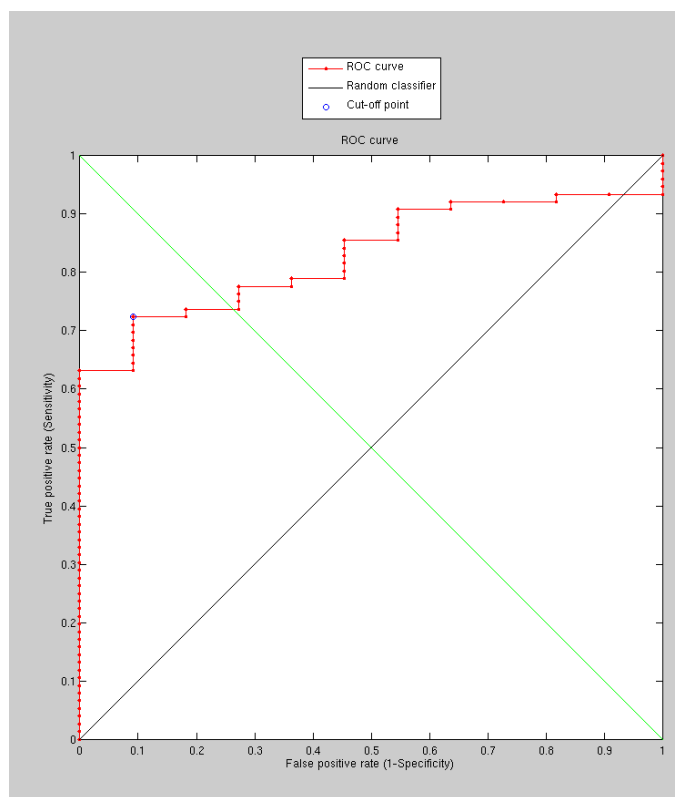


Figure 147: The performance attained by applying GMDS just to the nose region

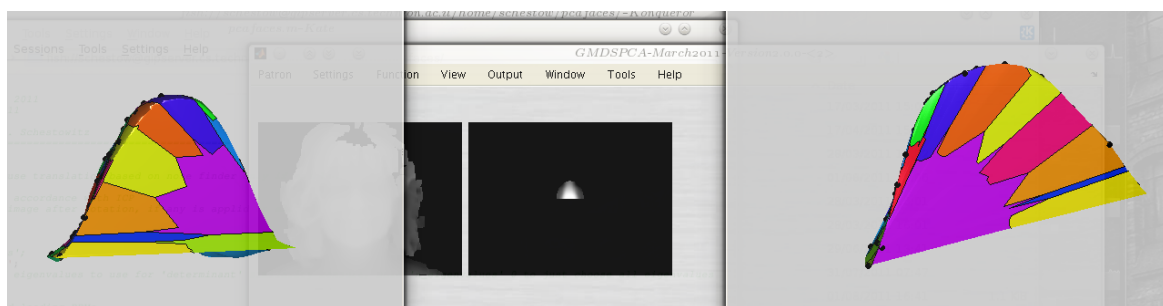


Figure 148: Example of the effect of ICP-induced rotation on the Voronoi cells

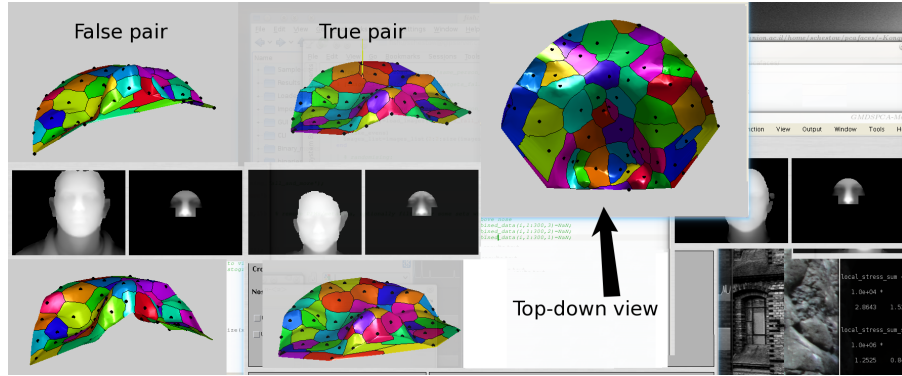


Figure 149: Return to the old mask with additional rotation, which does not yield better results than those at the region of 92%-98% recognition rate

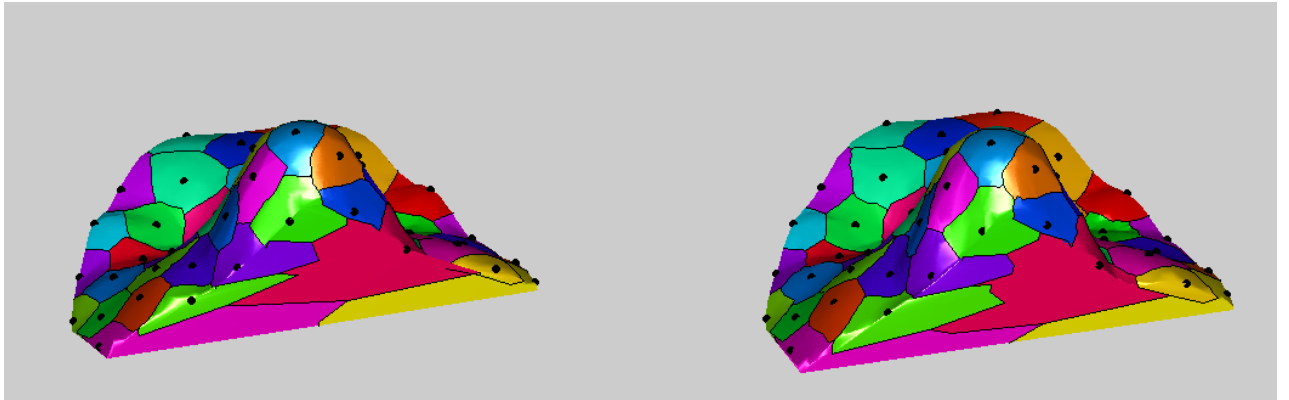


Figure 150: Cheek inclusion gradually staged in for understanding of its impact on recognition performance (geodesics and PCA)

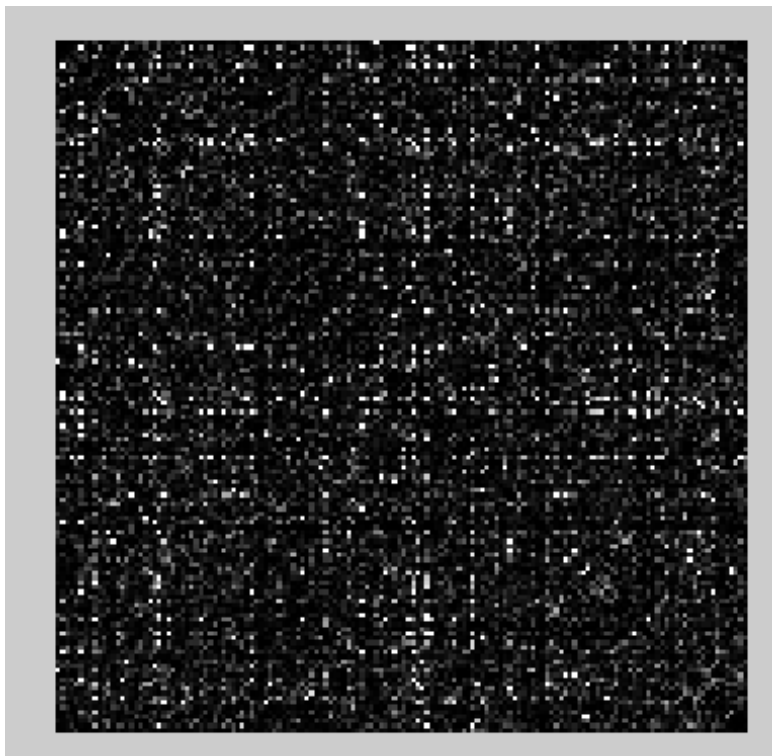


Figure 151: The stress map corresponding to the new binary mask (with 150 points for FMM)

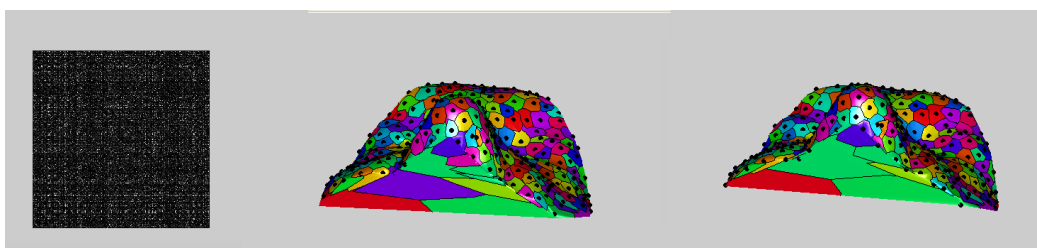


Figure 152: Stress map and the corresponding faces (looking from beneath the nose) from which it is derived

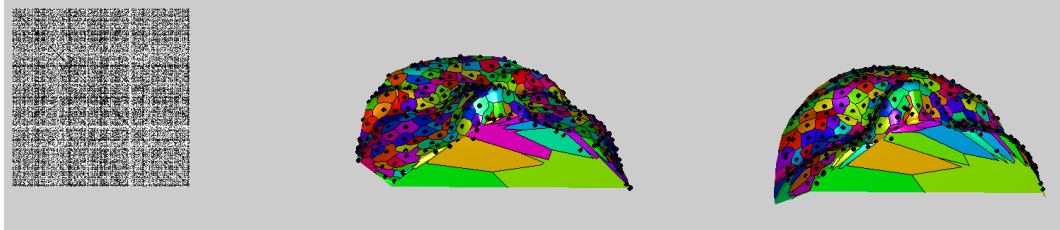


Figure 153: An example of a false pair (different people) and a cleaned up stress map showing some interesting patterns

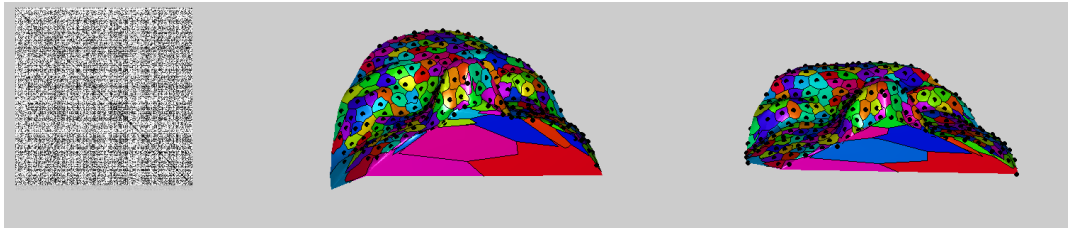


Figure 154: Another example of a false pair and the results of GMDS

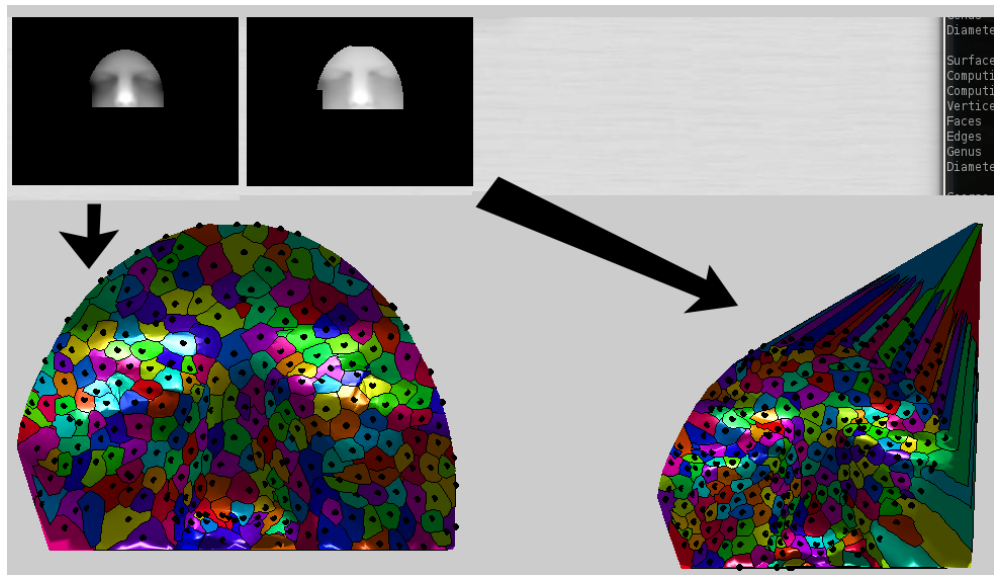


Figure 155: Example of a bug found in the program, leading to massively false correspondence upon the same person

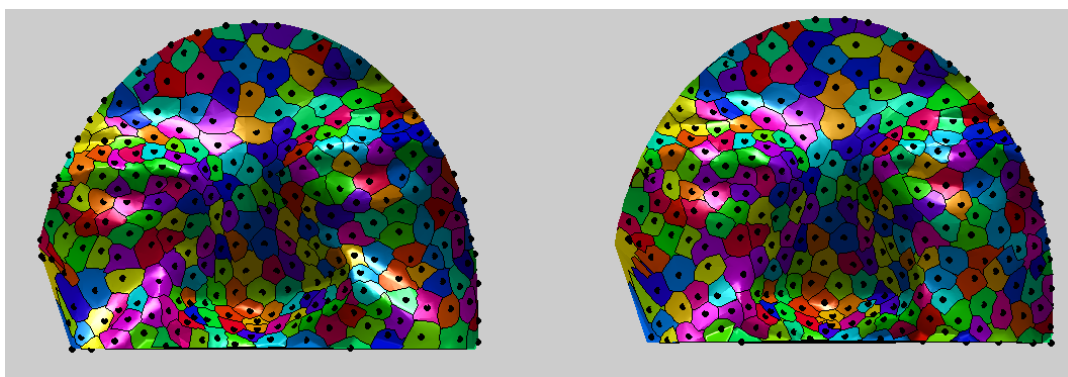


Figure 156: An example of acceptable matching between two poses of the same person

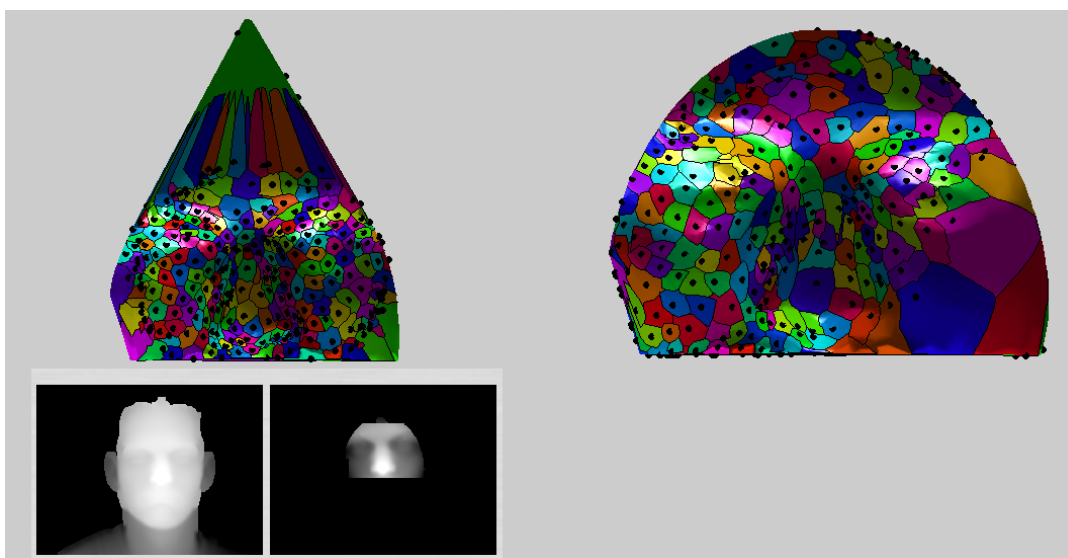


Figure 157: Another example of a bug found (and resolved) after it had proven problematic to recognition rates

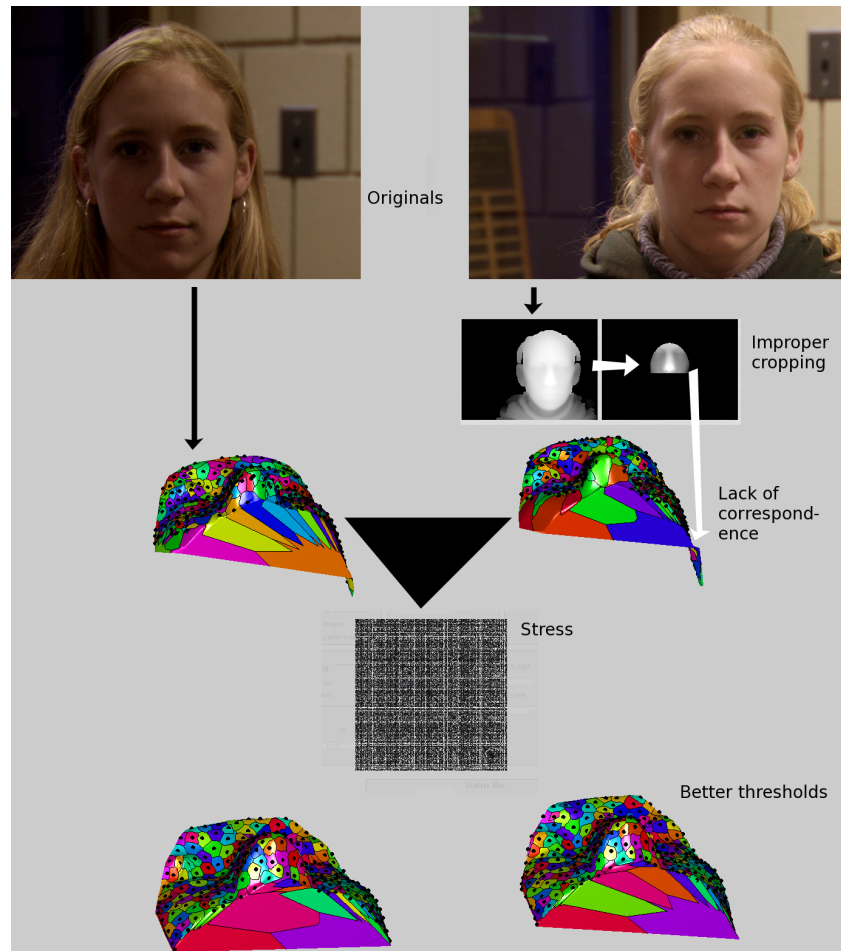


Figure 158: A look at the problem associated with narrow faces that lead to incompatible sampling

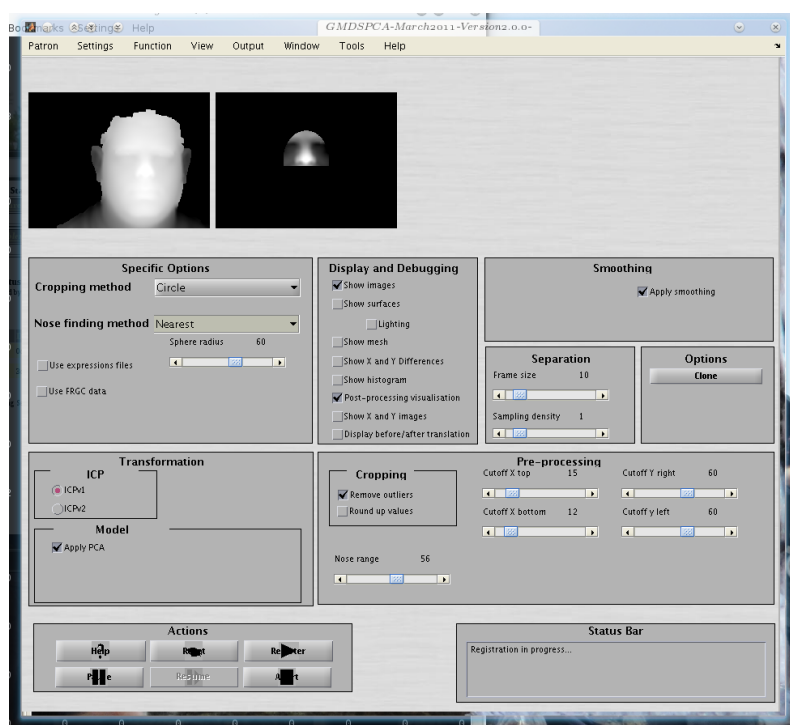


Figure 159: General program settings used for the subsequent experiments

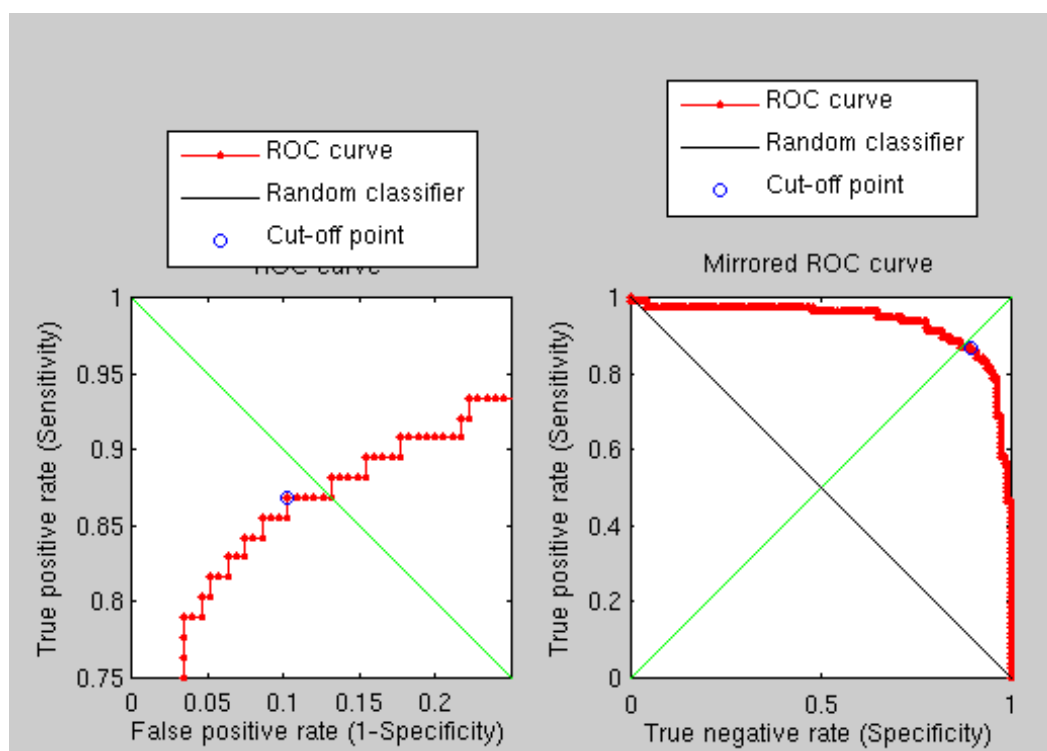


Figure 160: Results of a large-scale test after previous bugfixes

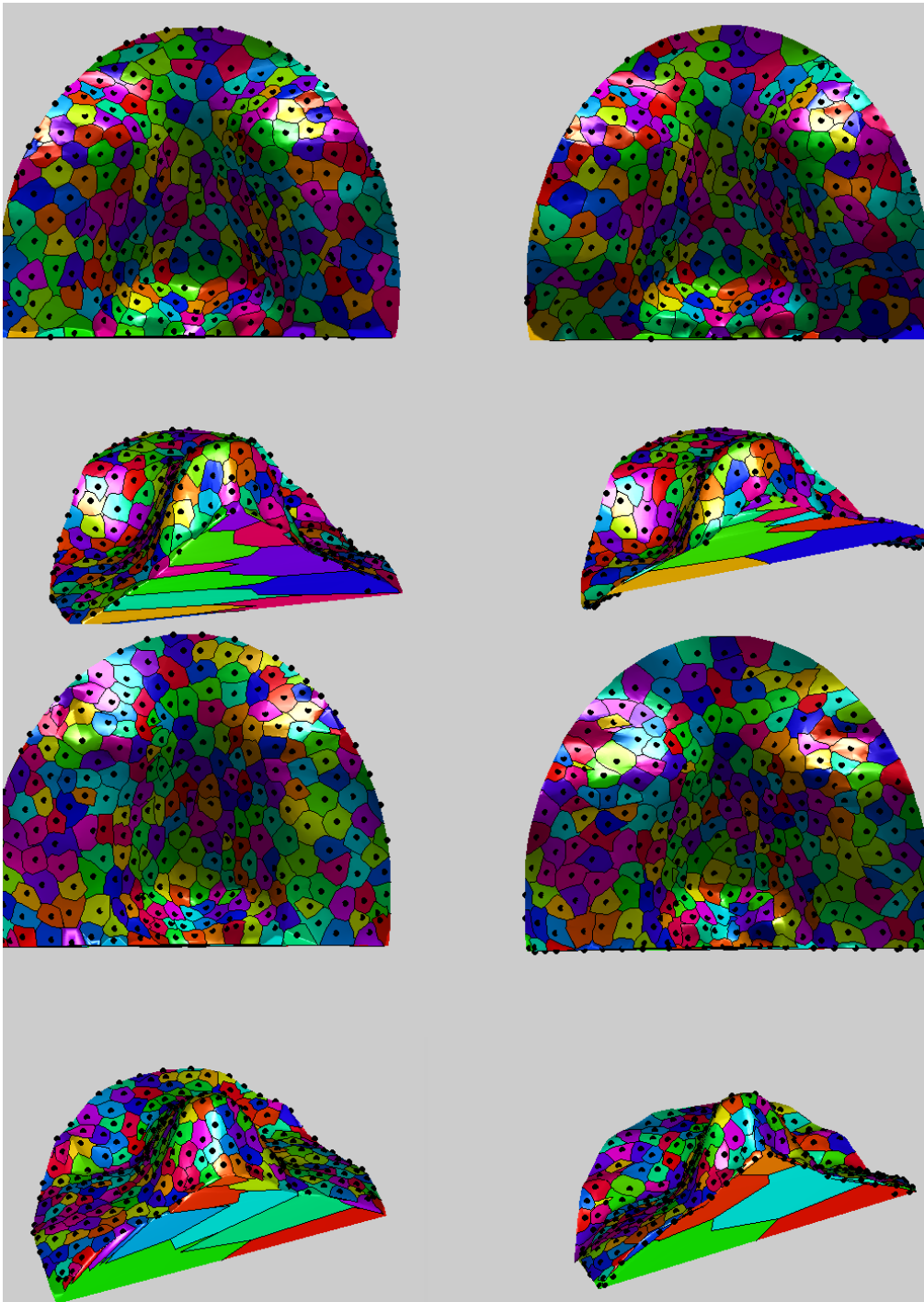


Figure 161: Example of a correspondence problem in a pair of images (one image on the left, another on the right). The top 4 images show the correspondence after the bugfix for one pair and the bottom 4 show the outcome of applying a fix to another pair.

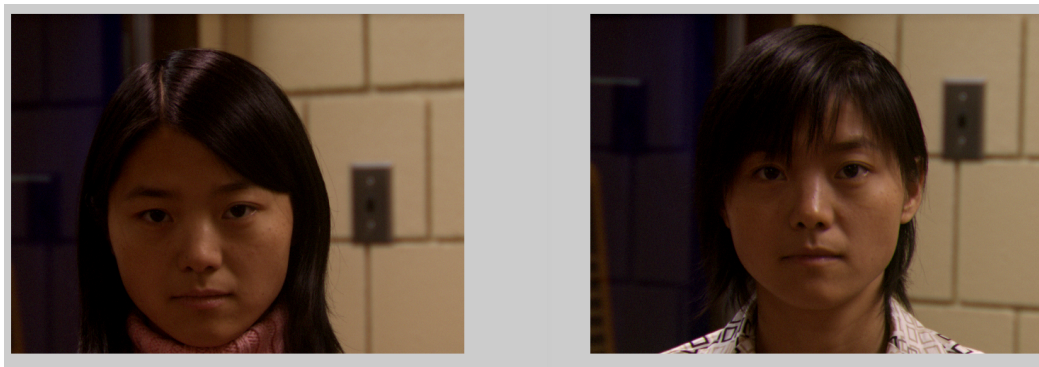


Figure 162: Example of a problematic pair where hair obstruction and nose position compared to the forehead caused an issue which is now properly addressed

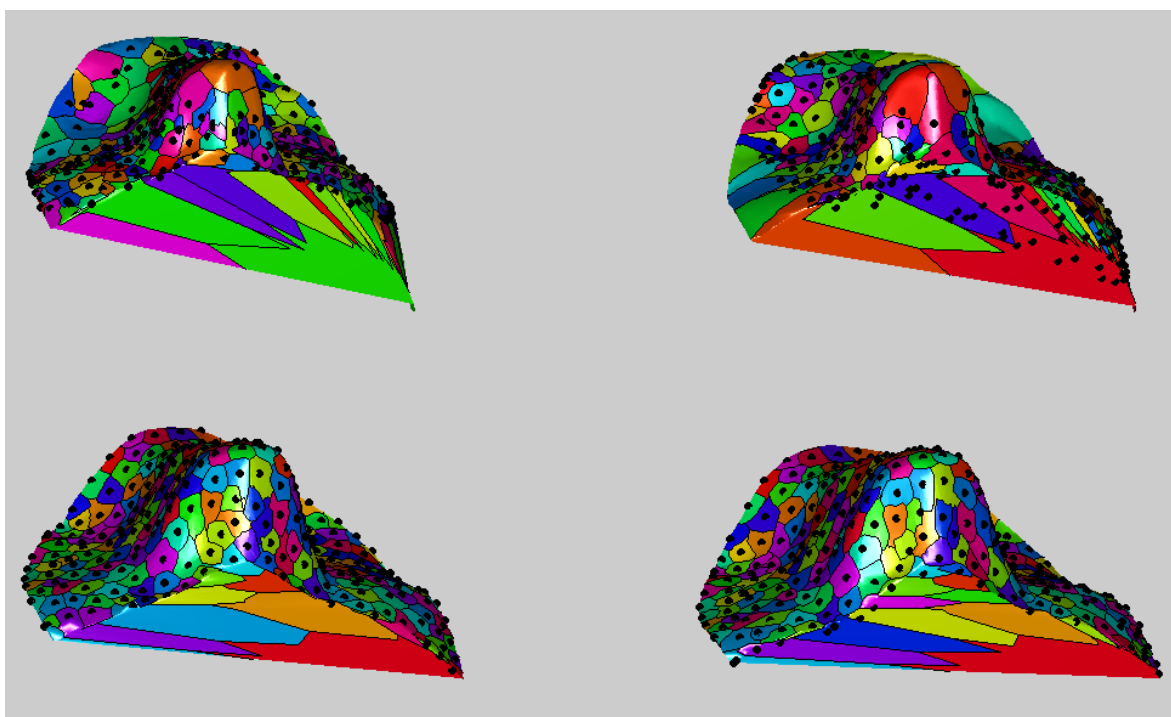


Figure 163: Example of a pair where the side of the face got sampled, leading to serious issues (top) before they got resolved (bottom)

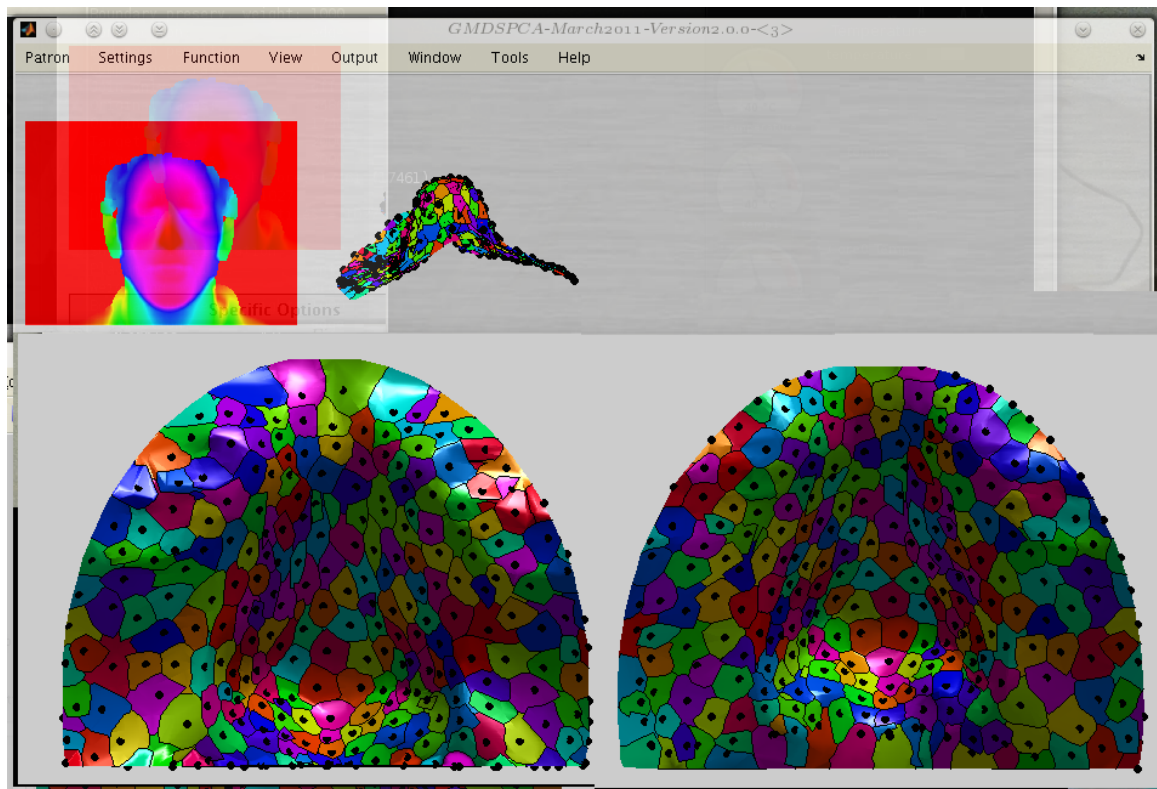


Figure 164: Comparison between images of the same person, where the height of the nose relative to the cropping is causing issues



Figure 165: The images corresponding to the above example (same person, different positions)

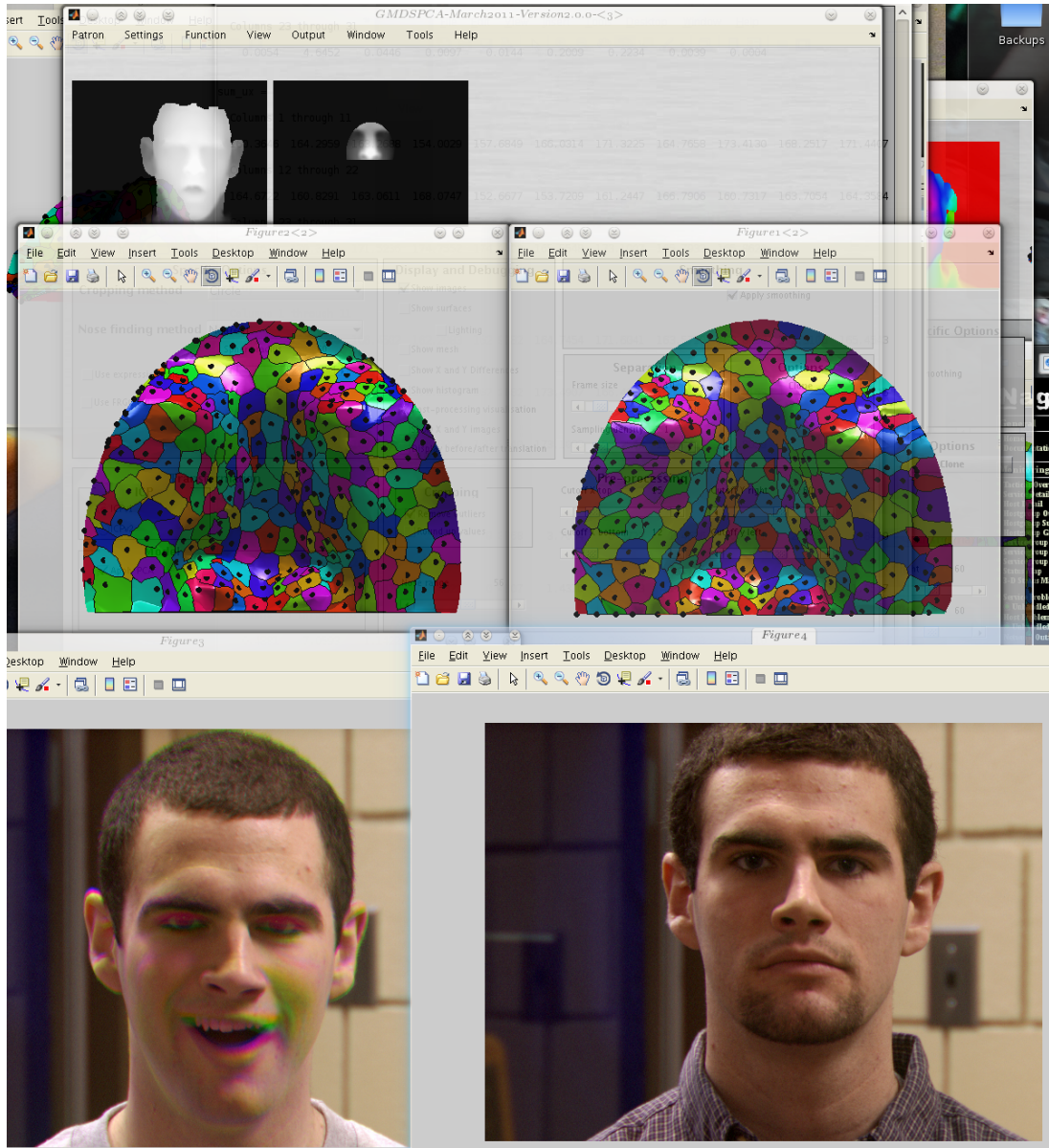


Figure 166: Another example of a problematic example where the score borders on being seen as “no match” even though it is

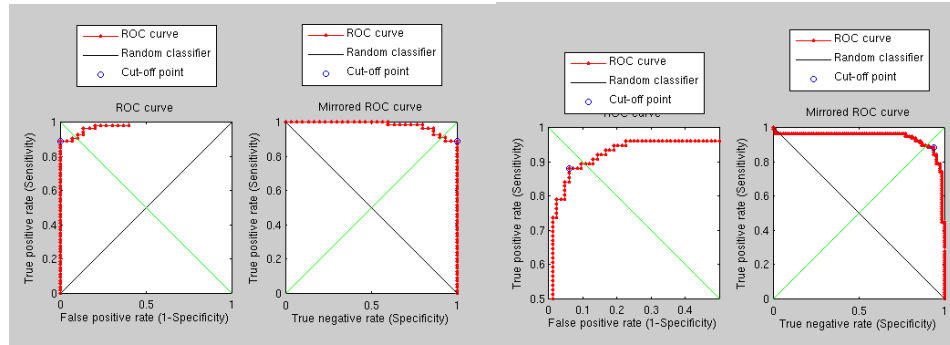


Figure 167: Some recognition results from the above experiments, with denser sample on the right where the cheeks were also remove to test their impact on performance (little impact)

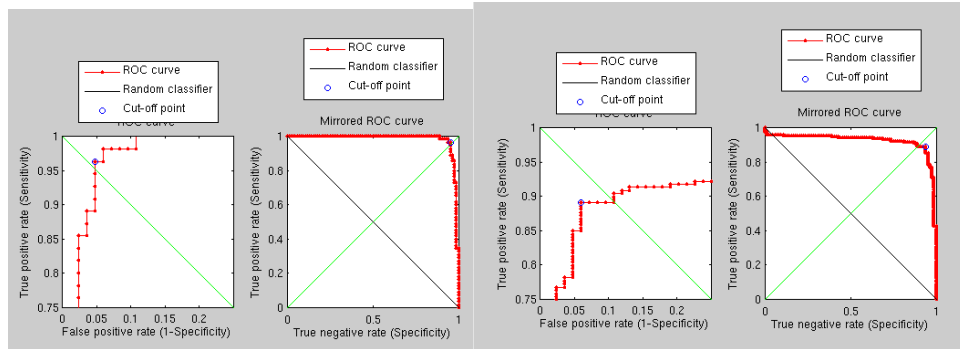


Figure 168: Smaller-scale and large-scale (right) experiments that look at how applying the methods only to the training set (many identical faces clustered together) changes the above results. It does not affect them much.

Data preparation works as before on the Texas dataset, having encoded the necessary adjustments and make them more modular (so as to keep the program compatible with GIP datasets and NIST datasets). Currently, the program crashes at GMDS sometimes, but that ought to be resolvable within days.

With caveats, after much work at increased pace, the GMDS-PCA code can now be applied to parts of the data from Texas, as shown in figures 169, 170, and 171. These are not matches between identical people but between different people. The code is not yet in a state good enough for benchmarks. After much debugging that includes visualisation, regression and repetition, it appears likely that the bug we have been investigating for weeks is in fact not truly a bug in my code but an initialisation problem associated with GMDS. When bouncing back and forth between the two surfaces trying to match one to another, GMDS sometimes appears to fall flat on its face. When it gets things right, the results are nearly perfect; when it does not, there will almost certainly be a detection error, which is at least predictable. The initialisation as it stands at the moment benefits from dense initial correspondence (based on the ordering of points) with ICP properly applied and its validity further verified. Unsurprisingly, all three datasets (GIP, NIST, Texas) are affected by this; without it, the PCA-GMDS approach would not work so well either (it just takes longer), as it heavily depends upon the finding of points between analogous points in almost every case. The built model would be without value unless there is consistency or contrariwise a

reduction of irregular observations (robust PCA).

The observed behaviour is curious. While the process is somewhat stochastic and non-deterministic, upon different runtimes the results are usually more or less the same, maybe with a variation of a few percentile differences up or down, perhaps a fraction of 1%. But with particular pairs of surfaces there appears to be inconsistency as those same two surfaces are in a bit of a limbo. It is possible that GMDS will get the segmentation/correspondence wrong many times in a row (with widely varying stress values) and then ultimately get them right somehow. Figures 172, 173, 174, 175, 176, 222, 178, 179, and 180 give some examples of the process of debugging and the issues encountered along the way.

Multiple Initialisations or Multi-Scale Approach GMDS is a non-convex problem, and indeed it would give the right solution only if we start at the basin of attraction of the global optimum. Multi-resolution and any other good initialisation should get us there. We then wondered. Have there been any attempts to apply an approach which attempts multiple initialisations and then selects the best match among those? It means something along the lines of simulated annealing? Could be interesting and some people did try something related for symmetry detection (i.e. used GMDS to refine several initial conditions determined by feature points). See Can Raviv's work as well as Anastasia Dubrovina's main efforts in her masters (again linking feature matching to GMDS relaxation). But it was not explored much beyond that.

We could plot the accumulated stress per point so that we see where the mapping fails. One might still suspect the cheeks are to blame. We have been looking at different segments of the face and how dealing with each of these in isolation leads to similar problems in cases where GMDS struggles. A look at the stress map reveals nothing too unusual and almost the entire time cheeks are included as well (only the earliest experiments excluded them because these were the default settings).

After further tests on several smaller areas of the face (assuming that a piecewise approach of GMDS-PCA with LDA and maybe fiducials for different distinct features might work), it seems like the issue is fundamentally related to the way correspondences are found, or in about 5-10% of the cases simply not found. The success of recognition in all 3 databases is hinged on the ability to always do this correctly.

A group of small-scale experiments were run with the aim of investigating how selecting different regions of the face would affect the success of GMDS. A systematic scale increase of about 10% at a time (in terms of the relative size of the region in question) was used to show, although not on a statistical basis, that for very small regions with very uninteresting structure the variation is too large, an order or magnitude apart sometimes, which made GMDS inadequate for the task. Thus, a an investigation of the problem domain at macro-scale was undertaken. There was also some experimentation with binary masks and smoothing, under the presumption that if more date and less pertinent details are available, then performance will improve. Part of

the problem is still be explored. The stress maps show no evident problem around the cheeks, which – even when included as shown in the attachment – are not exclusively where extreme stress is found.

In the chart, stress score for the pairs is (from left to right): 3.4866, 2.4497, 2.4718, 10.9726, and 173.6779 (see Figure 181). The images were flipped upside down where the correspondence found was asymmetric (left being right and vice versa). Since all are pairings belonging to the same person, it is expected GMDS that should succeed most of the time, rather than in about 90% of the cases. The main pitfall here is that there is no guarantee that correct correspondences will be found; the option which seems reasonable is reseeding the stochastic process and retrying, although that would be wasteful and quite computationally expensive. Perhaps some fiducial point can instead be used to improve the initialisations.

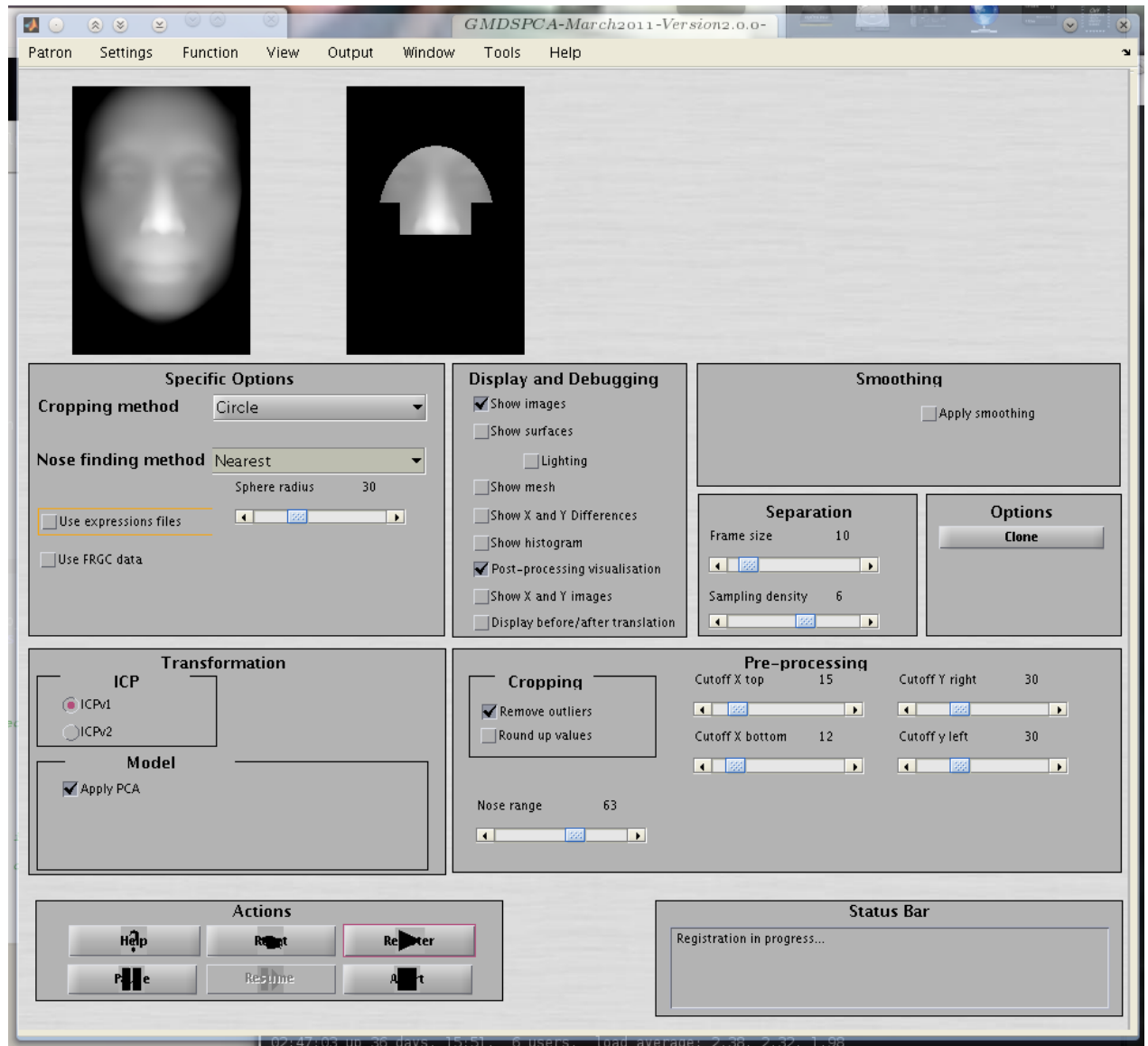


Figure 169: Standard program settings with which to run the Texas data

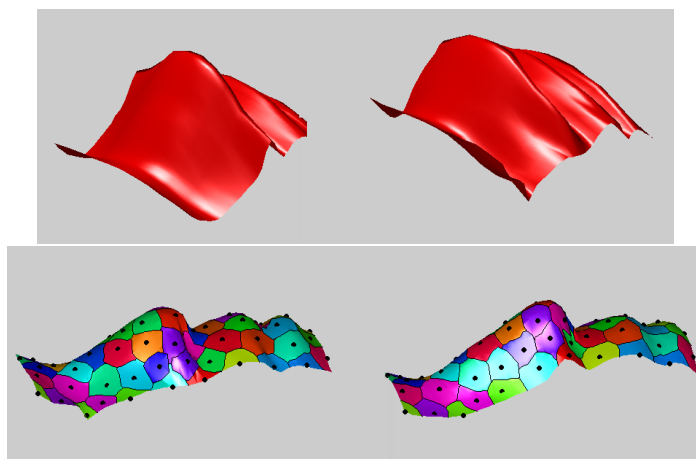


Figure 170: An example of GMDs applied to just a vertical slice of the data taken from different individuals

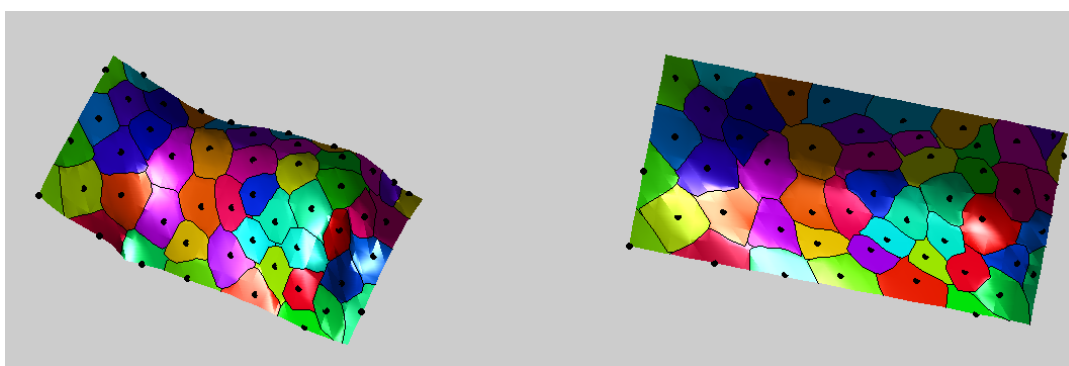


Figure 171: Exploratory work around GMDs applied solely to the nose region of different people (left and right), shown from different angles

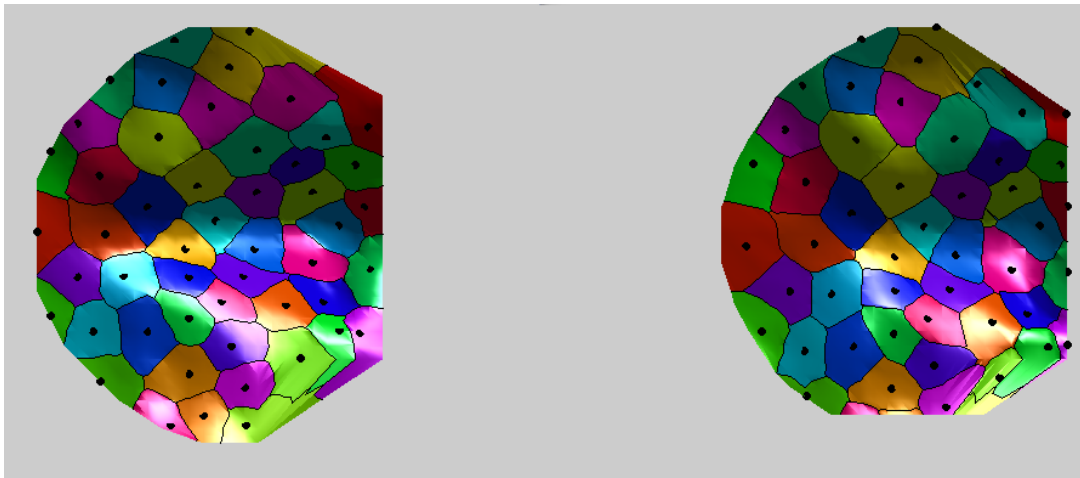


Figure 172: Initial experiments with the Texas3DFR Database excluded the cheeks, which were later added as various parameters were studied for their impact

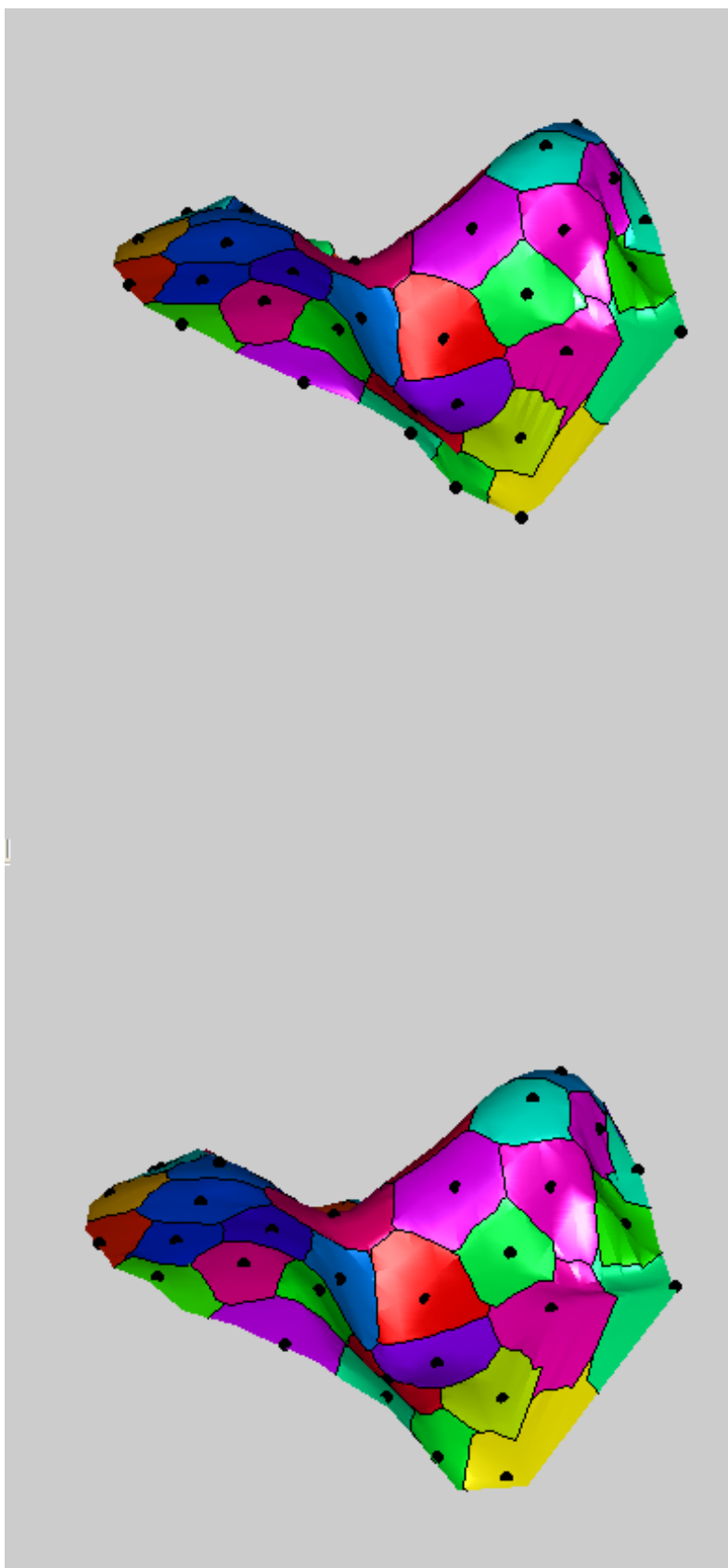


Figure 173: Texas3DFR Database pairs with the correct correspondence

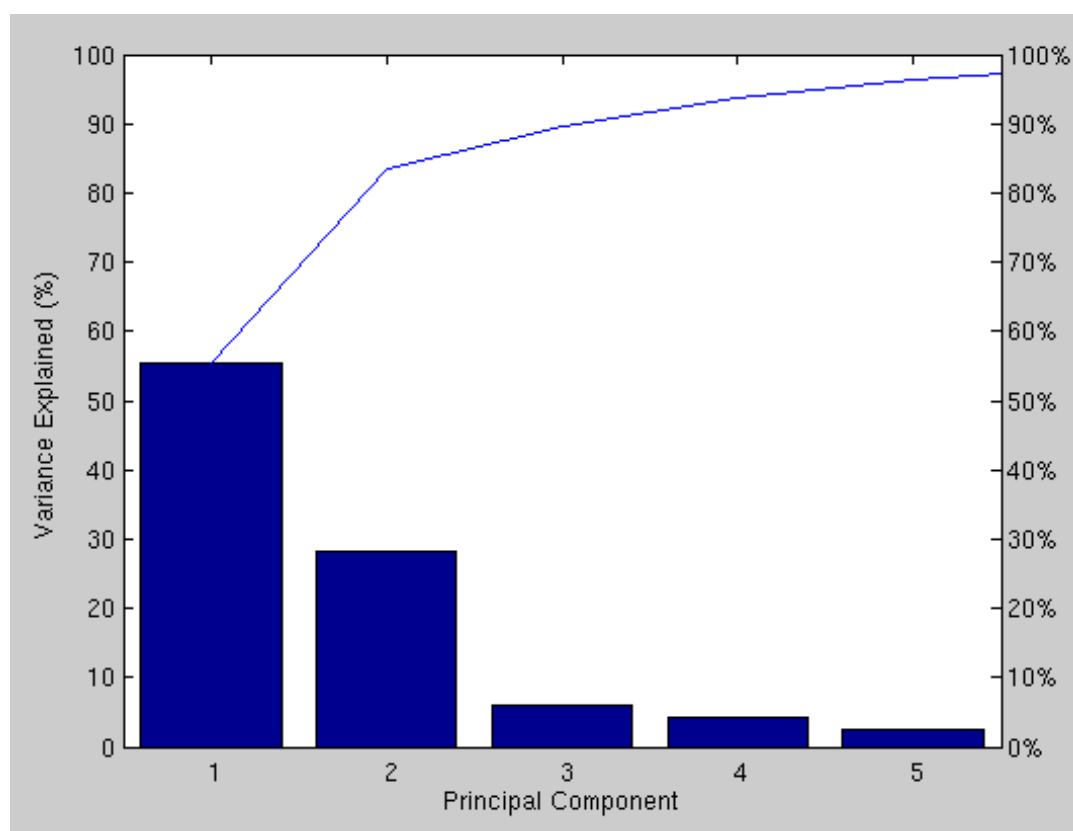


Figure 174: Model modes with more than 1% variation built from correct pairs

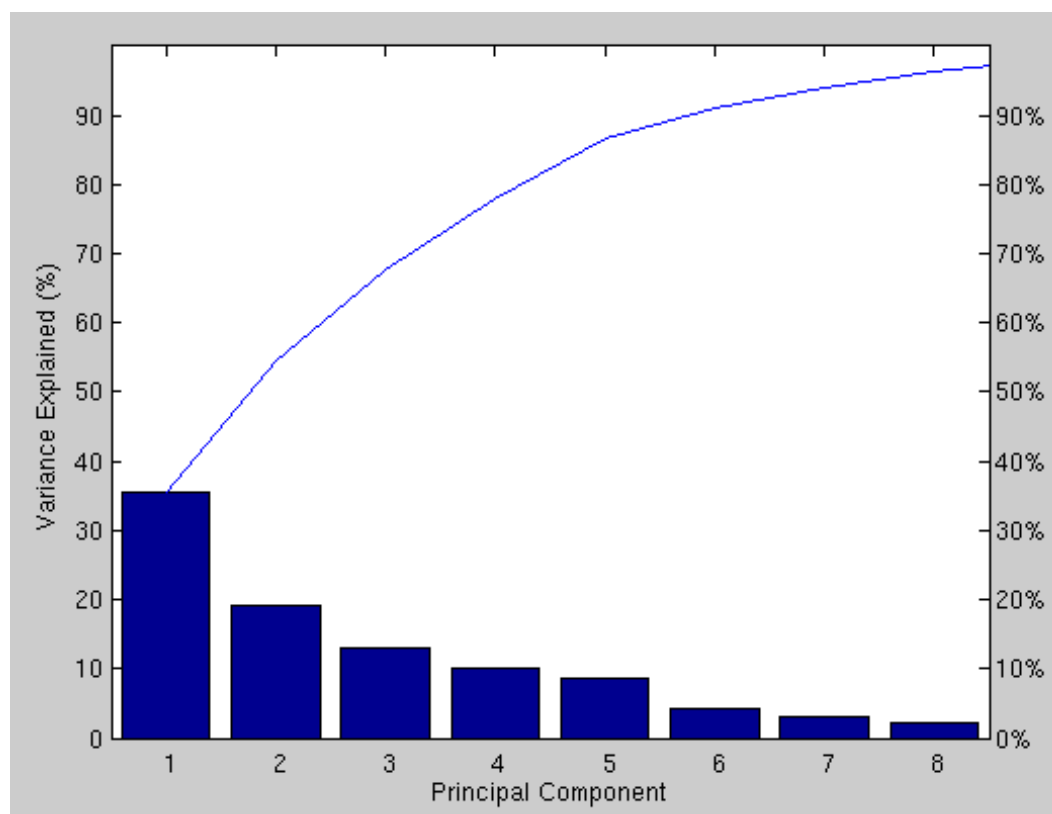


Figure 175: Model modes with more than 1% variation built from false pairs

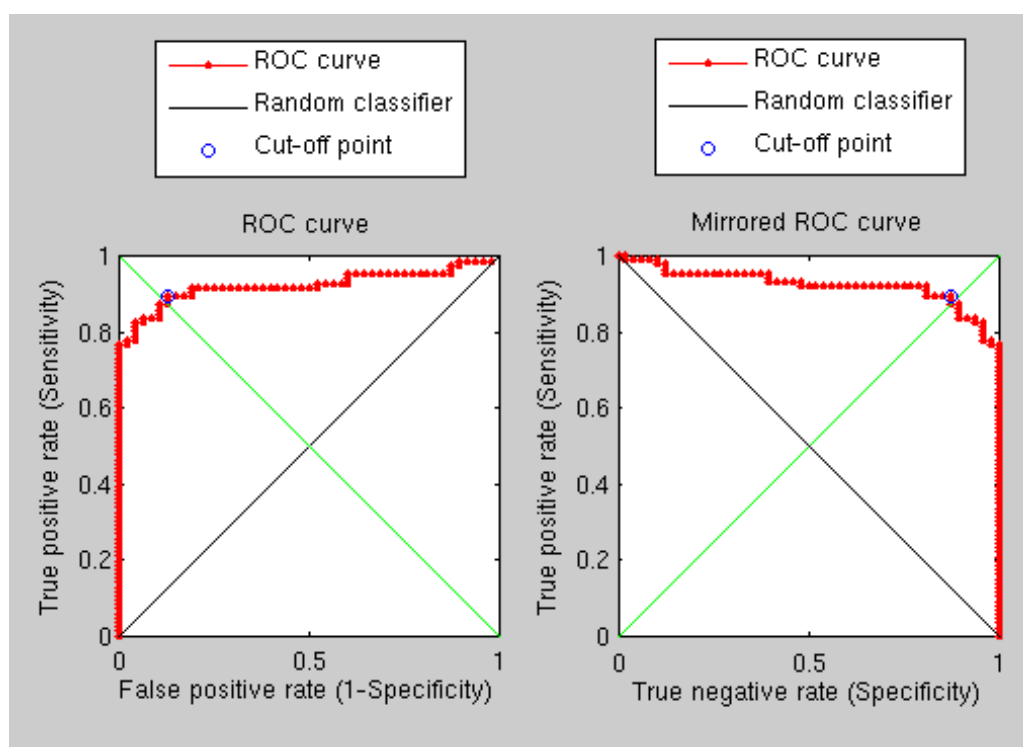


Figure 176: With GMDS issues still in tact, the ROC curve for recognition suffers

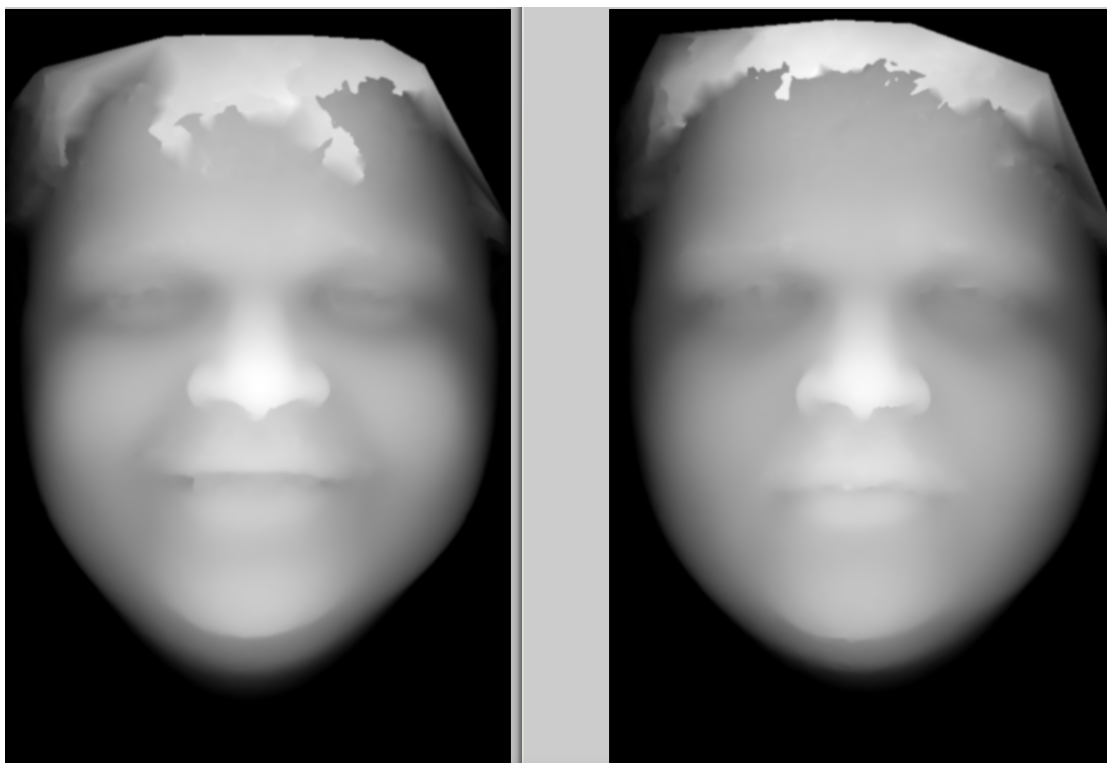


Figure 177: A pair that GMDS usually fails on

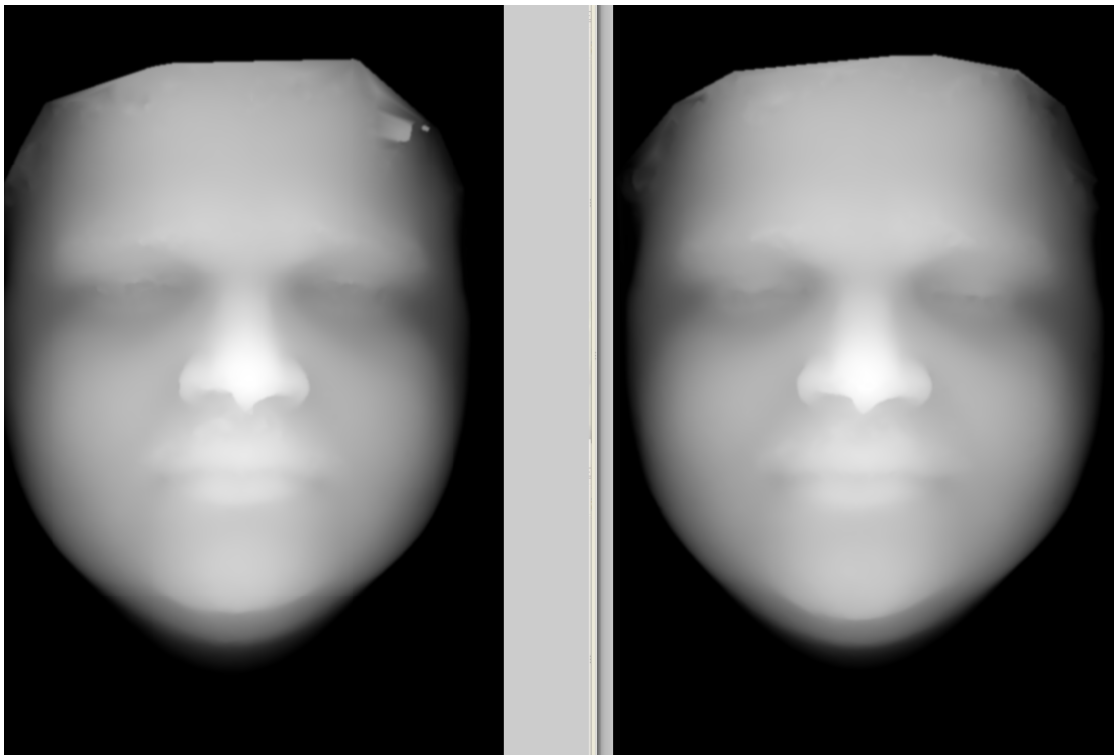


Figure 178: Another pair that GMDS usually fails on

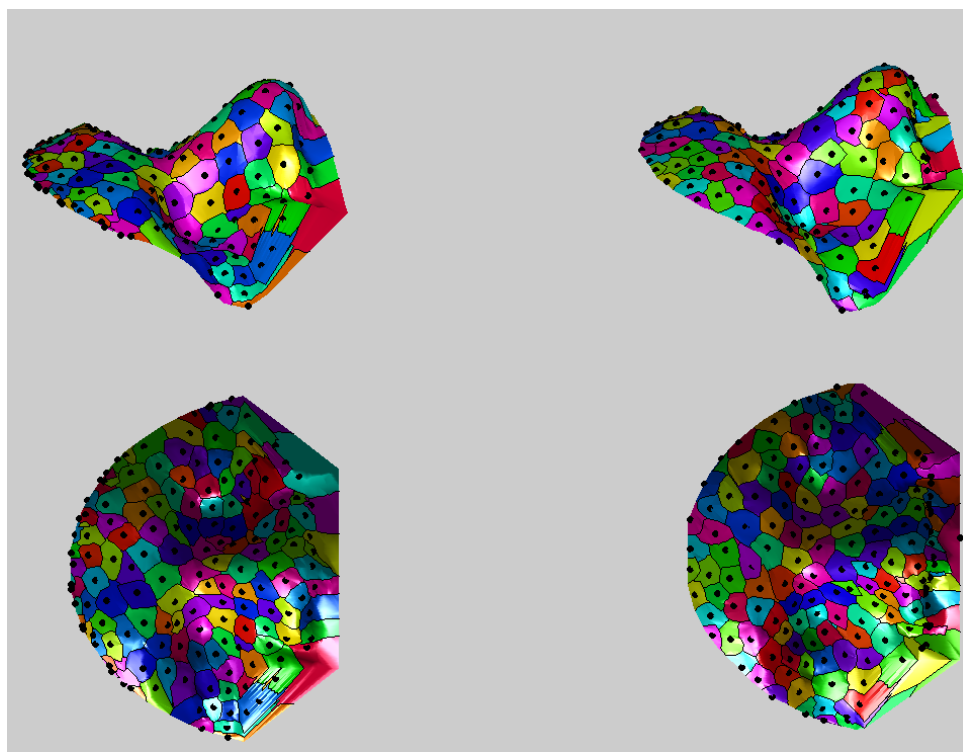


Figure 179: A closer, GMDS-style look on the very flawed correspondence-finding (example from Figure 222)

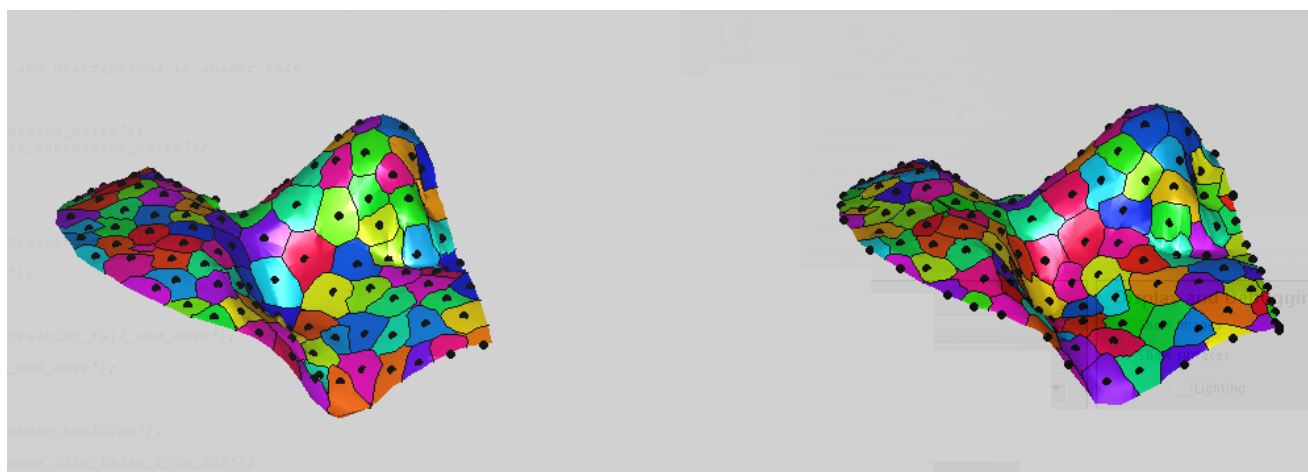


Figure 180: Another example of a GMDS-type comparison applied to a real pair and failing

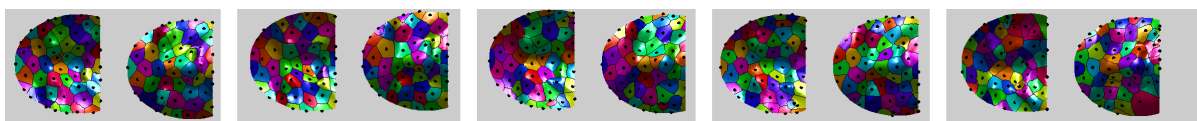


Figure 181: Pairs of facial expressions from the same person, cut in half beneath the nose and tilted sideways, then shown with GMDS applied. The score (from left to right): 3.4866, 2.4497, 2.4718, 10.9726, and 173.6779

It also depends which kind of smoothing gets used, if any. Note that it should be intrinsic. I.e. convolution with a Gaussian which is coordinates dependent (w.r.t. to plane) would not do the trick for between-expressions matching.

Smoothing was used at high levels of granularity when handling expression files that were GIP-formatted and exceptionally noisy. Without a fair deal of smoothing or averaging (or median) the data was hard to work with as any raw image was too noisy. This was part of an image sequence taken in the form of video. For FRGC data, smoothing was a lot more limited and most of the time it was not necessary because the images were generally of good quality. The same goes for the data from Texas, which probably offers the best image quality. Smoothing was only applied to it experimentally, although based on intuition it was not necessary. As smoothing was applied to range images, it was perpendicular to a fixed plane rather than tangential w.r.t. the shape's surface. It is abundantly clear that smoothing along the surface and not along the viewing angle would be the defensible filter to apply. In case smoothing is revisited, it will be implemented properly, but as it stands – given the high fidelity of images – smoothing only degrades signal and removes no apparent noise. It also does not appear to impact performance for the better, although more systematic experiments would be needed to validate this type of contention. These are probably not the right experiments to pursue at this stage.

It is obvious that the smaller the region the less discriminative it would be. This is why one could be hoping that at the end of the day we will add

as much region as possible. Having spent at least an hour or two viewing the results of ICP-aligned, full-face GMDS, it seemed evident that the same problems persist. This was after a set of systematic debugging/scoring round on two desktops and two 8-core servers. At first it appeared to be achieving perfect detection, but later on some false negatives and false positives could be spotted. This was due to GMDS failing to find the correct correspondences. In these experiments, areas around the hair, neck and ears had been culled out. The way this was implemented did not make a geodesic criterion for culling though. GMDS discrimination per scale would require that underlying issues in GMDS are first resolved. Using geodesic circles we can make the surfaces more consistent. We could slice out a union of geodesic circles about features rather than the "mushroom" on the projection plane? One might suspect wrong support could also lead to distortions. And the support should be intrinsic rather than projection plane dependent.

Also the graph I was referring to is a texture mapping of the integration over the distortion between a point and the rest of the points. It would give you an indications of where the largest distortion happens...

looking again at the examples you sent, the support is a problem in the last example. Wrong support leads to wrong alignment.

We shall work on each of the suggested areas of improvement in turn. The correct fix seems to be within arm's reach and the addition of a new menu option for geodesic circles sure seems necessary. The results at the moment

are not as negative as that last illustrative set of images may suggest; this was supposed to accentuate the differences between good matches (single digit) and those which are 1 order of magnitude or even 2 above the rest.

As means of showing that GMDS on the full face does not work reliably, shown are a bunch of results (including stress maps with overall scores) obtained from pairs of the same person. Noteworthy are the observations where left and right need to be flipped and in one case, the one where the score is extremely high, the matching is found to be upside down (chin matching forehead). See Figure 182 and Figure 183. Although the author's experience with GMDS is somewhat limited, it does not appear as though carving out a on a geodetically-defined boundary will magically resolve the issue, so we will explore some other tricks until the missing code can be fetched (or alternatively be re-implemented).

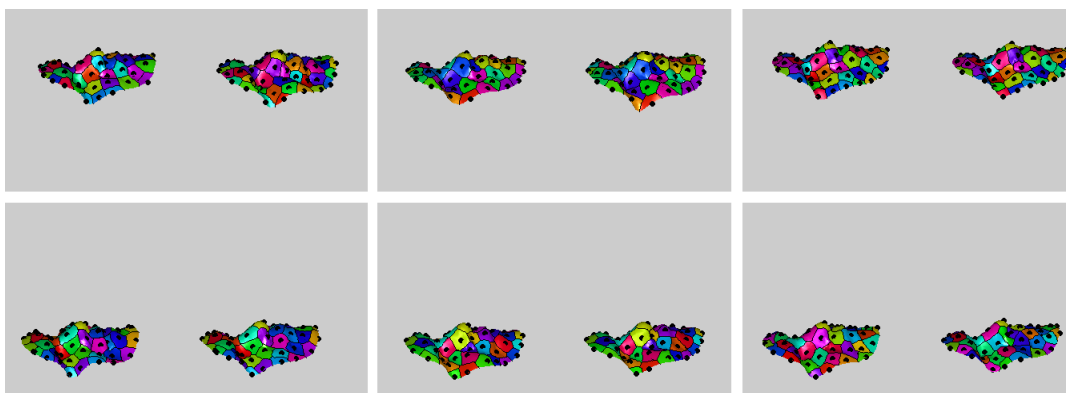


Figure 182: Some of the raw images (full face) after GMDS with 50 Voronoi cells displayed

The images show how stress is distributed over the shapes/surfaces. But

stress mapping is something else. For each point on the surface we plot a gray level or red level that is equal to the sum of its distortion from the rest of the points. The total stress of each point on the surface, p_i , w.r.t. to all points in P . If one views the matrix of geodesic distances between each point to another, then the desired outcome would be a figure, e.g. with Voronoi tessellation, where for each point we know how much distortion (or total distance) it has. This would help debug GMDS as applied to faces, but such functionality might not exist yet. To clarify, we have two shapes, X and Y, with sets of N points on them, $\{x_i\}$ and $\{y_i\}$ accordingly. Say these points were found by GMDS, that is pairs (x_i, y_i) are corresponding points. Now, one can take these points on one shape, say x_i on X, and use Fast Marching to find distances from them to the rest of the points on X. The same you can do for Y. Basically, all that's required is that for each such point, $\{x_i\}$, the total of all those distances from $\{x_i\}$ to all points on X (or Y) should somehow be shown graphically. I wonder if the visualisation tools for GMDS – not necessarily those with Voronoi cells – can be used to show us the stress per point, $\{x_i\}$. What we try to achieve is, basically, something that tells us where the fitting between two surfaces (faces) fails. At the moment we can plot the matrix of distances between disparate points, but we do not have a way of visualising where each point number exists on the actual surface. We hope to find a simple way to find or visualise which point shown in surf/mesh/plot3 corresponds to point i . This is needed for debugging or tweaking.

The next step will properly look at areas of difficulties. In the mean time, we will explore the effect of the algorithms on some synthetic data, at the very least in order to improve our understanding of how and why things fail. A blob that varies in terms of its size, location, and aspect ratio will be used very briefly as a form of validation data, a fourth dataset even. Later on it might help reason about the validity of this approach, showing its ability to recover the correct solution and the ground-truth correspondence.

Synthesis is shown in Figure 184. The short set of experiments used data consisting of three types of consistent patterns with different scales (width), location in Y and rotational orientation (small variation). This was designed to test using well-behaved data how GMDS copes. The results are shown in increasing order of difficulty and they seem to suggest that the appearance of new features around edges causes confusion, so a stress analysis and geodetically-defined cutoff will be the next logical step.

It may not be understand what is going on in the examples. The synthetic data should ideally be richer, i.e. the surface should contain effective Gaussian curvature, else the mapping is ambiguous.

The examples in Figure 185 show the general results one gets when running the algorithm on "true pairs" (same person) and "false pairs" (different people). There is one case (shown by the score) where there is a misdetection, meaning that for a correct/real pair the correct correspondence is not being found. By looking at a geodesic slicing mechanism (ring/circle) and inspect-

ing the spatial distribution of stress we can hopefully overcome this issue for all datasets at once, hitting several birds with one stone. The value of the method is very much hinged on our ability to resolve this ambiguity, so it is worth the extra effort and continued research. Having looked at other possible selections of a sub-surface, they don't seem to offer a noticeable advantage, so for the time being the top half of the face will be used, centred around the nose.

We've finished implementing a variant of the Voronoi code that displays stress as colours, where the brighter the colour, the greater the stress (see Figure 186). As rightly argued at the start, if cheeks are culled out, then the stress around them is very high.

Worth noting is the inability to access more GMDS code (we use the TOSCA demo code only).

The colours of dots have been improved so as to take full advantage of the whole range of greyscale levels and shown as an example in Figure 187 is the difference – stress-wide – between inclusion and exclusion of the cheeks.

The next thing to implement is a geodesic mask. Geodesic masks are indeed necessary. This would be sensitive to the source location and, still, with partial matching it could do the job. While trying to avert some mysterious problems with unknown program crashes (hardware exception, not consistently reproducible) the results are approaching what needs to be achieved (several features with union of geodesic circles around them).

In Figure 188 are some of the debugging artefacts.

The next step concentrates on finding an adequate cropping methodology which preserves surface area based on the anatomy and not Euclidean measures (the former is invariant, whereas the latter changes with expression and pose). Having found out that Fast Marching was causing the crashes as deprecated interfaces had been used, this step ought to be a simpler one, but for qualitative results it might be necessary to run many different experiments like those tested for Figure 189 and Figure 190. Previously it was found out that just taking the top half of the face and removing clutter along the sides worked better than taking the entire face, which sometimes led to flipping (matching upside-down). Taking a smaller portion than the 'half face' led to lack of signal (lacking gradient).

7.11.1 Geodesic Cutoff

The core supposition or hypothesis here is that by better selecting boundaries of the surfaces – using geodesic means – the innately geodesic method which is GMDS will perform a lot better and hardly suffer from insufficient information, commonly caused due to Euclidean cropping criteria (as attempted beforehand). For consistency, the experiments that follow will adhere to this 'half face' approach as a given. Realistically, taking a hybrid of features and applying LDA or integrating over them might work better. This is also well supposed by the geodesic masking function. Rather than scoring based on

just two surfaces, it is possible to compare subsets of these, with or without fiducial points as a key component to latch the sub-surfaces onto (e.g. eye corners).

Early exploration of our problem revolves around the effect of varying the point around which surfaces get carved. It can be shown that, even for the same person (i.e. same face), the effect of moving that point is as severe as comparing two different faces, so this point is very prone to change the results (a sensitivity-related issue). An interesting experiment might be to distance oneself from this point by controlled and gradually-increasing amounts and then rerun experiments of recognition, seeing how accuracy of this point's allocation affects performance and how much degradation is caused by missing it even very slightly. The location of the features with sub-pixel accuracy cannot be guaranteed, so this might be important to do as a form of sanity check.

Then we have a case of Euclidean plus geodesic cutoff depending on the side, with a fixed point below the eye. The stress/mismatch/distance too is being shown at each point. It helps show the difference, but doing this more properly with geodesic distances all around is probably best. In terms of results, it seems promising, but more elaborate experiments are needed before arriving at any such a conclusion. Using just the nose area for matching has proven to be a poor approach based on some rudimentary tests, but maybe this can be improved shall the need for piece-wise feature-based comparison arise. The geodesic cutoff examples are shown in figures 191, 192, 193, and

194.

We shall use the following feature point selection and geodesic circle (support) strategy. If without loss of generality we select the nose and eyes as the features and our geodesic distance ambiguity of the feature selection is δ_a , then for the gallery (target) surface, we define the support as the union of $r + \delta_a$ geodesic circles about each feature point (δ_a and r could be different for each feature). The probe surface should then be defined as the union of radius r -geodesic circles about the features.

This way we try to embed a smaller surface into a larger one, where the larger one is large by the amount of ambiguity of the selected feature (we would like to assume it would be small, but not too small, one would say, 5mm).

3 points were spread rigidly around the image to mark the centres of points which define geodetically-bounded surfaces. One of these is the tip of the nose. In order to prevent the mouth, for example, from entering the surface (it depends on the length of the nose and its vertical component prevents sufficient point sampling due to the camera's angle), there is a Euclidean concern around there, which explains the flat boundary at the bottom (Figure 195). The results can be properly explored once a tolerance component is added to the probe or the gallery (consistently for score stability) and another quick set of evaluations was run on a set where the carving was based on the forehead and nose rather than the eyes. (Figure 196) There is probably too little information of high entropy around the forehead, though. Sometimes

there is hair there.

In terms of performance, with just 600 vertices it does a lot worse than the Euclidean approach with a lot more vertices.

As we rerun with high density, the results change. Increasing the number of vertices to 2420 improves the results considerably. Some results are shown in Figure u198 and Figure 199 show some early results.

It is hard to make the case for surface sizes that are not as equal as possible, based on the signal (similarity which is still too noisy). At a coarse and fine levels too, the similarity is not great when geodesic measures are used to carve the surface. By going much higher in terms of resolution we can approach or exceed 95% recognition rate, but to reach 99% or thereabouts there will be further exploration of what needs tweaking, based on pairs where misclassification occurs.

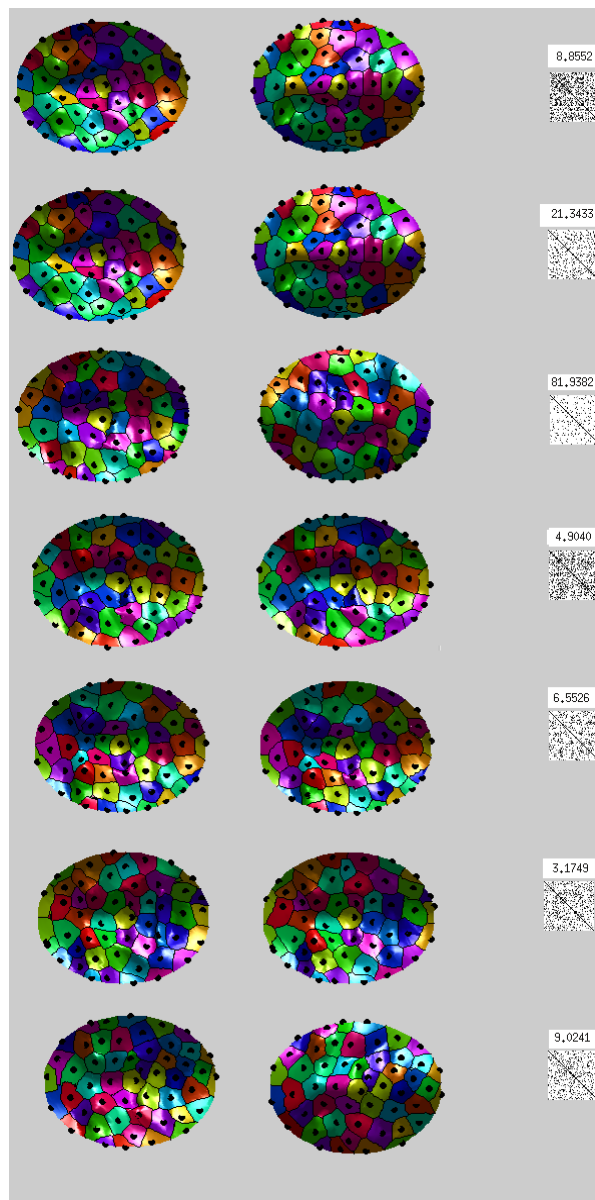


Figure 183: A top-down view showing the matching and the corresponding score. with flipping manually corrected. The third example from the top got the topology completely upside down.

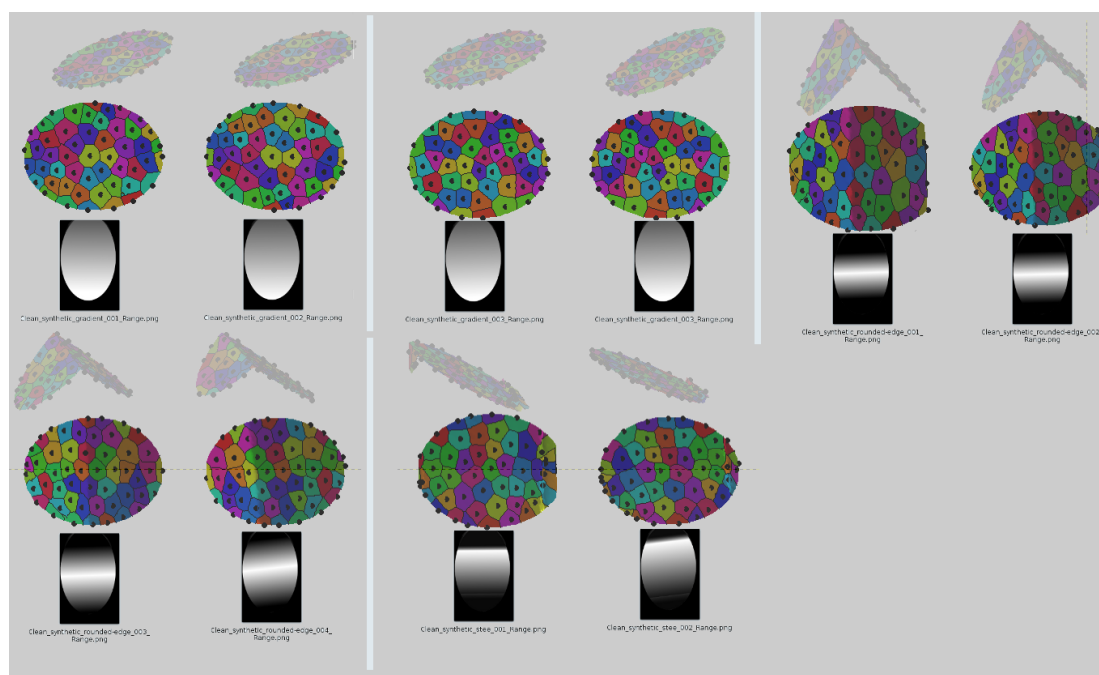


Figure 184: 5 examples of experiments with synthetic data, where the top part shows the pair of images in their classic form, the middle shows a top-down view, and the bottom part is the range image

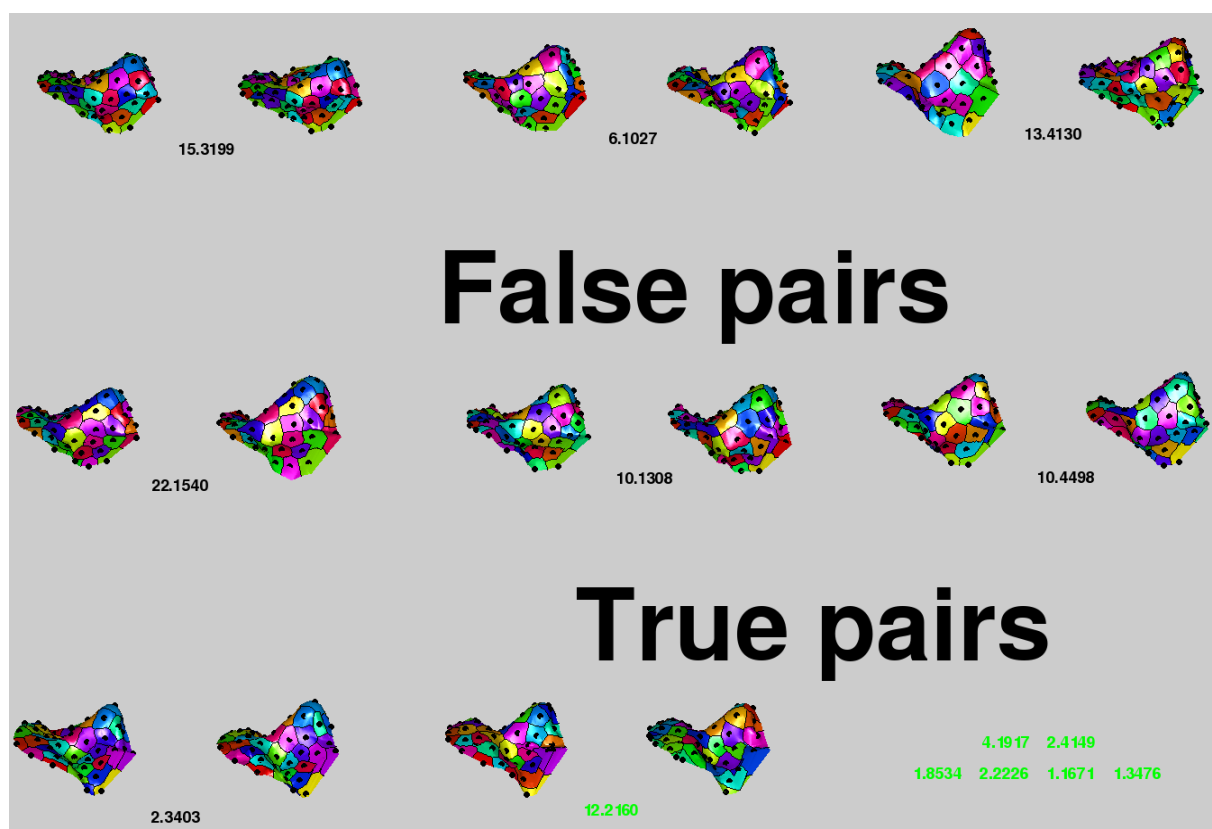


Figure 185: The scores in black show the pairings between different people and in green are the scores of matches between the same person

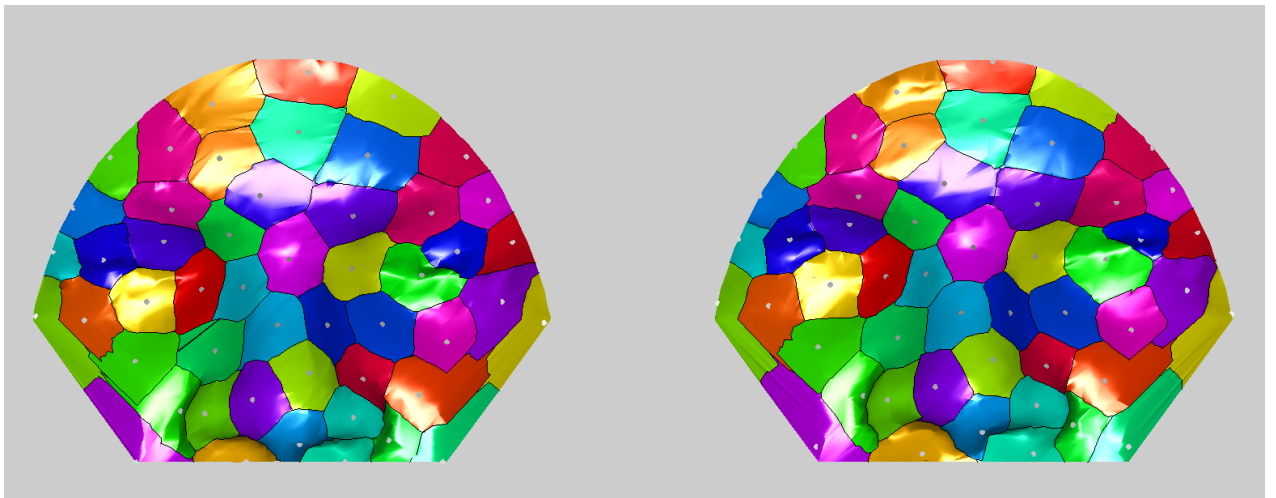


Figure 186: A new visualisation form where the dots signify stress at the given point

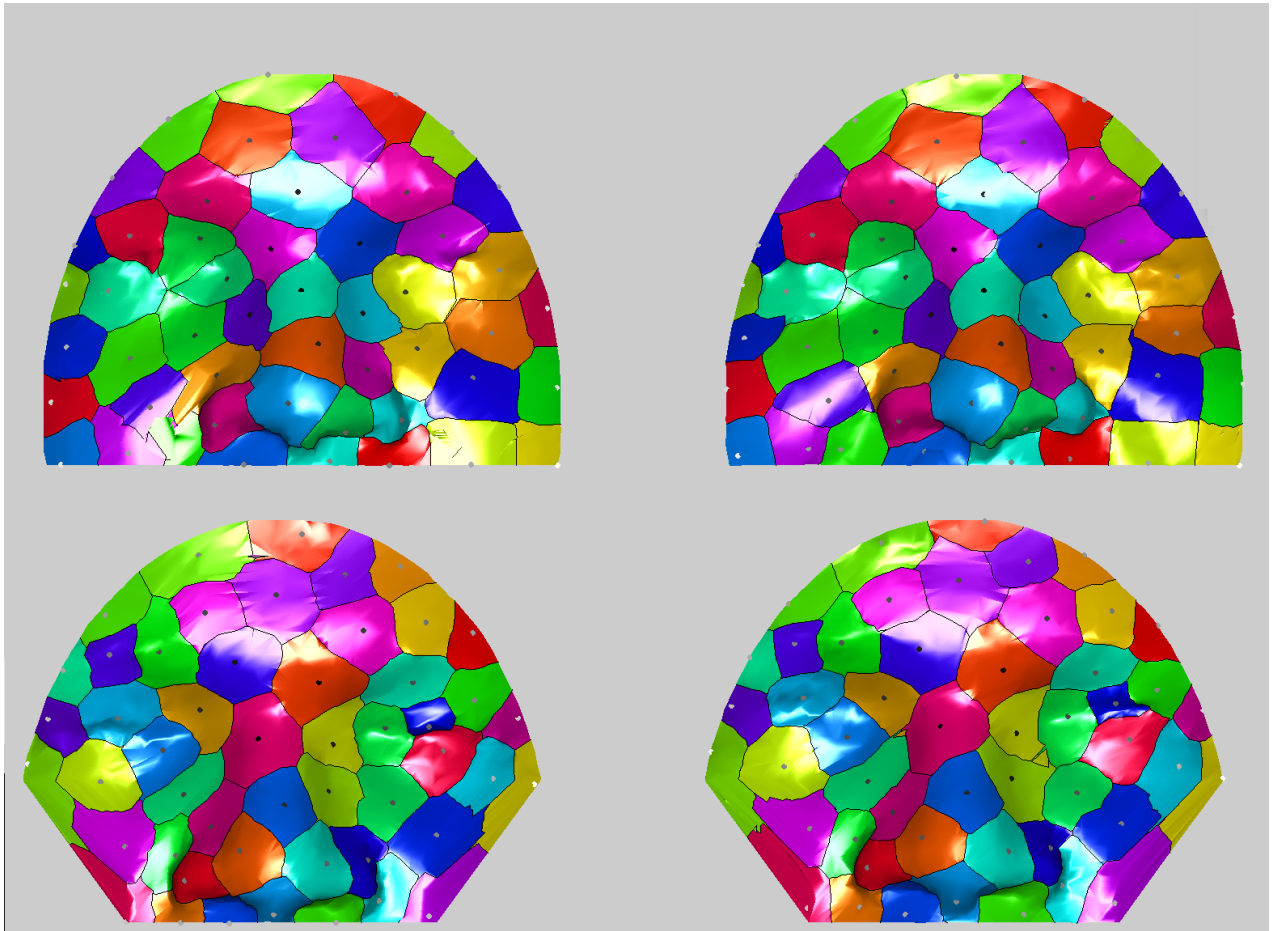


Figure 187: **Top:** A mapping of GMDs stress when cheeks are included in match-finding. **Bottom:** same as above but cheeks excluded. One can assume dark means low stress and white is high stress.

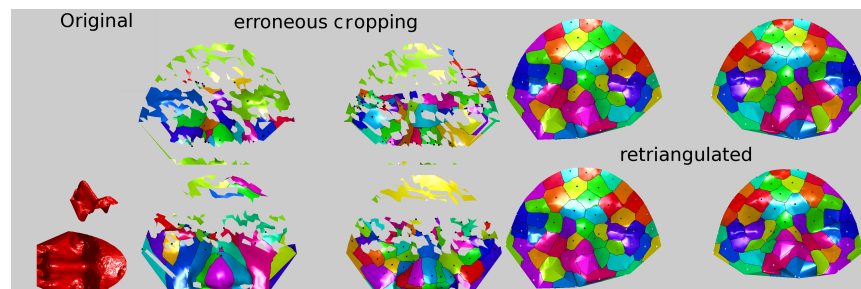


Figure 188: Original images, erroneous cropping effects (still in the process of debugging) and retriangulation of the points after omission.

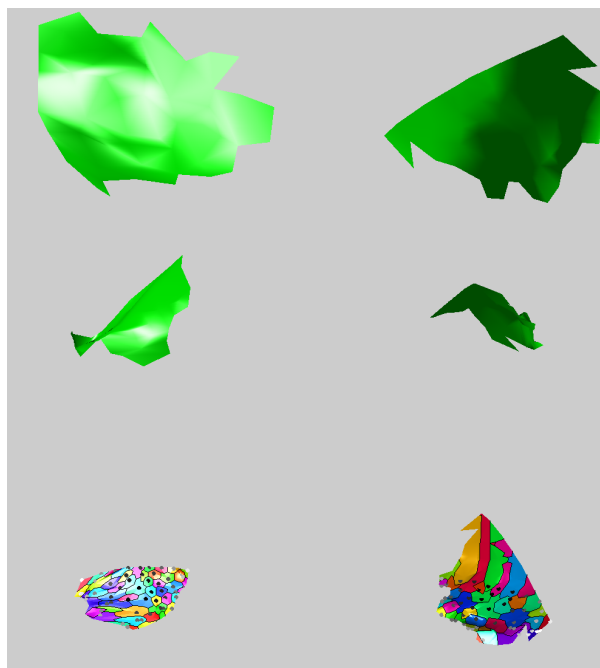


Figure 189: A toy example of a very small couple of surfaces cropped from the centre of a face of the same person, where the pairs shown correspond to top-down view and GMDS' results

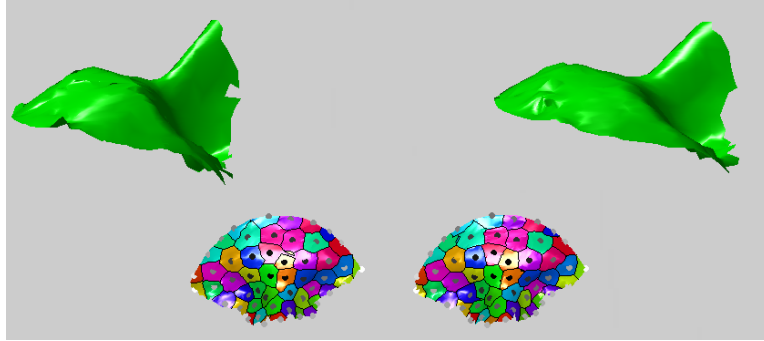


Figure 190: Example of an augmented slice from a pair of faces and GMDS applied to these

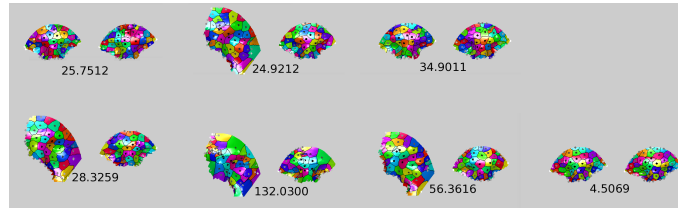


Figure 191: Results of a comparison between arbitrary bits where some boundaries are a Euclidean cutoff and some are geodesic

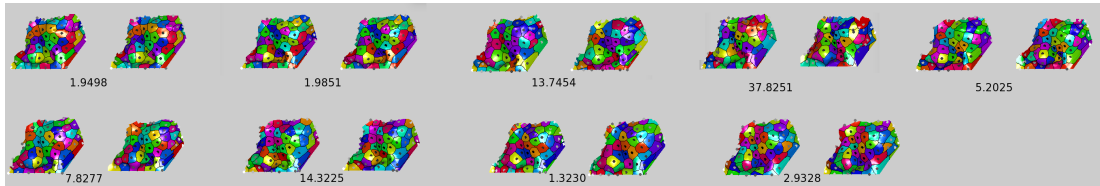


Figure 192: Results of a comparison between consistently chosen bits (near the eye) where some boundaries are a Euclidean cutoff and some are geodesic

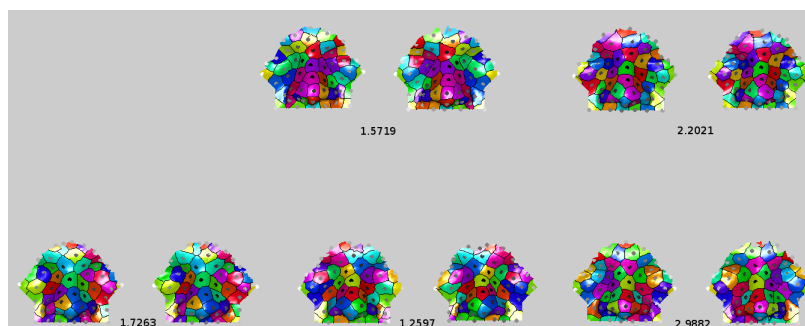


Figure 193: Results of a comparison between surfaces that are mostly carved out of a geodesic boundary

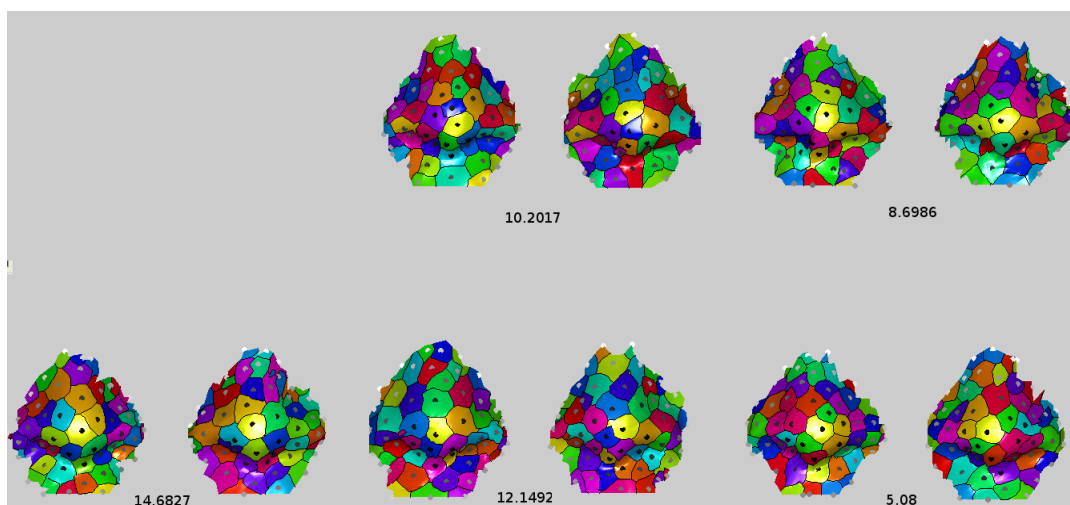


Figure 194: Results of a comparison between noses with a boundary defined by geodesic distance constraints

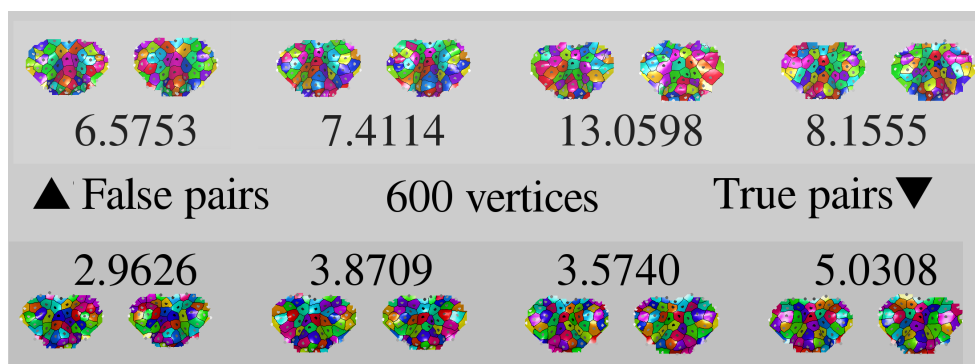


Figure 195: A preliminary look at a predominantly geodesic mask and how it separates pairs from different people (top) and pairs from the same person (bottom)

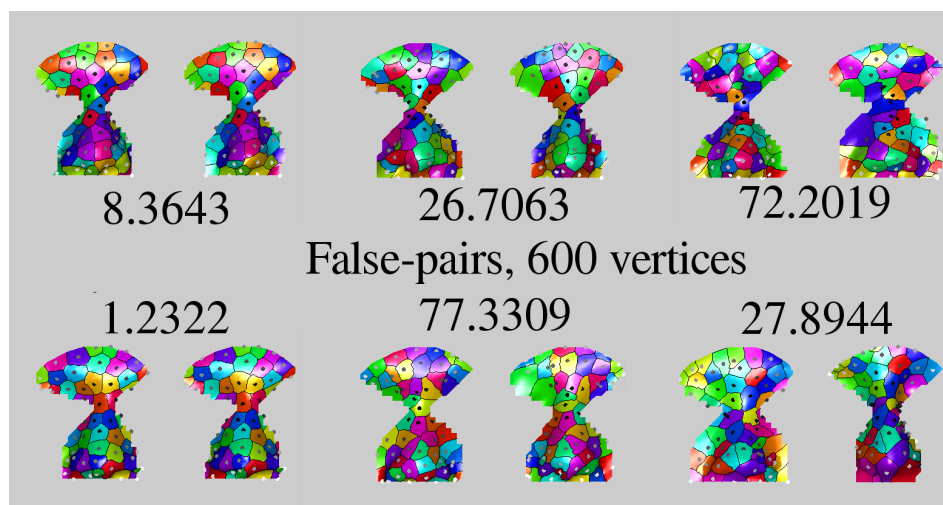


Figure 196: An experimentation with a mask that includes points around the forehead and around the nose

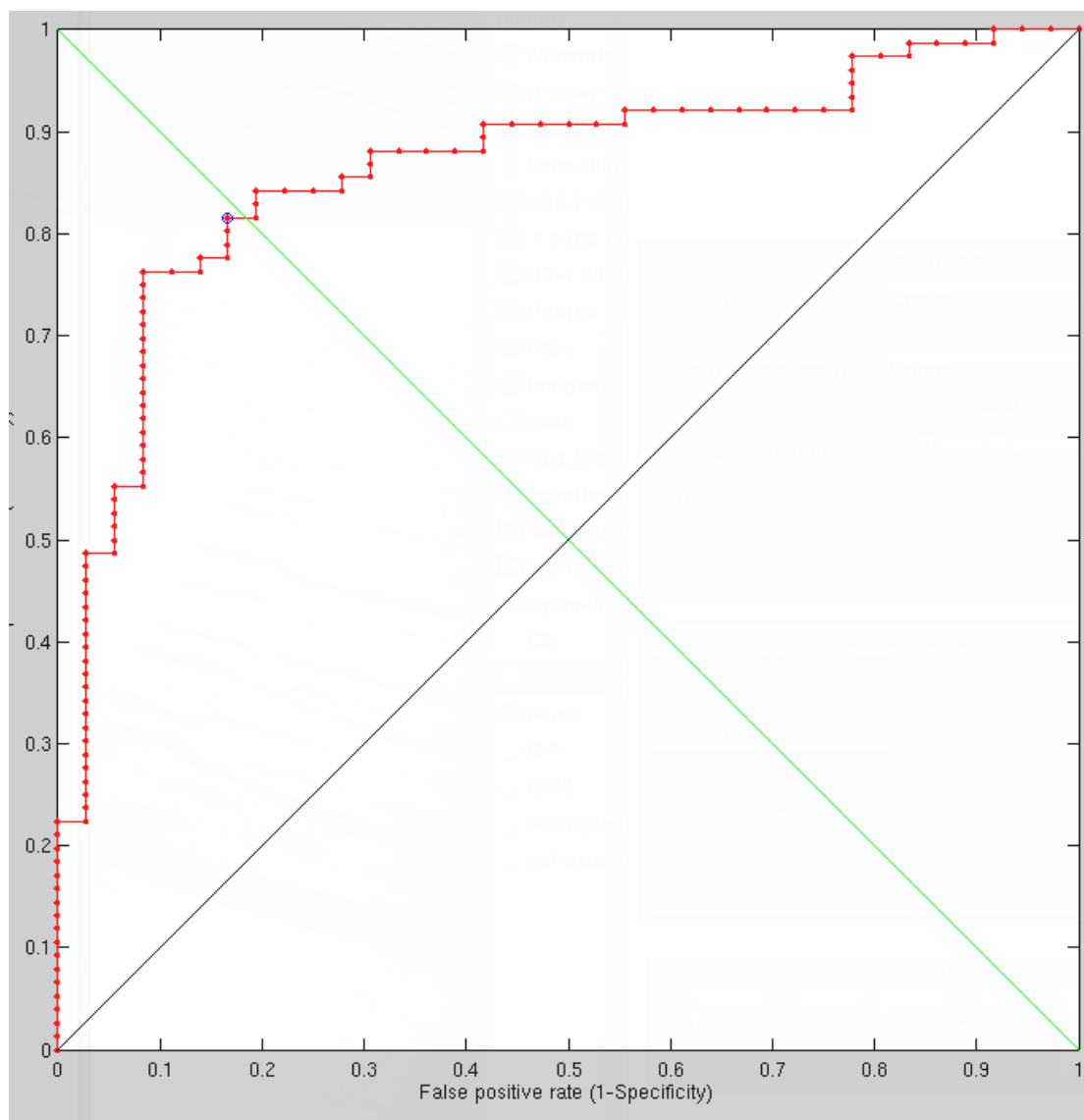


Figure 197: With just 600 vertices, the ROC curve shows unimpressive ability to distinguish between true pairs and false pairs

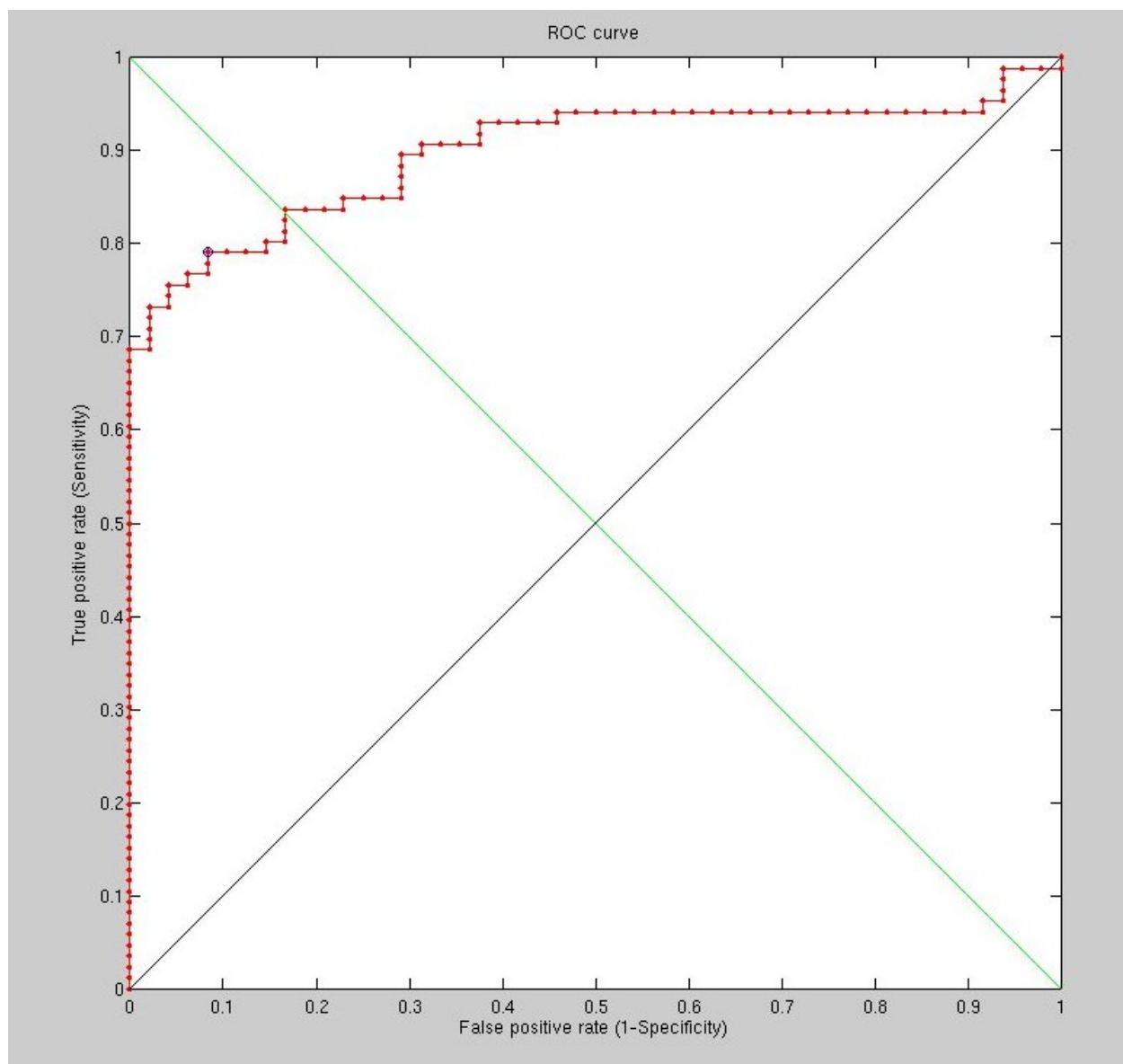


Figure 198: The effect of increasing the number of vertices to 2420.

To define "much higher" resolution, there are several things here and they are all program parameters. First we have raw image signal. Then there is the sampling of the image and also the size of triangles we turn the image into, inversely proportional to their number. The triangulation goes through remeshing and then fed into GMDS, which also can find N correspondences, yielding a matrix smaller than the FMM results. All of these parameters affect speed and some affect stability too. The latest results use 2420 vertices, but the next experiments will look a little at how this can be improved. For instance, changing the geodesic thresholds helps refine the area of consideration. It's unclear how exactly.

The sampling density is still not so high as it ought to be as this leads to some freezes and other issues (if this is attempted, there are also RAM constraints). However, two additional experiments were run to study the effects of changing the geodesic distances around the eyes and nose (see Figure 200). When pushed too far outwards, performance dropped to under 90% recognition rate (Figure 201). It is easy to get a 'feel' for what works better and what works poorly based on about 100 test pairs, assessing the results on a comparative basis. Ideally, however, if stability issues can be improved, then the overnight experiments can be run reliably rather than just halt some time over the course of the night.

At the moment, surfaces are being sampled by taking only a separation of 5 pixels between points. In the past that was 2 pixel (range image) and to overcome this loss of signal one can take a local average, which we have not

done yet in the experiments. It would make sense to try this.

Oddly enough, experiments which test the averaging of range images (for the sake of better sampling before triangulation) actually indicate that the averaging reduces recognition rates (Figure 202. This performance difference has a direct correlation because all the other conditions remained identical. This seems to concur with previous such experiments – those with PCA. How quaint.

Before scaling everything using PCA (to improve the results) it is probably important to ensure that GMDS is performing as well as possible. Right now there is no spatial scaling applied to the score, notably based on the variability of one area compared to another (e.g. rigidity around the nose, unlike the centre of the eye).

To re-define the averaging process we apply, what we mean by averaging range images is basically down-sampling the images, or at least the logical equivalent of that.

When the range image is turned into a mesh of triangles there needs to be a discrete sampling of points and the way this is done at the moment involves taking the points within the geodesic mask, then either sub-sampling those (picking with gaps) or taking all of them – about 50,000-60,000 vertices – then remeshing to make things more scalable. The vertices used vary in terms of their number, usually between 600 to 3000 depending on the experiment. It's $O(N^2)$ for FMM, so there need to be reasonable limits. GMDS is

sensitive and prone to crashes when the points are picked very densely.

The grid on which you we the $N \times N$ operation has nothing to do with the number of points, i.e. we work with the original resolution of the mesh and pick a small number preferably with farthest point sampling points to embed.

To clarify, there are two phases where FMM is involved. One is the selection of the surfaces on which to perform a comparison. This requires a Fast Marching operation which then helps define what makes the cut and what doesn't. Following that phase we are left with fewer sample points (or fewer triangles) to actually run GMDS on. It is then that GMDS in general (encompassing FMM) depends on the amount of data available to it. What's unclear is, what number of triangles are desirable on each surface? Would 1,000 suffice? Or perhaps more that are smaller? The transition from pixels to triangles is key to preserving all the signal. Triangles are a coarser description of the original data.

We keep the original number of triangles for the whole process. But we can have GMDS work on 5 points distributed (preferably with FPS) on a triangulated surface with 1000000 triangles. It's true that in the preparation step, in principle, one may need to compute the all to all geodesic distances within the surface. But this is wrong to do and could be done on the fly. That is, if indeed we have 5 points trying to re-locate themselves on a 1M triangles surface, then at each iteration we need to compute only $3 \times 5 \times 15$ inter-geodesic distances. If we sub-sample the surface, we usually introduce

a non-geometric process that could destroy the similarity.

The high vertex/faces density we require is due to very fine features that are identified and paired/matched when GMDS is used as a similarity measure. For a region of about $300 \times 300 \approx 100,000$ pixels to limit ourselves to just 1,000-9,000 vertices might not be good enough. Recognition rates are not satisfactory in comparison with state-of-the-art method that take account of denser, finer images with higher entropy. If we compete using an essentially down-sampled image (due to scalability inherent in GMDS), then we throw away a lot of information that can distinguish between anatomically-identical surfaces (be these derived from volumetric data/voxels, camera, or whatever). The bottom line is, after just over a month working with GMDS (since end of July) we are unable to get around this caveat of scale. Many workarounds can perhaps be tested, but none will actually permit the application of GMDS to high-resolution range images, which are what leading algorithms utilise for greater accuracy.

7.11.2 Scale Issues

As for the limitation of number of vertices - one can safely increase the limit in the program, but only to some degree. When it's too much, FMM crashes for some reason. The C++ version allocates a /contiguous/ chunk of $0.5N^2$ doubles for the distance matrices, which allows SSE but is a serious disadvantage for high values of N , because of memory issues (N is

the number of vertices on the mesh). The largest mesh previously used was ~14,000 vertices, and only allocating the memory took at least 10 minutes (the operating system had to move stuff out of the way). That's obviously not the way to get (distances should be computed on demand and cached), but it is what the coder wrote at the time.

14,000 would give us roughly 120x120 points. It will be good to see if this can improve our results and, if so, to what extent. The artificial cap on the number has been raised and it does take a lot of time to allocate memory. From a purely research-oriented point of view, it will be interesting to at least check the feasibility and suitability of this approach, regardless of performance in terms of efficiency, at least for the time being. In some previous experiments we needed to essentially down-sample the surface into a 30x30 or so grid (not grid per se but vertices), which makes a high-resolution image almost 'iconised' (32x32 by some conventions) and therefore performance was poor. This is similar to the limitation encountered when applying PCA in image space, having to reduce the covariance matrix to something manageable by the available memory, e.g. 1000 observations or 30 pixels/points by 30 pixels/points. The PCA approach was also applied to the gradient, we tried a hybrid of signals, e.g. intensity/depth signal with derivative, and ideally we foresee a combination of GMDS and PCA, enabling scaling along particular dimensions to be carried out dimensionality reduction the classic PCA way. This is still a subject of debate.

One might guess that a modern machine with 8 GB ram cannot run GMDS

on a mesh with more than $\sim 100,000$ vertices. One possible solution would be a multi-resolution or coarse-to-fine approach. We might, for instance, wish to first view the problem from above, watching general topology, then pick particular segmented (by GMDS) sub-parts and apply GMDS to those for a finer similarity assessment and score (ensemble). FMM is already being used to carve out surfaces from the whole, depending on the location of easily-identifiable points and geodesic circle/s around these.

Having spent several hours learning the behaviour scalability-wise, it seems safe to say that: 1) stress minimisation is faster by a factor of almost 10; 2) stability issues that are caused by the remainder of the GMDS process limit one's ability to run unsupervised experiments, especially when the number of vertices is high; 3) increasing the number of vertices may often lead to a sort of program freeze, from which the only apparent route out is killing of the whole MATLAB process (PID) or disconnecting. This is not necessarily related to the aforementioned stability issues. Going above 4,000 or so vertices almost always results in this problem, which is not trivial to debug. It is likely correlated to RAM being almost fully exhausted (it is a shared server) and swapping being used spuriously.

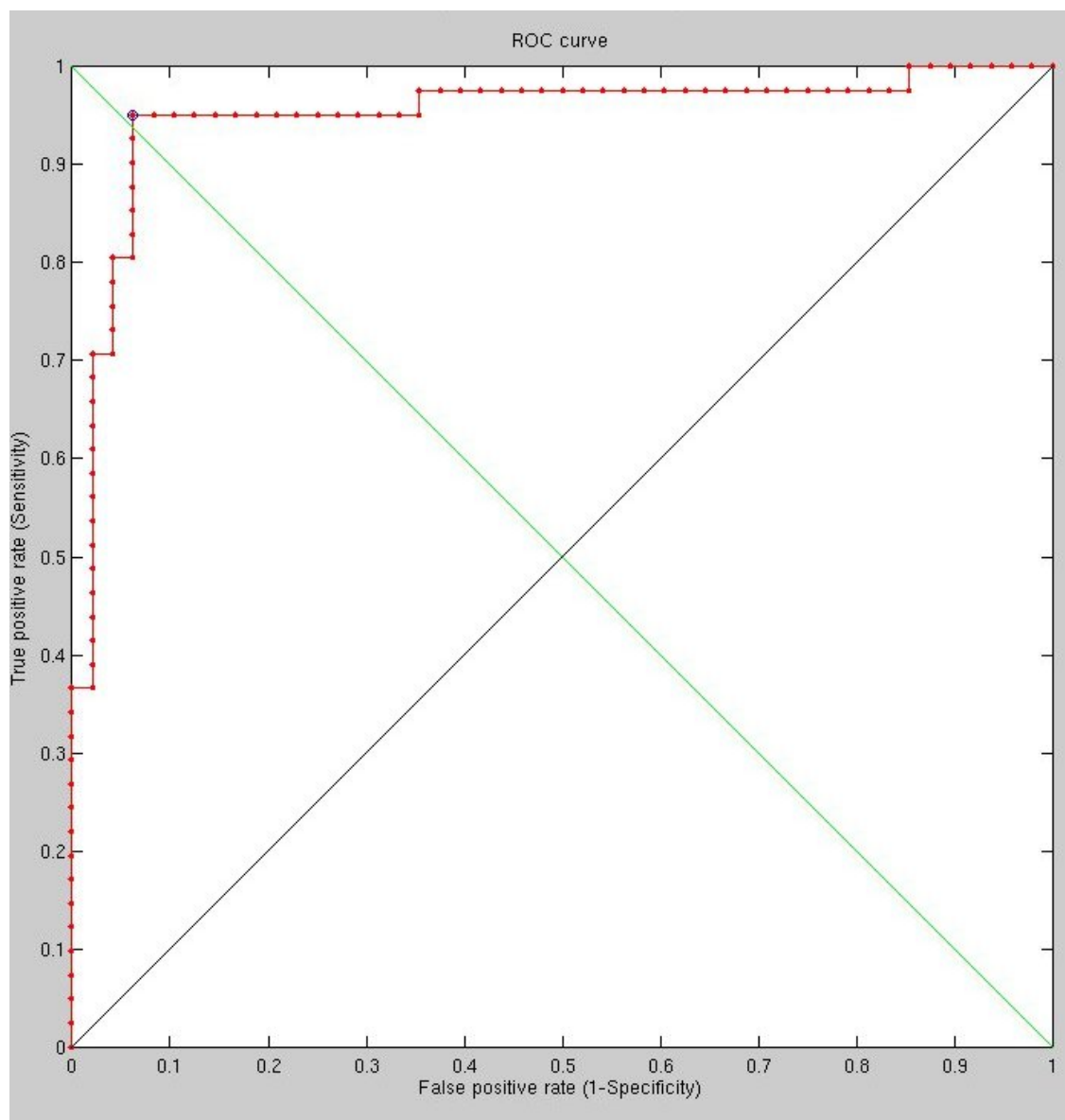


Figure 199: Improved performance with slight changes in surface size for the gallery

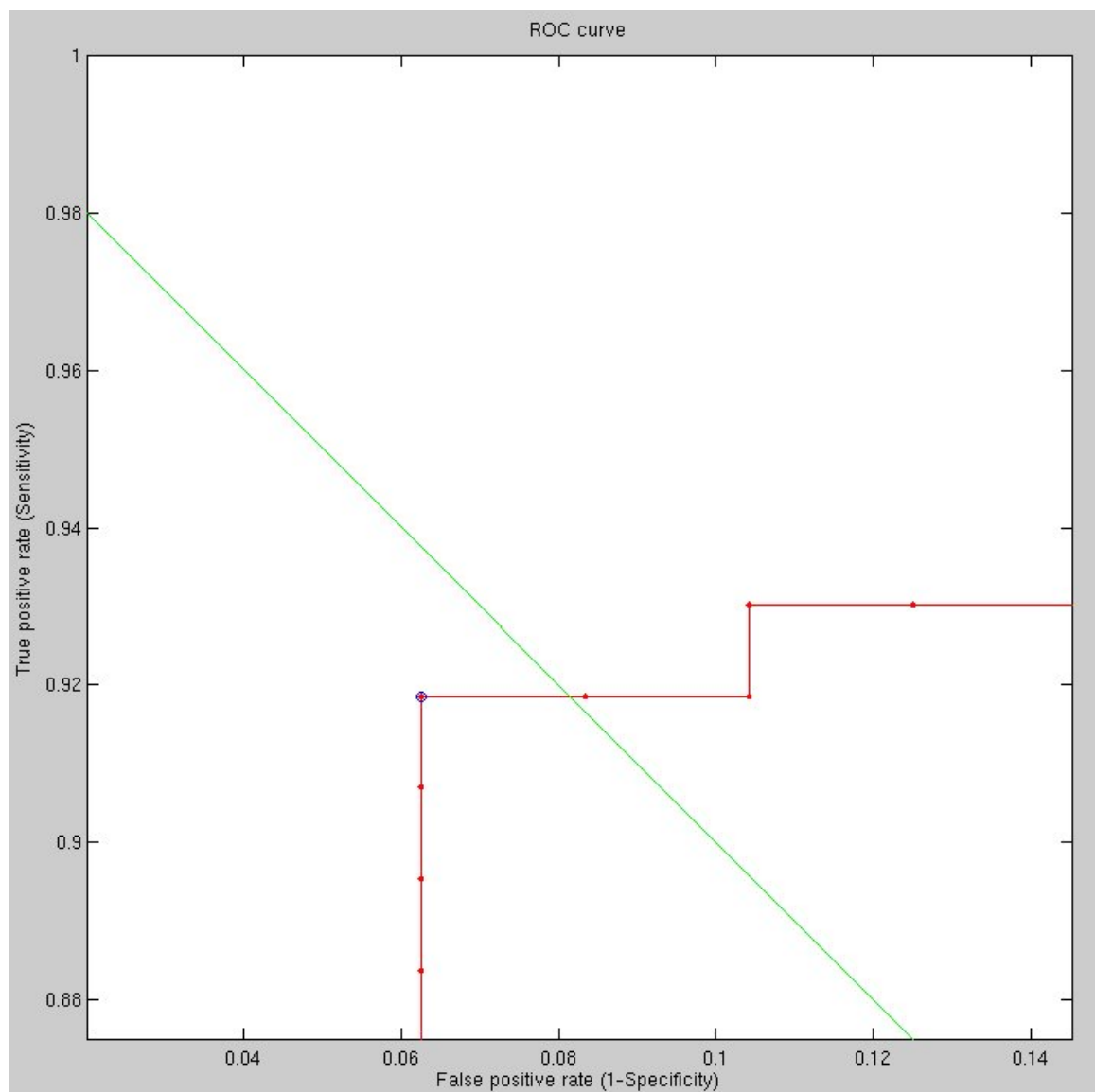


Figure 200: The result of changing the border threshold for surface carving

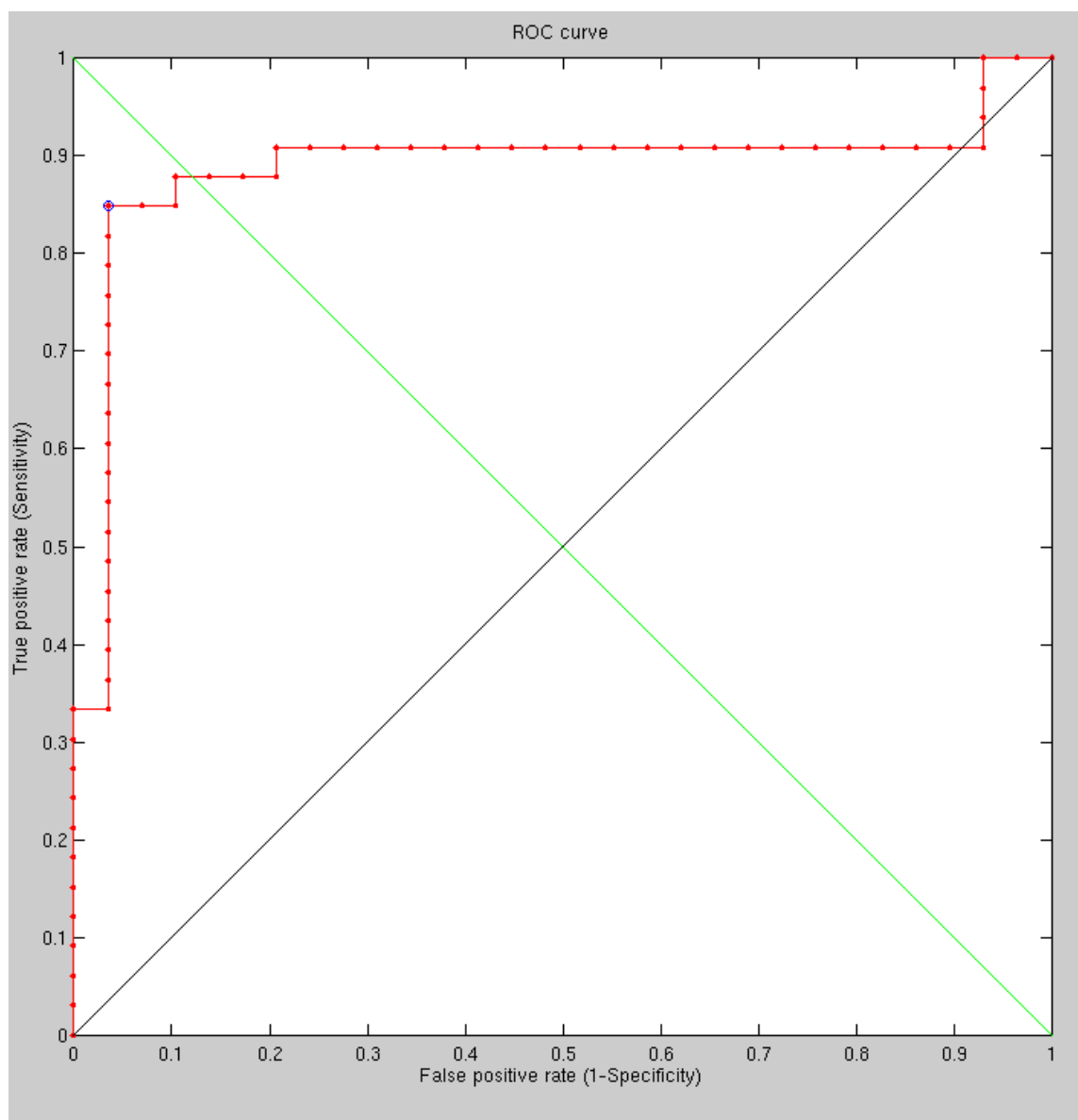


Figure 201: The result of growing the surface too big

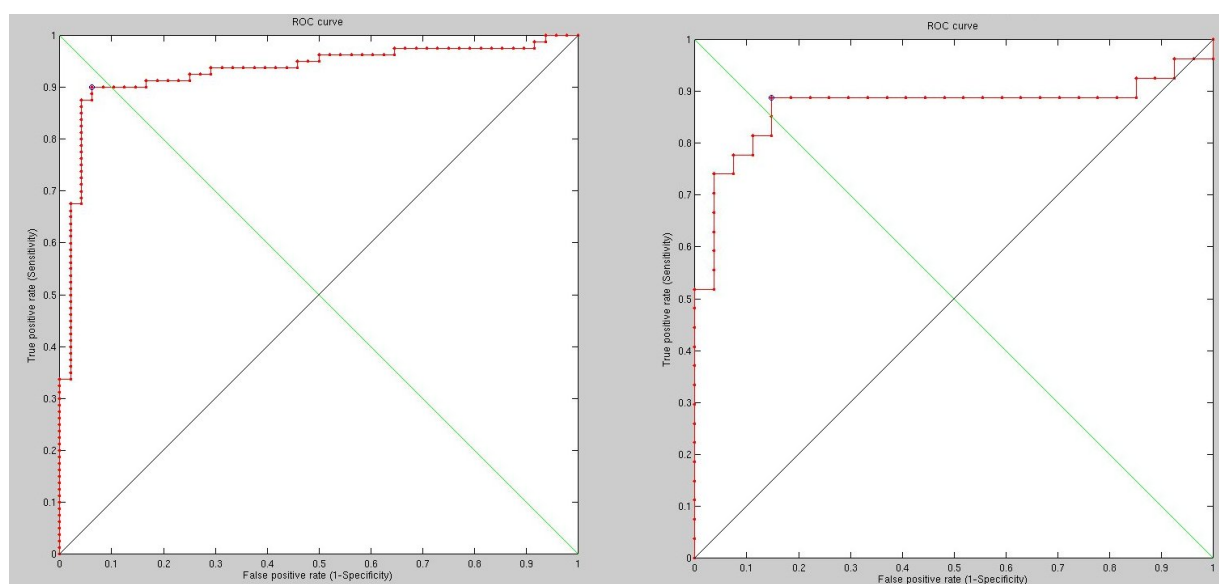


Figure 202: Performance with (left) and without averaging (right) of the arrange image for better sampling of the GMDS process

Our performance depends heavily on our ability to handle full resolution images, as opposed to a sub-sampling of them (without interpolation). GMDS almost always gets the general topology right (PCA too performs at similar levels), but for inter-person distinction we must begin to think of a better approach that scales gracefully. A lot of information is unaccounted for.

The distance computation should be performed by demand and not as a pre-process that computes all to all distances on the given mesh. This would immediately make the whole procedure function at a complexity related to the number of points and not the mesh resolution.

7.11.3 Improving GMDS

While we don't have a readily available code for GMDS with on-demand distance computation, the plan is to implement this. We do have the code for efficient fast marching and software cache that combined together and plugged into the old code (or better, the C++/new code) can do the job. At present, the way we utilise GMDS is the hybridisation of the C++-based stress minimiser, the C implementation of FMM, and `gmfs.m` as the wrapper which puts together those fast implementations of key pertinent parts (substitution of analogous MATLAB implementations with slightly different interfaces). Currently the bottleneck is associated only with the number of sample points that make up the two surfaces. The MATLAB profiler might provide insight into what part exactly takes up all the RAM/CPU and to

what extent. Another concern has been (and verified by others too) that when the number of vertices is increased, the program becomes more prone to freezes/crashes that are neither consistent nor easily reproducible. Carmi and I never looked too deeply into it, but overcoming these limitations would definitely make GMDS more featureful and robust, not to mention suitable for a plethora of new applications where resolution is high (detail at micro-level and not macro-level/topology).

it remains to be decided what language we will want this implemented in. MATLAB would obviously not be as fast, but currently, considering the scale of the experiments and the access to 2 computational servers with 8 cores each, it probably would be simpler and also work seamlessly across platforms as it's being interpreted. Modifications have already been made to some GMDS code, e.g. to highlight stress in Voronoi-style visualisation, so the task of handling MATLAB code is less daunting than education oneself about the object/class system in the C++ case.

To summarise the situation, hitherto we've investigated the potency of GMDS as a fine-level similarity measure for pairs of surfaces that are carved around a central reference point with geodesic circles around it, or even several such points (the union of those). There are two unique aspects to this approach, one of which is the implicit assignment of weights to distance variation based on PCA, applied by studying local variation in a model-building stage. One other aspect is the hybridisation of measurements, e.g. Euclidean and geodesic using principal component analysis. In GMDS, the main draw-

backs are scale and speed. To an extent, speed was improved significantly owing to the C++ implementation, however it remained unclear how to handle a very dense tessellation of the surfaces – dense enough to preserve the amount of information inherent in the original range images. One caveat was the repeated computation – and optimisation over – geodesic distances. By reducing the complexity of this process, e.g. by stochastic selection of points or by caching some certain distances (maybe propagating one onto the other more efficiently, where appropriate), it seems likely if not just hoped that the process will scale better and perhaps deal with even 50,000 triangles. At the moment, trying more than a few thousands leads to technical problems. This problem is reaffirmed by others who encountered the same bottlenecks. The limitation means that general topology can be determined very well, but at a very fine scale where distances are minor and not topological, the problem needs to be downsized, i.e. the triangulation made considerably coarser. While we get promising recognition (pairing) results, in order for them to be close to perfect it is probably necessary to use the full information and not a sub-sampling of it. Right now we use about a tenth of the available entropy which tells about one surface from the other.

Alex B. ran few experiments with the ideas of software caching and on-demand computation of distance fields and was able to cope with around 10K vertices in MATLAB without particular difficulties. He thinks we ought to explore this direction. If we have some of the modified/hacked code around, even in very rudimentary form, then this is a reasonable path ahead. Maybe

we can try to re-plug that in and tidy up a bit, then return it for consideration in upstream inclusion.

According to institutional regulations, the end of October is the final date for your temporary employment. There are two options for continuing this project and one is to continue making gradual progress. We need outside support for that resolution issue. that is partly implementational We have been speaking to Alex regarding resolution and are quite dependent on knowing whether we get it done because other paths of exploration – while viable – are not ideal given the trajectory we are after, namely PCA and GMDS combined and a performance decent enough to bring about something publishable.

The only limitation at the moment is the resolution and it is not just a performance/duration issue if RAM constraints stand in the way (which they do). Had it been possible to run full-resolution experiments very slowly as proof of concept, then the practical benefits would be possible to assure, i.e. it would be possible to show that given necessary optimisations, the required outcome will be good and also rapidly reachable. As attempted have already been made to code the necessary improvements, it would be unwise to code them from scratch again, which is why we waited for Alex and gently reminded him about it.

It ought to be added that all prior results and ROC curves also used low resolution images because in the case of PCA, for instance, RAM was exhausted

if the sample of observations, I , was too large for principal component analysis (covariance matrices get vast due to a quadratically-increasing order of complexity). Values of i were selected based on a down-sampled and at times averaged representation of surfaces. This served as a baseline and was consistent with the emulated method.

Dealing with multi-dimensional problems at such a high scale and still amassing all the fine details may require a multi-grid approach and it is easy to envision how this would be implemented.

At an imminent point we have to change you status from temporary to research fellow (at the end of October. The cost is a bit higher and there are some local social benefits associated with it. Progress needs to be done and seen.

Main ROCs or something of that sort should have been our base line. All ROC curves seen so far are produced based on a recognition task of the order of magnitude of icons, i.e. $32 \times 32 \approx 1000$, where those sample points are either triangles or cloudpoints (i.e. 3-D coordinates, not just pixel value). The way we approach this problem is adopting the supposition that we can encode surfaces compactly and then use these compact descriptions (e.g. geodesic distances) to tell surfaces apart. The problem in this case is clear; while it should be possible to sample a distance on the surface, without having high resolution at hand the distances are sampled on a coarse grid and therefore they become imprecise. In order to overcome this, density needs to come into

play such that it uses the high-resolution image and then yields a shorter description. Moving from micro to macro would help here, not topology-wise, but measurement-wise. Multi-grid for GMDS is something we would definitely like to explore. One possibility that would be easy to consider is as follows: we take an overview of the problem to get topological information. This can be done reliably in many different ways, even with GMDS. Then, by taking smaller chunks of the problem and applying GMDS to them we can potentially perform measurements with the full resolution (more triangles) in tact.

There are some issues with the current way we interpolate distances within the GMDS. Distance interpolation may violate the distance properties. This could happen at a much smaller scale than being significant, yet for delicate comparisons it could be a pain. There are some remedies we are aware of (like replacing geodesic distances by diffusion distances, in which case interpolation is done on the eigenfunctions of the Laplace-Beltrami operators rather than the actual distance, so distance property is always preserved).

While diffusion distances would be more robust, if they take as their starting point something which is spaced out too sparsely they will fail to discern anatomical differences that are not just sub-pixel scale but also multi-pixel scale. The curse of dimensionality has always been our greatest enemy here and perhaps de-emphasised all along. The way GMDS works quite strictly requires that this limitation gets taken into account, at least in its current implementation. The fact that others suffered or at least encountered this limitation (with ongoing implementation attempted) means that it's a real caveat that, if properly addressed, widens the applicability of GMDS to a plethora of new tasks.

There is still work to be done even with pure geodesics. The geodesic measures may be fantastic, but it is not their fault that they operate on a poor gridding system or a coarse mesh when handling a task where topology is almost always the same and the real changes are very minor.

7.12 Planning for Final Stages

In the following brief, action items are listed for the near future, explaining how we finish the face recognition project successfully.

The steps to explore next are as follows:

- ▷ The implementation and use of caching for scalability.
- ▷ Experiments with an increased number of triangles using:

1. FRGC database
 2. Texas3DFR Database
 3. Potentially expressions data from GIP and synthetic data too
- ▷ Increase the number of vertices from at most 2420 to around 24200 and maybe 242 in order to demonstrate, graphically, the effect of changing resolution on the ability to distinguish between surfaces such as faces, where variation is milder.
- ▷ Perform an analysis based upon results from GMDS where Fast Marching is applied to a greater number of points.
- ▷ Combine GMDS with PCA and potentially with Euclidean distances too, hopefully demonstrating the superiority of this approach where a model is trained to learn sources of variation in entire sets of surfaces, e.g. what local regions vary more than others across different individuals as opposed than within an individual (intra-person, mostly expression).
- ▷ Provide a detailed (and yet rather concise) explanation of:
1. Why and how the curse of dimensionality affects the problem, how it can be overcome
 2. What other approaches were attempted, notably the purely PCA-based ones

3. How the algorithms work and what results are obtained, bearing pitfalls like great scale in mind
4. Which perimeters were altered and which values gave improved results

The code for caching is an adaptation of the predominantly C-based implementation of FMM, with hooks that use as interface a caching query (written in C++), all wrapped together in MATLAB for calls from the outside. Based on a quick run through some debugging code, backporting the changes to a 2009 GMDS implementation should be doable. The code is very much separated from GMDS-specific functions that are sufficiently modular.

The past fortnight was mostly spent waiting for the caching work. It was not done in a matter of days as had been wrongly assumed, so we have begun implementing a multi-scale approach where slicing of regions gets done at one level of granularity. The edges of the triangles are very clear to see in Figure 203 because of the small number (1000) of triangles for the face region as a whole. It should then be possible to increase the resolution and work on regions in a piece-wise fashion. GMDS would scale better. The tricky part then becomes the accurate annulment of triangles outside the geodesic circle of choice. That's why the piece-wise approach is likely to lack accuracy, but the imminent results will provide some insight.

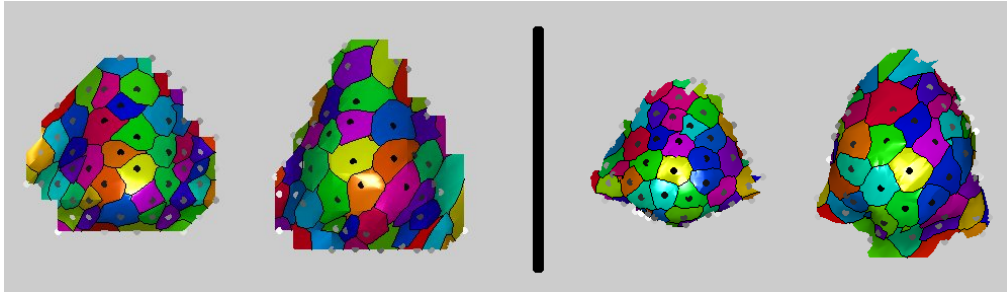


Figure 203: A look at slicing at geodesic boundaries around the nose tip, with coarse resolution on the left and improved resolution that isolates regions on the right

7.12.1 Caching Code

Moving on to some proper caching, there are issues to overcome. As we don't have the code entirely, recovery was needed. A software cache in C with matlab interface was used along with fast marching in C (with matlab interface supporting separate grid initialization and computation). The idea is to get a better resolution approach working, but it requires studying our programs as we then need to merge these changes with the mainline. Having re-fetched the latest version, we shall merge.

To summarise modification with a quick report, changes were as follows. Taking one portion at the time, we have been find'ing, diff'ing and grep'ing corresponding files that come from GIP SVN and the SVN tree at TAU, which comprises ANN, Cache, FastMarching, and Remesh (a subset of the whole).

Cache and FastMarching are the main components that have been adapted

to facilitate caching and upon a cursory check, other parts too have some interesting inherent changes that might cause incompatibilities, probably a result of many people – myself included – changing the code somewhat and hacking our code accordingly. Compiling on the server is slightly complicated by the fact that the server uses 64-bit Ubuntu and not Windows (ANN is already compiled and remeshing works too). Currently, code is being changed somewhat to make the rest compilable as everything else has been prepared and tested, so it ought to take hours to mend.

We finally compiled FMM successfully, although with the following compromises:

`__inline`, `__forceinline` etc. got removed as the compiler does not support them. It ought to be possible force 'inline' with GCC or ICC, but it would take further work that may not be a top priority.

The `StrCmpI` function (Windows) is not supported, so we use `StrCmp` instead. The same goes for some preprocessors and type definitions that needed to be rewritten in an ISO C++-compatible way.

Finally, attempts to cast fmm to 'unsigned int' made the compiler very angry, insisting that it loses precision perhaps because it assumes we are compiling 32-bit code on 64-bit machine. So anyway, we changed that to long.

If any of this is critically problematic, this might resurface later.

All the relevant code has been successfully compiled on GNU/Linux (after many minor modifications) and joined together with older code, including

the C++ stress minimiser. Debugging is necessary as there are still program crashes (binaries taking down the whole MATLAB session) and some opaque errors at times. An 'init' called in Fast Marching works as expected, but 'march' causes issues. Heading back to the case example tested with a sample surface, it gives the crash reports which are still under investigation.

Debugging

The fast marching code hosted by TAU insists on a triangles array that contains int32 data, unlike the analogous GIP versions (there are several lying around, which can be very confusing). Some use double and casts spuriously. Insisting on trying to make it work resulted in freezing (infinitely-running loop or another unrecoverable error) as opposed to a crash (both fatal), so there is no point to debugging using that path.

Trying to investigate things a little differently, another dataset – this time the Swiss role – is tested using the newer/enhanced FMM binary.

```
>> surface=swiss
```

```
surface = TRIV: [1024x3 double] X: [561x1 double] Y: [561x1 double]
Z: [561x1 double] D: [561x561 double]
```

```
>> which fastmarchmex
```

```
/home/schestow/pcafaces/Imported/cached-GMDS/FastMarching/fastmarchmex.mexa64
```

```
>> f = fastmarchmex('init', int32(surface.TRIV-1), int32(surface.X(:)),
int32(surface.Y(:)), int32(surface.Z(:)));
```

this one too 'freezes', but works OK when cast as double. In fact, the whole process works OK, so the binary I created this morning (based on Kimmel's implementation with support that separates initialisation and computation) certainly works OK for certain data, just not for Michael0, at least not at this stage. The important thing is that it does work and helps produce nice graphics (more on that later).

Experimentation with the improved Fast Marching binary showed a consistent behaviour where the expected results are sometimes obtained and sometimes the program just crashes. More specifically, while it can be shown that everything works for simple data such as the Swiss roll, for more complex data such as faces from the NIST (FRGCv2) and Texas datasets (newer Texas3DFR), the program sometimes uses up 100% of the CPU core and hangs in there with an error indicating non-ending loop in the B&B algorithm (see below). Upon a second attempt – after recovery by interruption

of the process – the program crashes at an earlier stage (apparent hardware exception), which is probably not as fatal as by that stage the program is already in an unfamiliar state. The successful runs with simpler data are reproducible, so further testing is likely to help resolve this.

Narrowing data paths down a bit, it is clear that while FMM produce the expected results (shown in Figure 204 is one triangulated mesh, which is as coarse as it has always been), it is the coarse correspondence stage which struggles to complete. This is most likely a compatibility issue which further testing will resolve shortly. Some types and interfaces have changed a little and this needs merging elegantly. It is not entirely clear how to deal with the case of "multiple" or "single", for example, without building two separate binaries (as currently stored in SVN).

Note that the continuation that you make by linking the cheeks (with very long) triangles may be painful for the FMM.

When running experiments I always move the GUI sliders to include the cheeks (convex and not concave). Some toy examples still exclude the cheeks. Today was spent narrowing down the cause of the crash probably to a recursion whose exit condition is never reached. This recursive procedure can be limited by two or three unfolding of the triangles. If these are good triangles, then it should work. Else, something could happen with the boundaries. It takes a while to debug because each failure is fatal and requires restarting the whole of MATLAB. Two desktops (spread across three monitors) and one

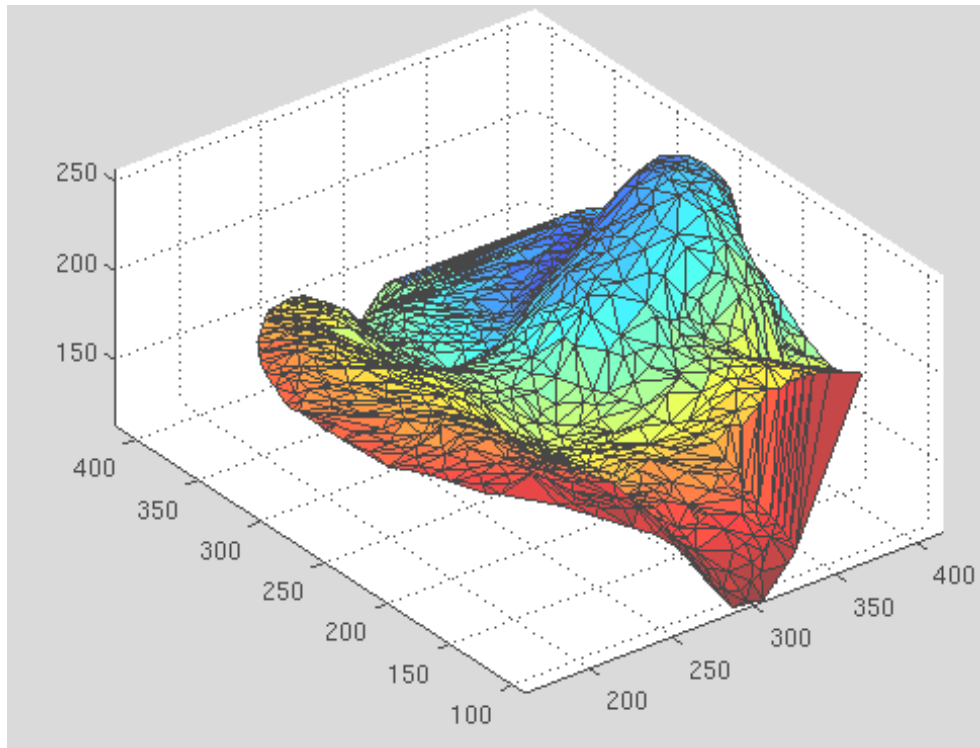


Figure 204: Newer Fast Marching algorithm as applies to a face from the Texas database

GIP server are used for debugging this. While one crash occurs, the other desktop is meanwhile loading up another instance of MATLAB. Each crash is verbose but gives very little information that is coherent or programmer-readable. So the testing is very trial-and-error-ish by nature.

One is a binary optimised to calculate geodesic distances of every mesh vertex from one source. The other case deals with distance matrices of pair-wise geodesic distances. In the implementations, "multiple" and "single" means distance map from a source to all points vs all-to-all. We need the first one among those two. They are the same essentially, and it is important that we

init once, and then do the computations. Grid init takes a lot of time and is valid for the entire life of a mesh.

While the full history of implementations diverging is unknown without having spent years watching the code, the most glaring difference between the version in GIP repositories and the one in TAU is that the interfaces vary somewhat. One uses an explicit init call, whereas in the other it is implicit/internal and there is also a subdivision into the two not-so-distinct cases, at least at a binary level (the MATLAB interfaces parse input and interpret which case is to be invoked).

The way the code has been modified ensures that the intermediary MATLAB file initialises explicitly and then calls the new binaries with the casting of types so as to make everything compatible. What remains unclear though is whether or not other helper functions too have been made inconsistent. Having spent the first hours/days looking at differences between identically-named files (e.g. in meshing), it seems clear that there are some unilateral changes.

Additional complication arises from the fact that the C++ stress minimiser has its own 'hacked' version of surface initialisation, which basically adds some surface property specific to the operations it needs to perform. A lot of the functionality is hookable, but the hooking mechanism is not modular, which means that the symbiosis requires manually adding/replacing lines of code. In the case of surface initialisation, there are several variants out

there all sharing the exact same filename. Ideally, all code changes would be applied to one single point, perhaps with conditional statements that facilitate everyone's needs. Otherwise it increases complexity and complicates testing in general. The FMM MATLAB executable, for example, has several names, such as `fastmarchmex`, `fastmarch_mex`, `fastmarch1_mex` (each doing something different and using different input arguments, requiring multiple stages/phases/function calls, etc).

Initialising moderate sized surfaces for `fastmarchmex` (TAU SVN) does not take long. Currently, provided nothing crashes, the problem occurs in `bnb`, which never terminates and does not provide much feedback indicative of the actual problem. Trying to add simple debugging cruft to the function somehow results in crashes every single time, which is mysterious and time-consuming. Attempting to debug in other functions, even by altering to-be-interpreted code rather than to-be-compiled code, leads to more persistent crashes. This is mysterious behaviour, but a segmentation fault might explain it. Given that the state of a program (e.g. what it ran beforehand) can determine whether it will crash or not, reinforces the suspicion that there needs to be refactoring and it is likely that functions entirely separate from triangulated FMM are somewhat baffled, e.g. by input data types. One function deals with doubles and another with 32-bit integers. This is entirely separate from the aforementioned problem compiling the code for a 64-bit machine, where pointers to memory were insufficiently expressive (it would be best to use versatile `UInt_Ptr` or `Int_Ptr` for those).

Here is what is known for sure. The new Fast Marching implementation works correctly and without difficulties on simple data. This morning we managed to make it work correctly with michael0 as well (this is the full body surface, as shown in the screenshot inside Figure 205). Once the function is applied to Texas/NIST data and output of this process is piped into the coarse correspondence phase, something is amiss. Moreover, stability is somewhat of a concern, although it's not yet clear if memory allocation plays a role in that.

It would be most preferable to understand how implementations of FMM vary and how to tackle the fragmentation so as to reduce or altogether eliminate complexity (somewhat daunting to someone who is unfamiliar with these SVN trees). That probably ought to include some consensus on data types because functions that include precompiled binaries come to depend on these (preconditions). If the cache from Patrick Audley is 'merged' in (or made dependent rather), then it would be ideal to make it usable with all different FMM cases – everything in one fell swoop.

Management of the source code could probably be improved somewhat. It seems like everyone still uses the TOSCO implementation of GMDS (2009), which is essentially a demo. Corresponding files exist on the GIP storage server but probably not in SVN. This means that a lot of collaboration on merges remains a missed opportunity. Sometimes changes are applied to SVN which render particular source files impossible to compile (as happened a few months back), which shows the downsides too.

The nature of development when each person pushes a reference implementation in a particular direction to study the impact (researching by modification) means that a lot of code gets changed and stored locally without being pushed upstream. For some, this means having all sorts of mutually incompatible but similar chunks of files with only one central point (e.g. Web site, SVN tree, or several if different departments have different SVN repositories) from which to check out an elegant implementation which was more thoroughly tested. Part of the current project would benefit greatly from fusion of several pertinent things which may – in due course – include caching as standard and thus obliterate a known caveat (currently, GMDS limits itself by raising an exception based on the number of vertices).

Looking at the project which deals with GMDS as a similarity measure, it very much depends on annulling the restriction that leaves us making a comparison based on meshes as coarse as seen before (triangulated faces with only a thousand of so discrete sample points).

As hard as it may be to believe (that the bug was so fundamental), it turns out that it too was a result of an incompatibility – one FMM implementation treating column vectors and another row vectors. This was the cause of a fatal error, and needless to say one leading to a buffer overflow that jeopardises other memory segments (security/stability issue) and has the whole session terminated without useful details given.

The functions' interfaces will need to be made more similar as they implement

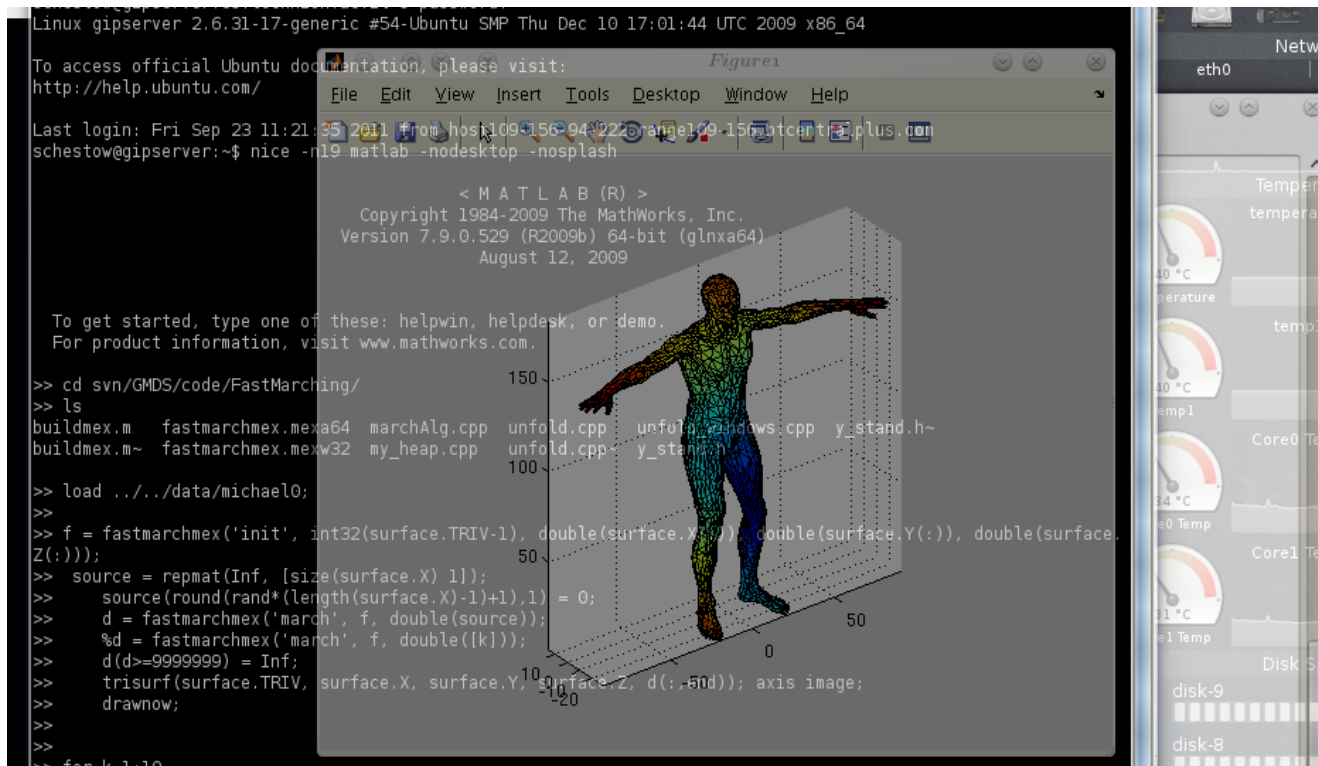


Figure 205: Newer Fast Marching algorithm as applies to TOSCO dataset

identical functionality in slightly divergent ways. The question is, however, in which way should these be standardised? Separation into initialisation and the rest seems more elegant, whereas one other implementation incorporates changes that add CUDA compatibility, among other things.

Another bug may have been spotted, which helps explain other bizarre and inconsistent (across runs) behaviour. The triangles themselves often get mangled, especially their faces. This breaks basic rules which can then cause crashes as well. See Figure 206.

It helps to be familiar with Dijkstra algorithm for shortest path computation.

It's a classic one in computer science courses (taking computer science years back to the middle of the 20th century). FMM is the same up to the update step. For most practical purposes one could ignore the unfolding part and then it's almost identical.

This is mentioned in the book "Numerical geometry of non-rigid shapes" (Ch. 4.2). The problem in this case seems to be purely technical although theoretical context helps identify where and why things go wrong. Once all the bugs are removed and both FMM versions work properly (either merged or separated by conditional statements), it ought to be possible to embed the caching. There are still fatal crashes, now occurring in the surface initialisation stage as investigations continue.

Figure 207 shows an example of how the triangles get all mangled (view from frontal observer).

In order to understand what we show in the middle (that "mangled" view) one should recognise this simply as a bug. This too turns out to have been the result of incompatible data types – a problem that was resolved after some changes to the code. Figure 208 shows the problem before the fix and Figure 209 shows the pair after the fix (the red highlight is the source point for \mathbf{D}). The next problem is less fatal in that sense that it does not crash the entire session; trying to make the code for stress minimisation (with its own surface initialiser) compatible with others may require careful merging.

TAU version has surface with properties as follows for TOSCO data (full

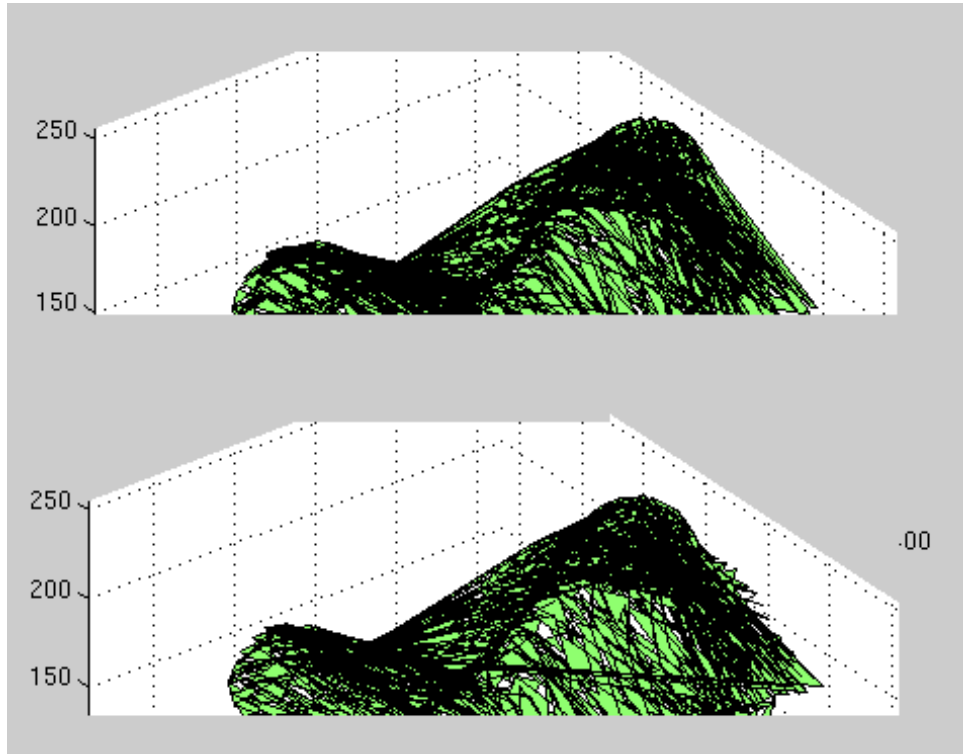


Figure 206: A bug with connected triangles

body): TRIV: [3987x3 double] X: [2000x1 double] Y: [2000x1 double] Z: [2000x1 double] TRIE: [3987x3 double] ETRI: [5988x2 double] E: [5988x2 double] VTRI: {2000x1 cell} ADJV: {2000x1 cell} D: || diam: 238.7711 nv: 2000 nt: 3987 ne: 5988 genus: 1.5000.

Consolidation is made harder by the fact that implementation-wise there are fundamental differences between the object-oriented commands-driven FMM and the other implementations which are CUDA-compatible and have no classes at all (mostly imperative), namely `unfold.c` and `unfold.cpp`.

It's really an open question; what should be taken from each to make them

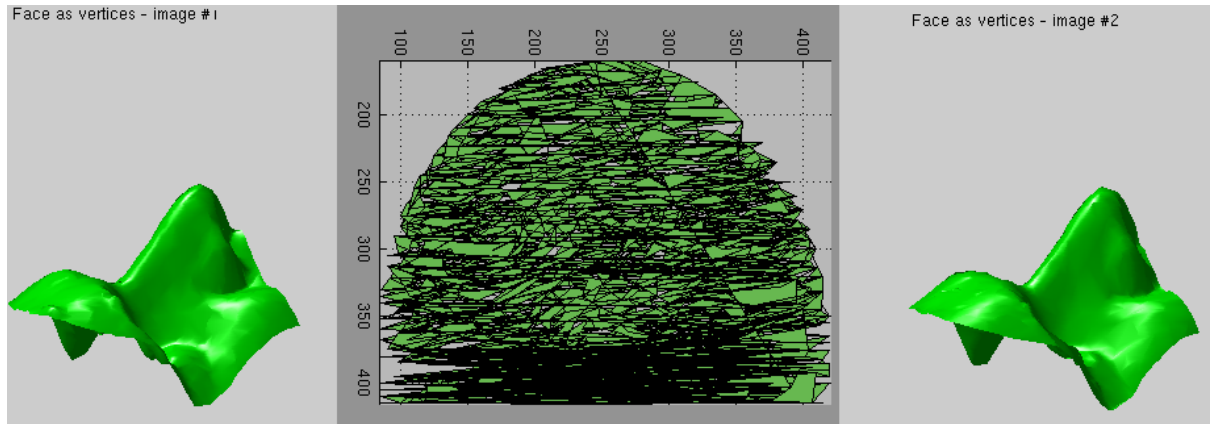


Figure 207: Two faces and the issue with triangulation

complementary? Which implementation should supersede the others? And moreover, there seems to be an issue related to the init and deinit process for the FMM class whose constructor and destructor do not always keep the data within boundaries (data is stored with a pointer to it). This seems to have been the sole cause of crashes so far. For simple cases where little memory is in use (toy examples), this always works fine, but inside a program that allocates and uses hundreds of megabytes this becomes a major peril and debugging for MEX inside MATLAB is hard.

7.12.2 Backporting

Trying to fit the new Fast Marching implementation into the framework maintained by GIP probably means coping with improvements made in TAU, some of which leading to compatibility issues that get resolved one by one *ad infinitum*. It might in fact be simpler to take the existing GIP implementation

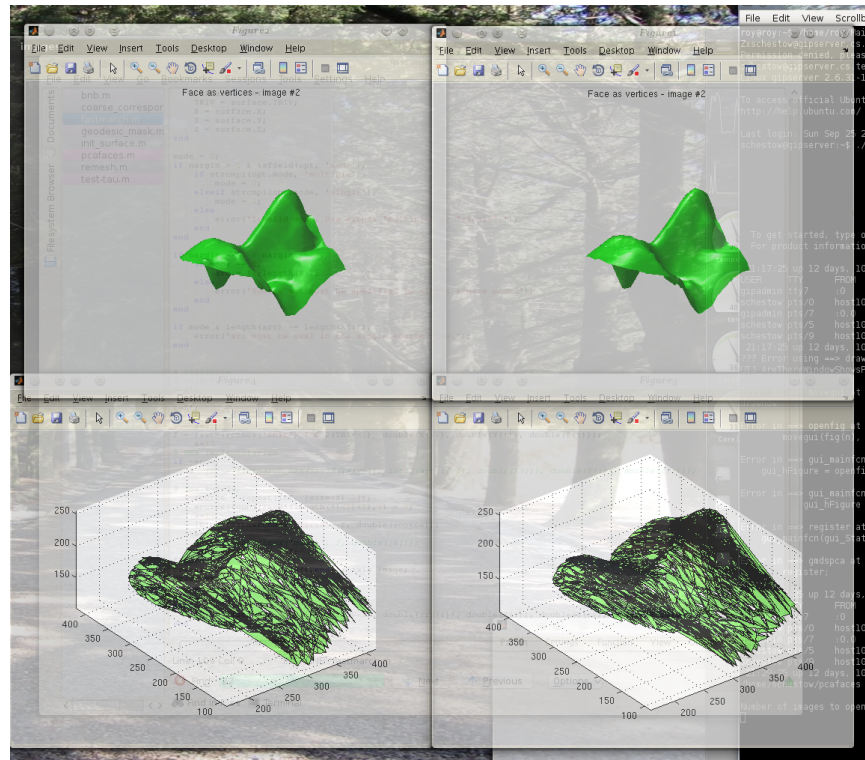


Figure 208: The data (top) and the inherent bug which incorrectly connects points (bottom)

(which works properly) and then backport everything that is of value in the TAU files, then add the cache to marching. This is not a small task, but it can at least ensure a slow and gradual deviation from what works rather than coping with many hundreds of program crashes. Nothing can mitigate the effects of future divergence unfortunately. Figure 210 helps illustrate GMDS and highlight areas of conflicts.

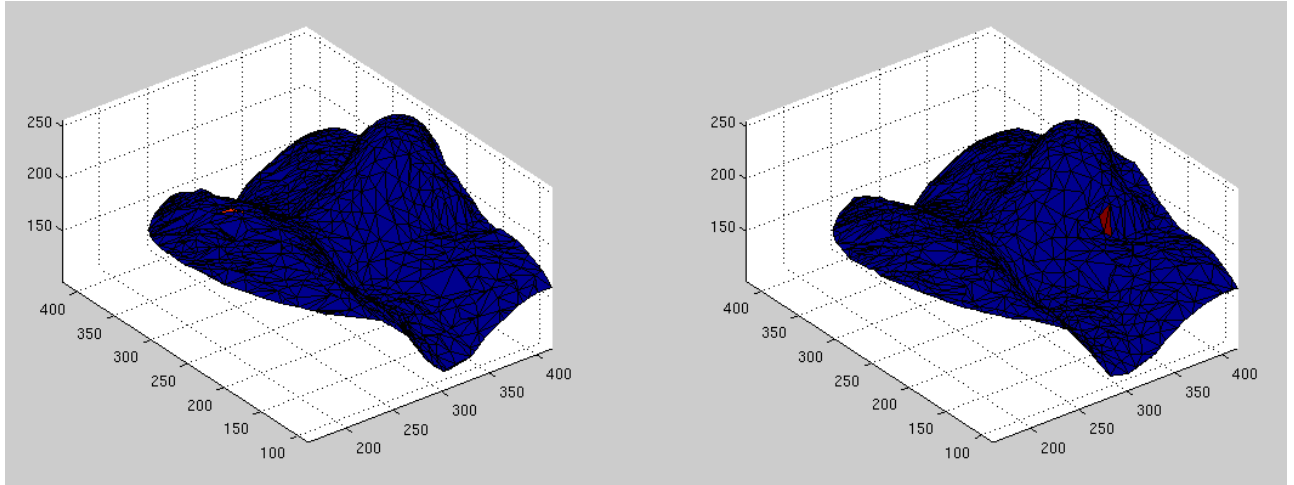


Figure 209: A correctly connected pair of faces with the source point highlighted

7.12.3 C++ FMM Debugging

We have just managed to stick together GIP code, TAU FMM, and the C++ stress minimiser, with the imminent goal of lumping the cache in to make things scale more gracefully. But there is still some bug that sometimes leaves distances between points in their initial value of infinity, which causes GMDS to cough out an error (at least not a crash). This requires some more debugging.

The causes of crashes are known now, so the task which remains to be tackled has involved modifying the GMDS code to satisfy other preconditions (for instance in farthest point sampling). Moreover, since we wish to recalculate distances the least number of times (for everything to scale properly), separation between initialisation and marching needs to be thought through,



Figure 210: GMDS using the older FMM implementation superimposed on top of new and incompatible code

with the ultimate goal of having a much faster and more scalable GMDS implementation, preferably backward-compatible with older interfaces. This has so far been a debugging problem delving deep into the guts of GMDS.

A stability problem persists in the newly-compiled version of FMM, which causes segmentation faults on seemingly random events, either at init, deinit, or march operations. When everything runs without crashes, the results are as expected. The crashes can also be reproduced using toy data, e.g. by running it many times until a segmentation fault occurs. The changes made to make the code compilable are minor and should be quite neutral.

About 5 hours were spent pinpointing the cause of the crash on GIP servers (both were attempted in case one had an unlikely malfunction). The memory allocation in Linux systems (GIP servers run Ubuntu) requires that casting gets changed in Ronnie Kimmel's adapted FMM files, which had been made more modular. The most mysterious thing about the crashes is that they are seemingly random in the sense that malloc'ing and pointing to a direct buffer would work fine for a given size/number of vertices and triangles, but on any subsequent allocations – and this cases happen at random with no obvious pattern found so far – a segmentation fault would be reported and the program suspend itself. The error occurs when accessing an FMM object (accessing a private parameter for example, through public functions, but not necessarily just that). This is being investigated with addresses being displayed as there is a strong suspicion that particular memory allocations with the old trick of casting (as unsigned integer, not `intptr_t`) throw a wobbly and make particular addresses not addressable (or occupied by another program). The whole process has required studying Ronnie's implementation, which was altered quite significantly and should probably be made compatible with the computational servers at the lab (unless another binary exists somewhere). It does generally yield the correct distances when it works, but the crashes prevent it from being testable in programs that recalculate distances (as they ought to). This generally slows down development of other parts of the programs and it needs to be resolved first, not brushed under the rug (as tempting as it may be).

We checked if the program been compiled for a GNU/Linux system before. Since the memory problem in this program is being debugged with several identical dataset in series (never to be predictably repeatable for crashes, as random parameters and new memory address are being instantiated), it might make sense to just rewrite this portion of the program, although it might inadvertently break other parts of the program, with which we need to become intimately familiar (although knowledge about it is vastly improved after many hours of debugging).

On a 64-bit Ubuntu server, creating/mallocing for an FMM sometimes has the object allocated a memory segment with address 28 bits long (0x0*****), but when it's fully 32 bits (e.g. 0xc3637a456) there is always a segfault. The problem was narrowed down to it. This seems like an architectural issue and it limits all other work. This needs to be runnable on the server.

7.12.4 Workaround implemented

To more effectively debug those pointer issues, valgrind might have to be installed on the 64-bit machine (I asked Yaron to use the admin account to achieve it yesterday afternoon). MEX has some really horrible debugging facilities and it leaves the user stranded in the land of GDB (based on the advice from MathWorks' site) or just trial-and-error. Debugging segmentation faults is hard without advanced debuggers or valgrind. Meanwhile, new workaround code was written to reject improper allocations of memory (ini-

talisation of FMM leaving pointers that are not accessible to the marching operations). The original code seems to have been written for 32-bit Windows, whereas the servers use swap and highmem/Physical Address Extension (PAE), so we have added a constraint for it never to use the swap rather than make a more permanent fix. This at least should facilitate crashless experiments and development.

7.12.5 Resolution Increases

Fusion of code portions gave more scalable code that uses a single source for FMM and enables one to see how GMDS fares as a face similarity measure, even if resolution is improved beyond the 600 or 1000 vertices, as shown in Figure 211. It is important to emphasise that all GMDS results we have ever gotten (and associated ROC curves) are based on just 600 or 1000 vertices. It should later on be possible to see the potential of increases, with or without cache (there are other tricks of the trade available for use).

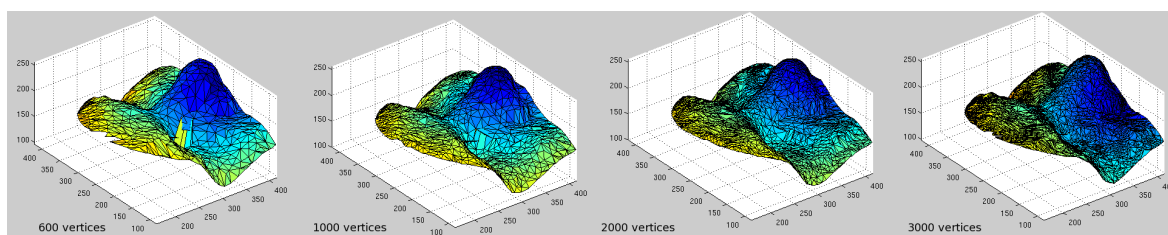


Figure 211: Visualisation of the increasing number of vertices

We have been pushing it hard to get high recognition rates (with increased resolution, refined boundaries, etc.) since the end of July to make it more

commercially viable and competitive wrt other algorithms (and thus publishable) and if these resolution gains do not help, there might be other things that can be more easily explored (without having to debug much).

Sticking with the objectives set in September, I shall try to produce a comparative analysis, e.g. using overlaid ROC curves, that gives us insight into impact on resolution on recognition performance. Curves should hopefully show that the higher the resolution, the higher the curve (nearer to the top-left corner), in which case it becomes clear that the problem is truly resolution-dependent, to an extent unknown to us at this stage.

Working our way upwards, so far it is simple to demonstrate the correlation between the number of vertices that make up the surface and the recognition performance, as shown in Figure 212's interim (coarse) plot that will have more samples added to it. So far, based on a very crude GMDS algorithm with just 50 points and between 100-400+ vertices, it can be seen that recognition performance is strongly linked to the amount of information available (no surprise there, but a sanity check at the very least). It will be interesting to find out at what point there is convergence/plateauing, i.e. at what stage it no longer helps to have resolution refined. A lot of the older experiments were run with 600-1000 vertices and those which were better optimised could peak at around 95% recognition rate.

I've reached 4,600 vertices. It's plateauing somewhat at this level, however with code from Alex come some nice tricks that will shortly be incorporated

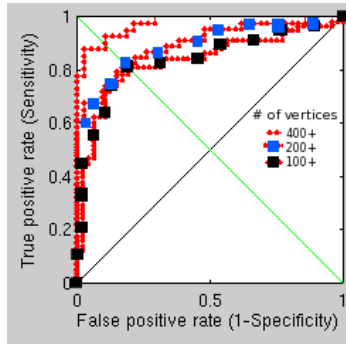


Figure 212: Coarse resolution performance compared

to improve performance. Some values are clearly indicating match failure at GMDS level, which merits another attempt at matching (exception) rather than a final classification.

These experiments took a long time to set up and run (manually) because of some freezing and stability issues. The important thing is that they provide insight into potential of particular paths of exploration. It is quite imperative, especially for 99% recognition rate (or anywhere in that region), to eliminate all situations of topological mismatch. If those cases are removed, then exceeding 95% should be easy.

Stability is a bit of a nightmare at a resolution which translates to 4,600 vertices, unless more tricks like caching are put to use. But all in all it is still possible to pull some values out and based on 10 true pairs and 10 false pairs (randomly selected), there is perfect detection rates (although with a small sample size). What would be worth implementing is a mechanism for detecting mismatch that it topological and then retrying with different ini-

tialisation, until the score falls within some sane range, as defined in advance. This would considerably improve performance.

The combination of C++ code for stress minimisation and some improved code (which does not run too robustly on the 64-bit servers yet) has the potential for acting as a similarity measure, with or without a training process (e.g. PCA). Performance in terms of time is not fantastic, but it is a tradeoff between accuracy and speed. The increase in the number of vertices does seem to play a real role.

Many incorrect classifications are the incorrect matching between correct pairs, due to GMDS errors. This leads to a high false negative rate (type II error), which can mostly be overcome with repetition using permutation, assuming the detection exceeds some certain threshold, which is something we were implementing. The ROC curves in Figure 213 do not yet use this methodology.

Interesting solution was seen developing, with hope of seeing if it works. Regarding the ROC, we have flip in performances as we increase the resolution. This is a baffling reversal, but given the size of the test set (due to crashes these experiments required literally hundreds of sessions) it's more or less expected to be around the same ballpark at high resolutions. Heightening the curve (vertically) is something which should be easily achievable and is worked on at this moment. There are problematic cases – however rare – where the same person in different poses is recognised as not belonging to

the same set of images.

It looks as if 3000 is worse than 800 vertices, which we do not get in 4600 vertices. 4600 vertices without some extra caching already pushes the program to its limit, which reduces the size of the test set, but all recognitions attempted so far provide perfect separation. It would be ideal to invest time in it once the algorithm has been amended to not account for GMDS errors as image mismatches (false negatives).

We could align coarse to fine if this is the case. This hasn't been tried here before, not in the experiments shown so far, Maybe by doing a multi-resolution test we can get more reliable similarity ensemble.

Areas of mismatch have been studied more closely in order to understand what causes them. Several large images were looked at along with the GUI (with previews of images enabled), showing quite clearly that we must remove hair from the surfaces as many mis-detections are caused by this. The hair just slips in sometimes, depending on the imaged person. We must also fix image alignment in cases where the nose tip is not accurately detected and use multiple runs, perhaps with a multi-resolution approach as an option, for ensuring correspondences have had more than one opportunity getting identified. The images from here onwards provide insight into the 'debugging' (adjustment) process, which currently utilises special cases (errors) to find essential workarounds. The aim is to get close to 99% recognition rate.

By cutting the surfaces above the eyes (still tweaking the levels), increasing

the resolution, and adding a multi-resolution approach among a few other improvements, the discriminative power is now greatly increased and the 'sweet spot' seems to be approachable in the sense that the erroneous classifications get looked at closely, whereupon it usually turns out the the metadata – not the data – has mistakes in it (incorrect pairs marked as correct ones and vice versa). The borderline cases are those which require tweaking for.

Figure 220 shows a densely-sampled surface without the forehead (the area above the eyebrows has components removed because of issues associated with hair).

The sample size is not large enough for an sufficiently informative ROC curve, but there are only a few wrong classifications. One is a borderline case where pairs from different people almost seem like belonging to the same person (just almost, so separability can be further improved). The other case is mostly a case of GMDS not working, not quite a wrong classification. At all resolutions attempted so far, one pair of faces (same person imaged) cannot be made correspondent. This gets detected as an error because the values are not sane. Other than that, there is almost an order of magnitude apart in terms of separation between correct pairs and incorrect pairs. One important issue to tackle is the rare case where GMDS hardly latches onto facial features at all, as shown in Figure 221 and Figure 222.

7.12.6 Smoothing

Following some further investigation it seemed reasonable to try smoothing of areas like the eyes, where local inconsistencies got GMDS preoccupied. So far the results suggest perfect separation (set size is about 30 and growing).

A problematic borderline case is shown in Figure 223 and also the sorts of surfaces (in 3-D, see Figure 224) where GMDS oddly enough failed, despite trying different resolutions and random seeds.

A GMDS-based face recognition task, with smoothed surfaces where the resolution is increased for accuracy and for improved performance, still works rather well (room remains for improvement). In the following experiment only one image was problematic, only slightly bordering the threshold because of pose variation on the face of it (still needs further investigation, see Figure 226). There was only one case where GMDS failed and the reason is yet unknown (Figure 225). The ROC curve is in Figure 227.

A kernel/window 13 pixels across, moving average (horizontal and vertical). The 2-D Gaussian filter is another option, but although it's in the program, it is not in the GUI yet, so this has not been attempted (there is a lot more that can be tested). Back in July-August it could be demonstrated that by smoothing the surfaces, errors could be reduced somewhat.

7.12.7 Resolution increased

With smoothing made better (covers a wider area in true 2-D) and the resolution increased somewhat, results so far show the threshold just approached by two images. It is premature to draw conclusion and too early to suggest that separability has been degraded/improved, but the good news is that GMDS has not failed in a major way, not for correct pairs anyway. This whole process could still use some tweaking and there are several ways remaining for improvement. Figure 228 shows this graphically.

The results based on first few iterations show just one minor error, as show in Figure 229

Basing the next experiment on the situations where detection merit/recognition value is on the margin (around 3 in this case), here are the 3 problematic image pairs (see Figure 230), which seem to suggest either a weakness in GMDS as a similarity measure or a gap in implementation. The good news is, by upping the resolution, cases of incompatible topology have been eliminated, at least for the test set in this case (about 30 pairs of images).

The false positive seems strange. GMDS is obviously not optimal, so we need to understand the sources of the these problems and try to program them out.

Interestingly enough, all the numbers are repeatable and reproducible despite the stochastic element and contrary to prior cases (coarse resolution). They stay the same across run, with a 3-decimal-point accuracy. In Figure 231 are

the three pairs shown previously, in context.

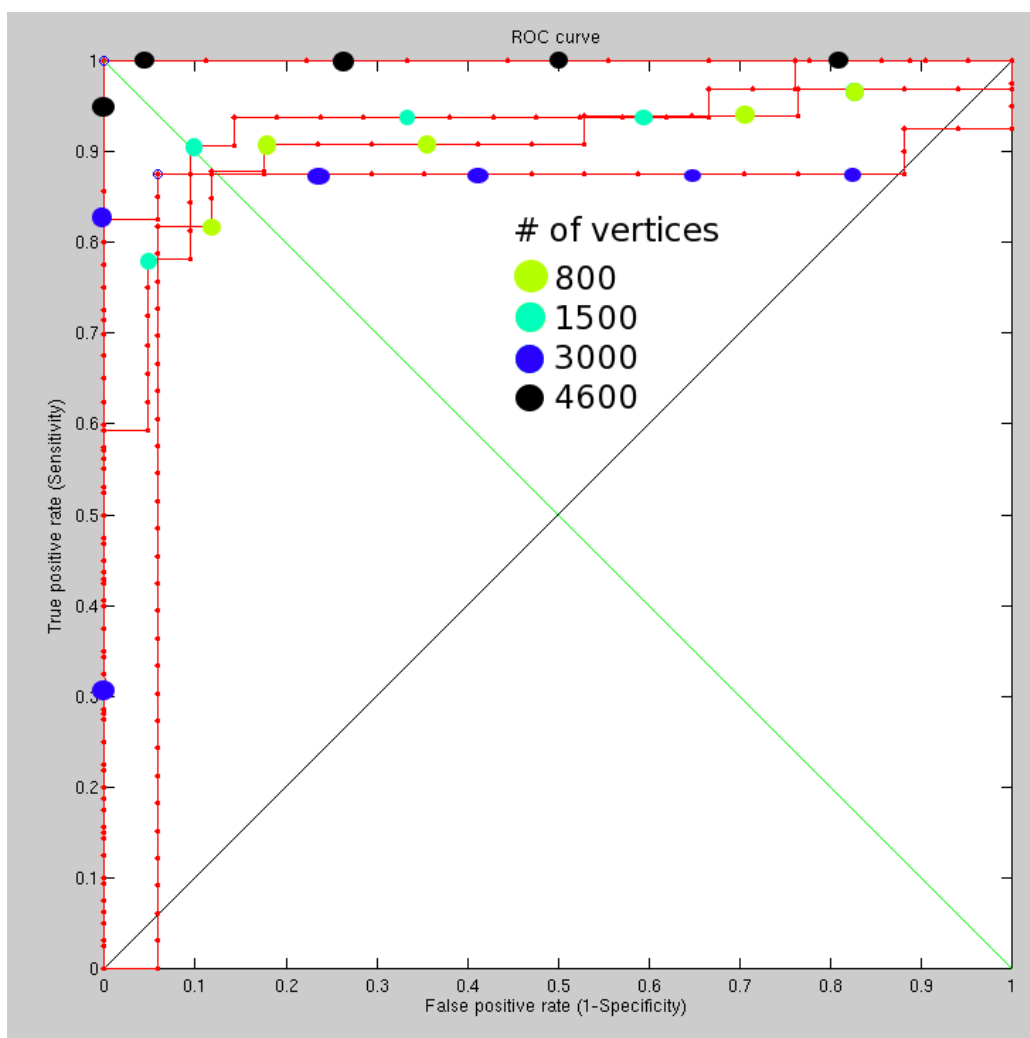


Figure 213: A finer resolution-oriented set of results obtained from fewer runs than before

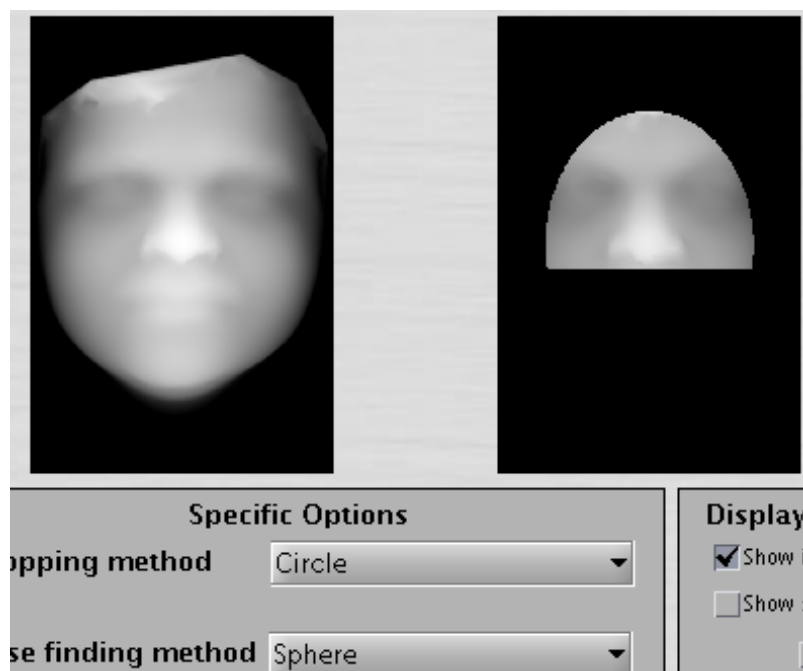


Figure 214: An example where the hair entering the surfaces can interfere with GMDS-based recognition (GMDS as a similarity measure)

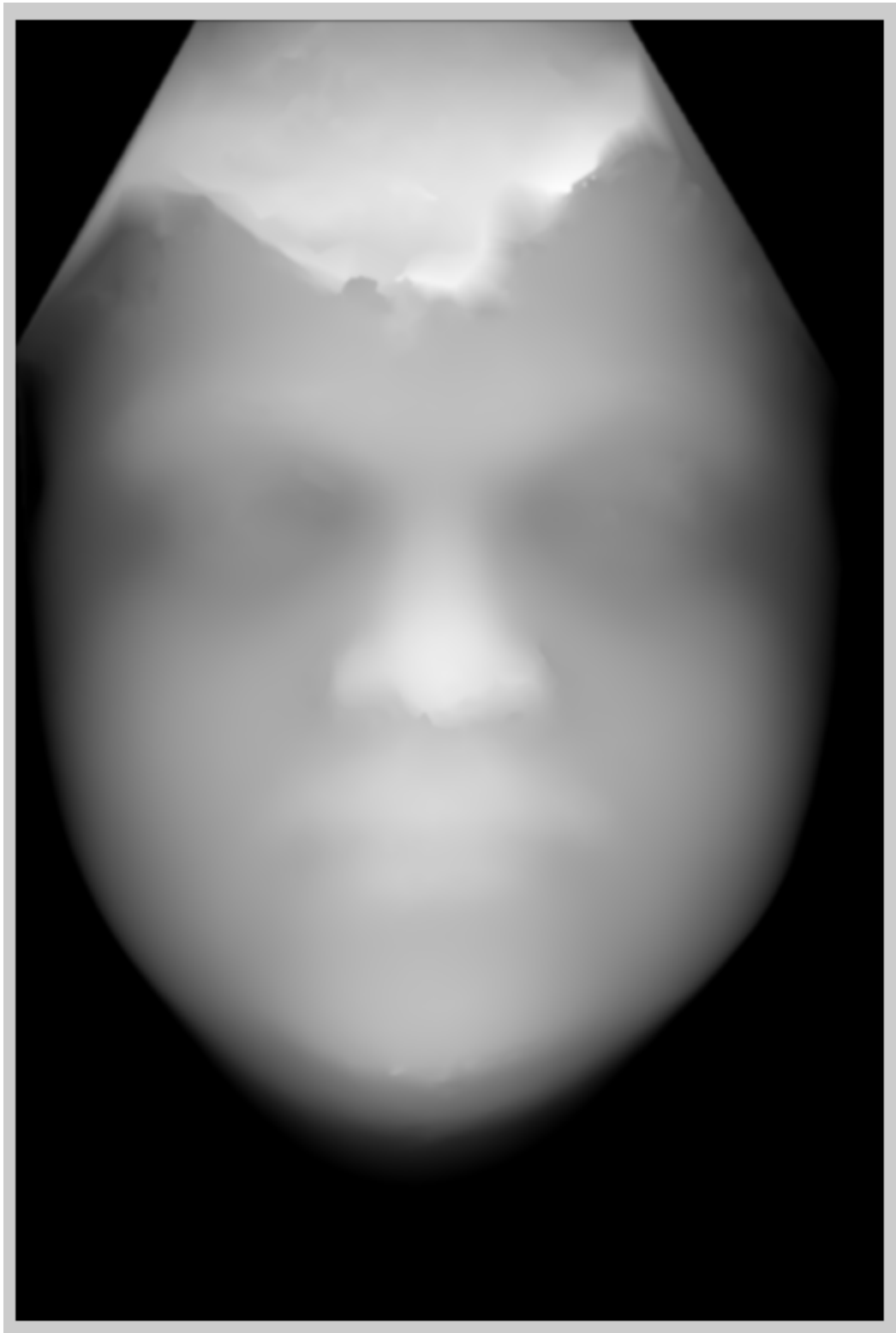


Figure 215: Example where hair is at risk as being treated like skin surface, depending on the mask/s

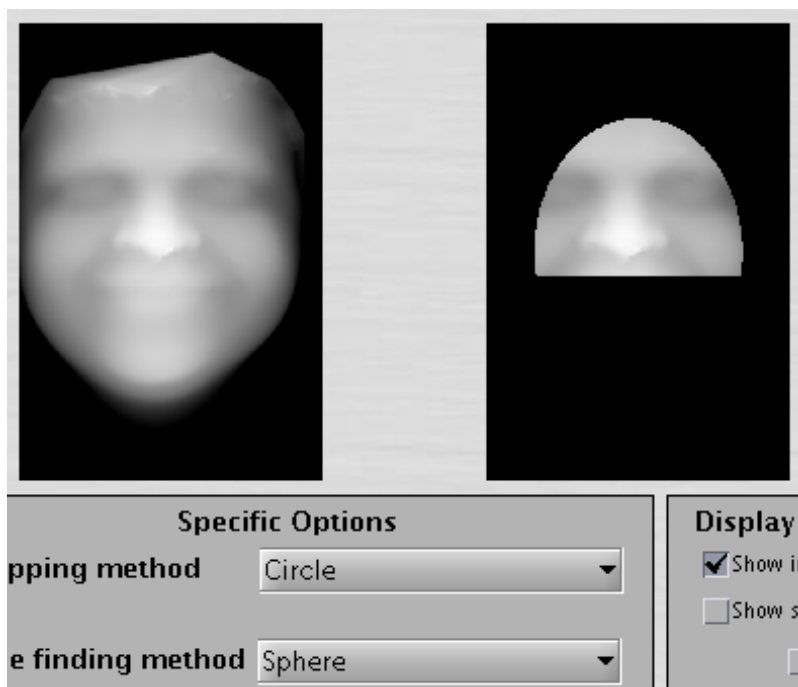


Figure 216: The result of the nose tip being misplaced (original on the left, after masking on the right)

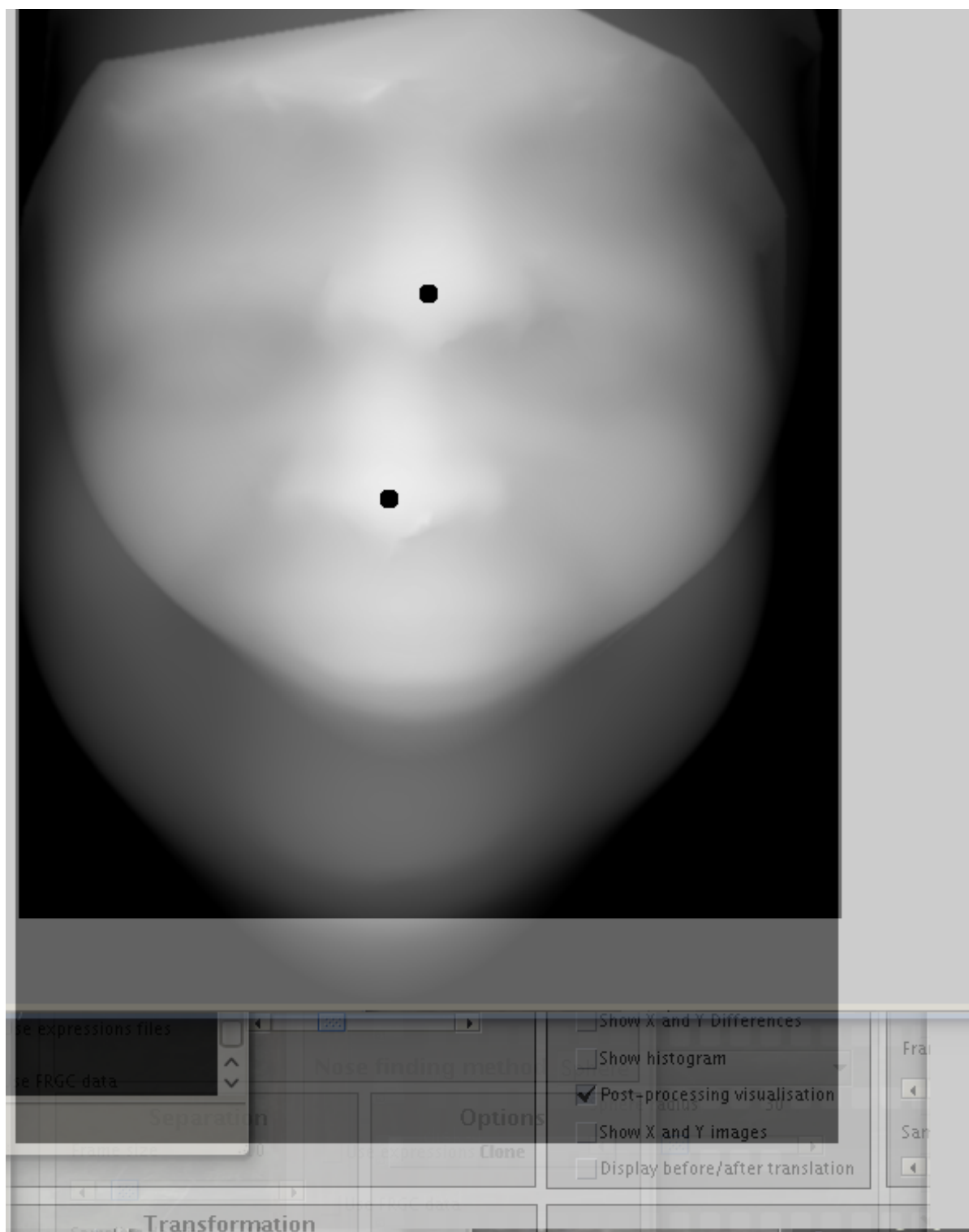


Figure 217: The problem of non-overlapping faces, a result of misregistration/misalignment

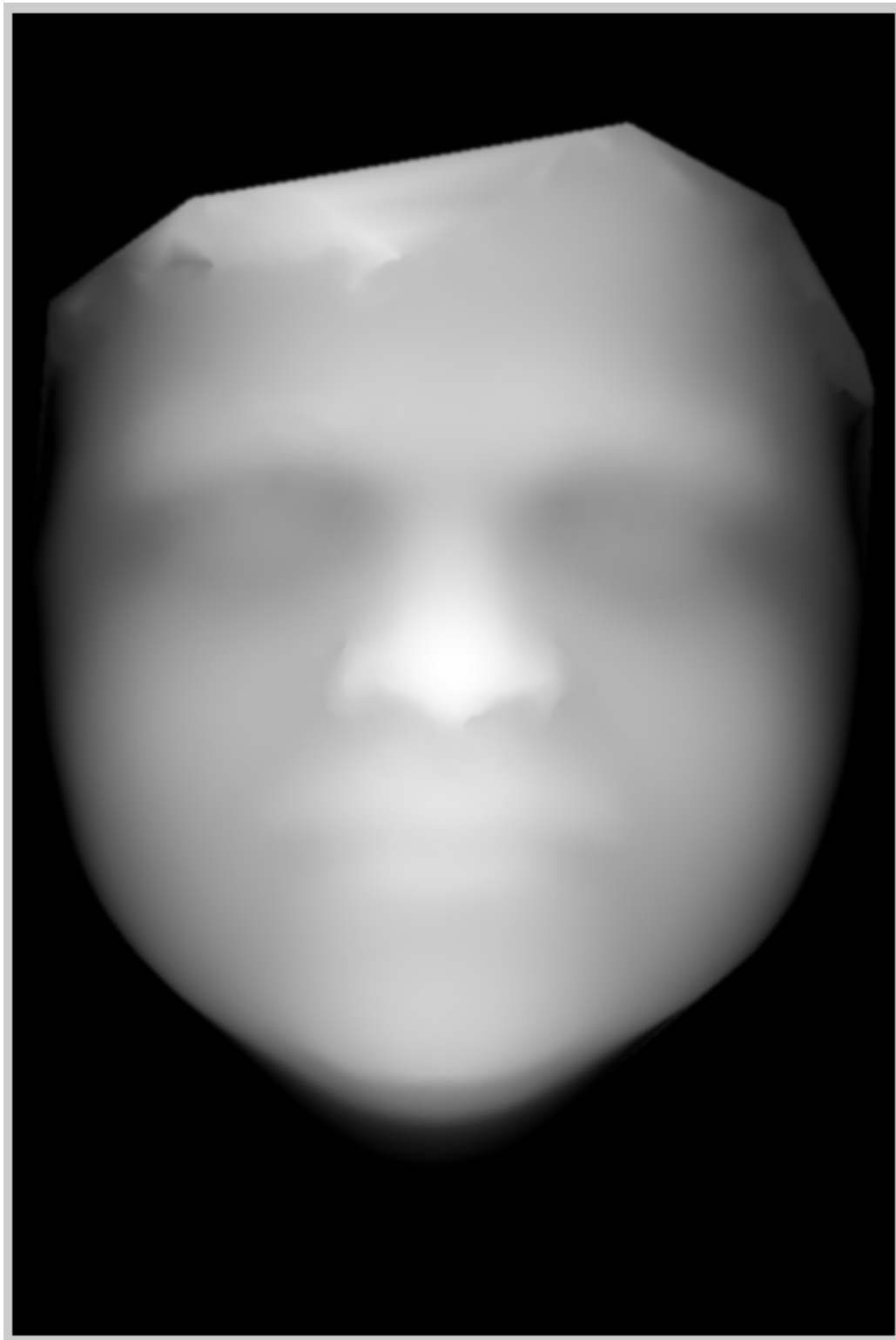


Figure 218: Registered and correctly aligned image

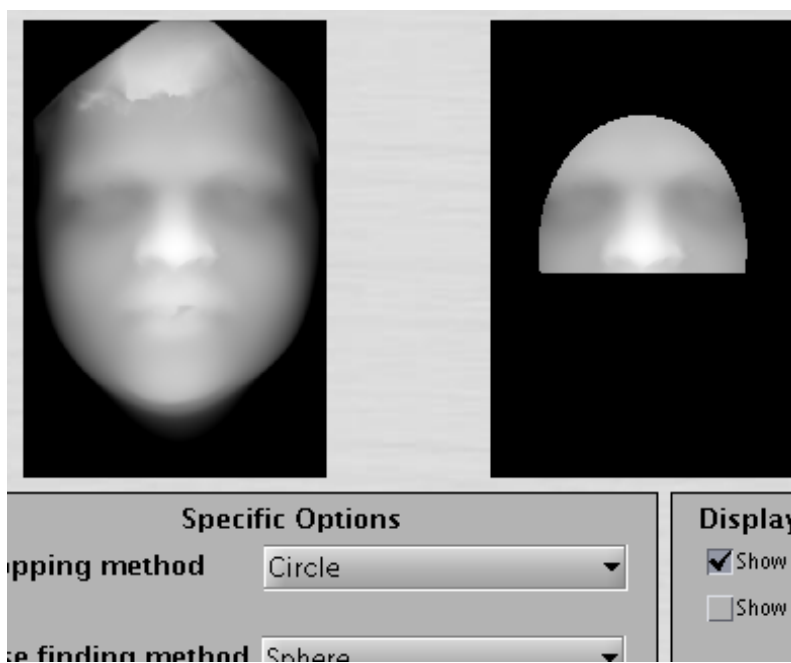


Figure 219: Example of a correctly sliced image subset (before geodesic boundaries cutoff)

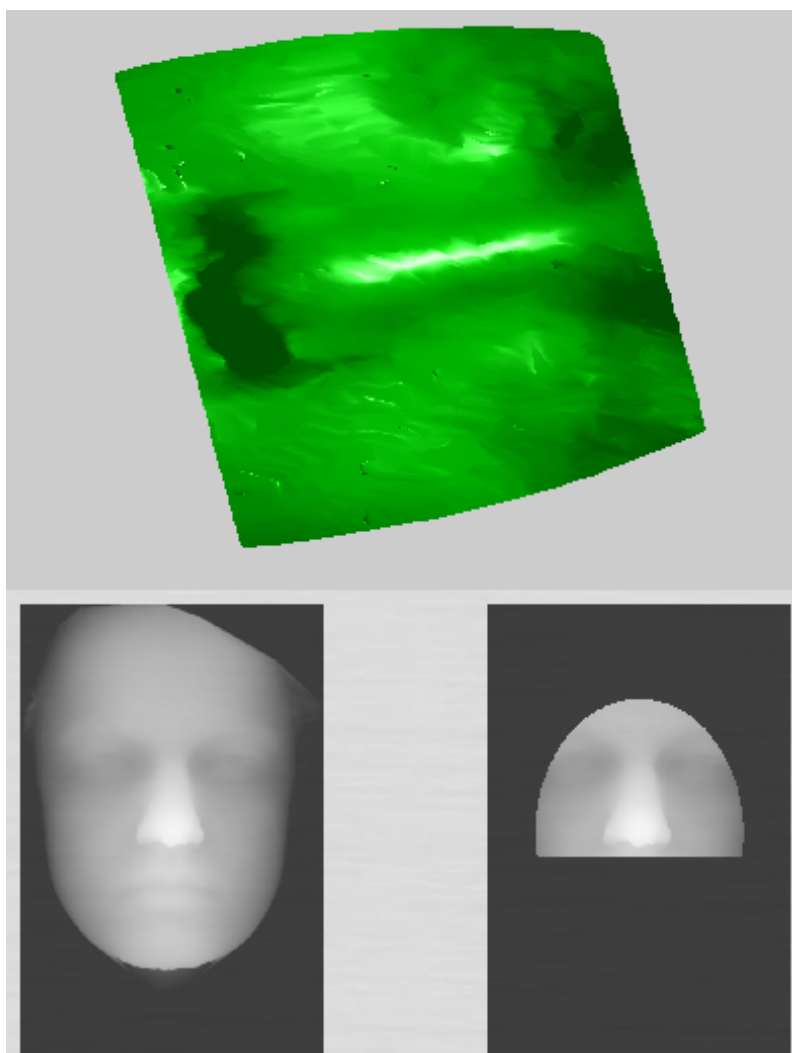


Figure 220: High-density (vertices) surface and the images it is carved off

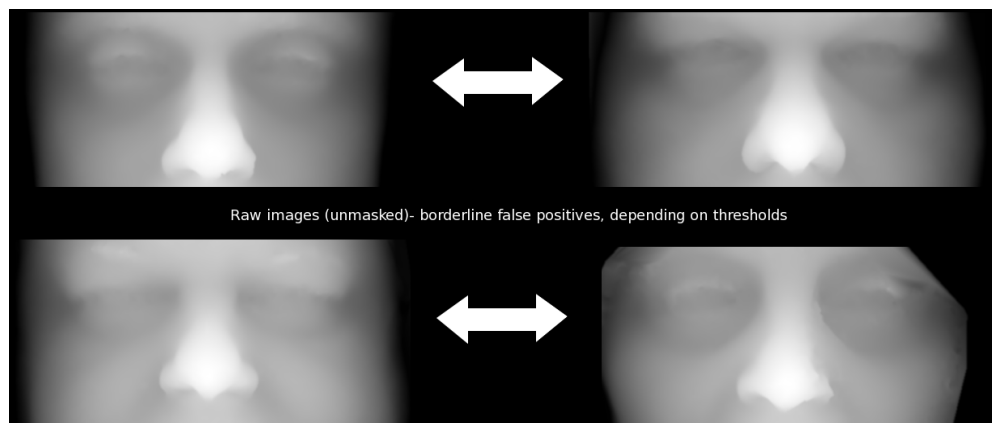


Figure 221: Problematic image pairs

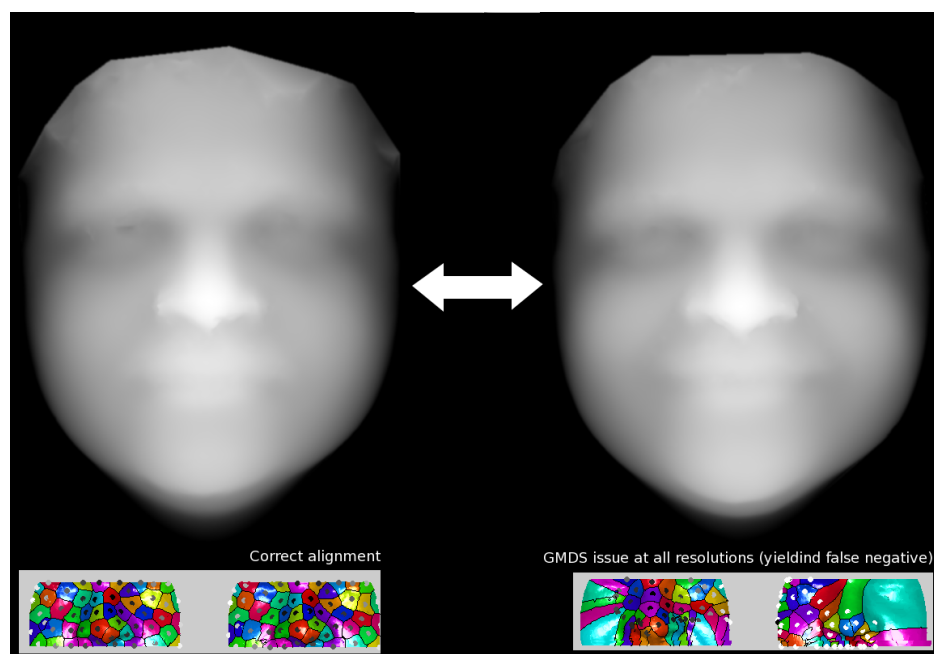


Figure 222: A pair that causes GMDS to fail

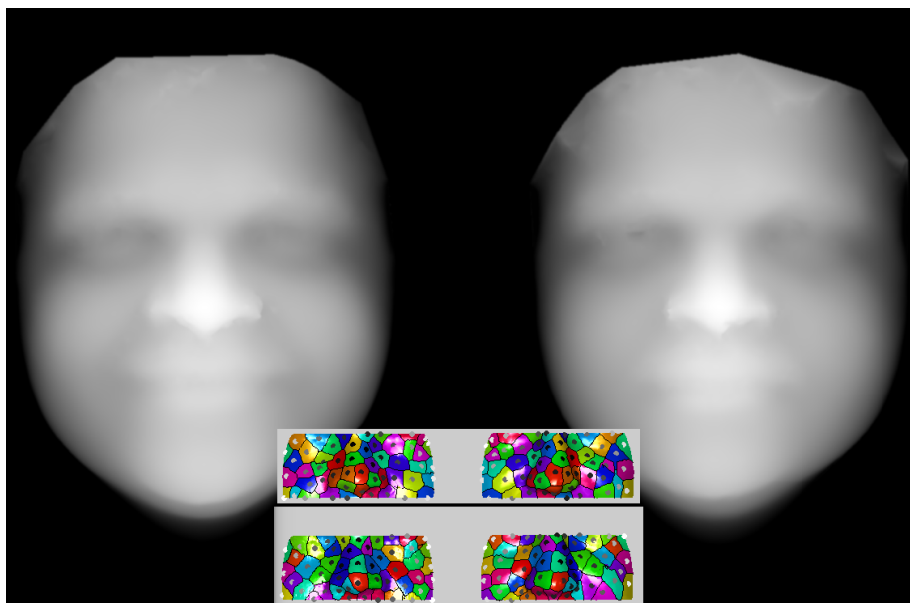


Figure 223: Problematic real pair (same person) where GMDS works but poorly so

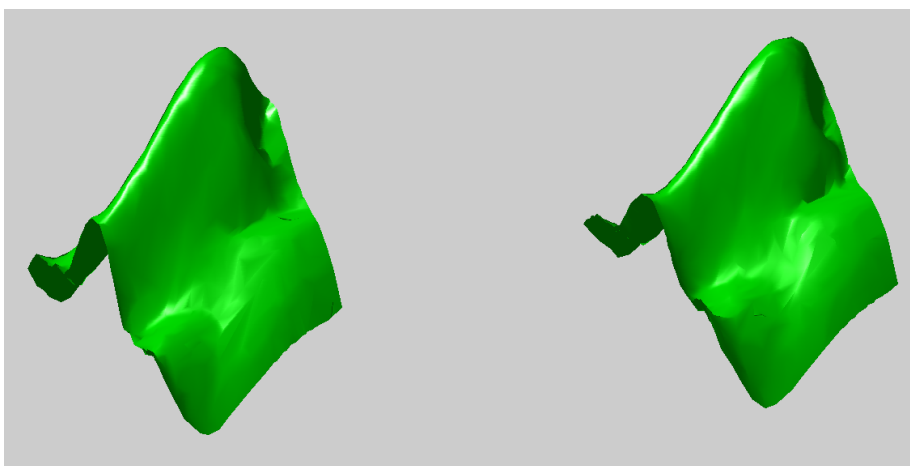


Figure 224: A 3-D representation of a pair of images from the same person

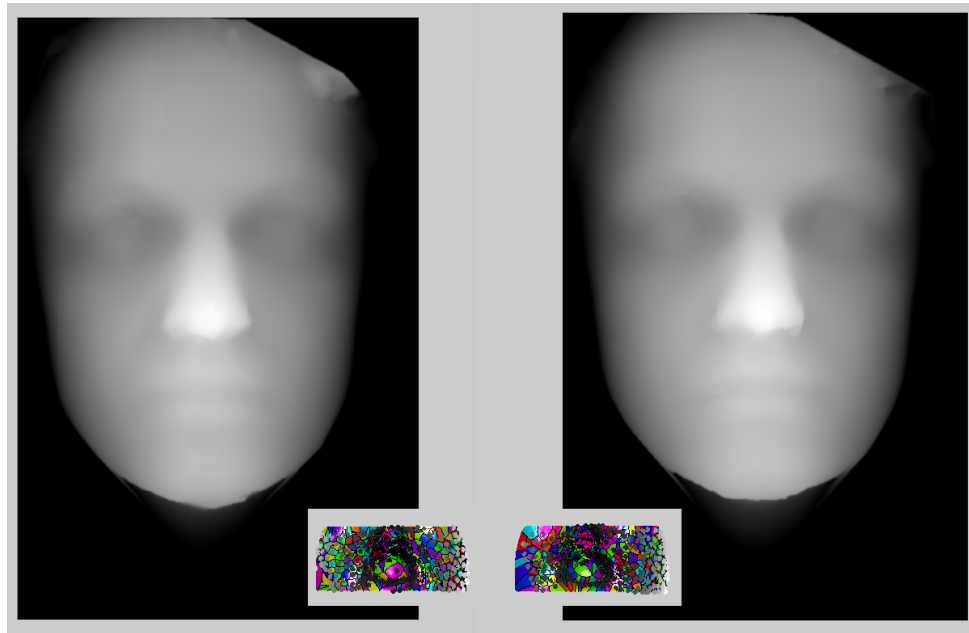


Figure 225: GMDS failing to work as expected



Figure 226: A problematic pair which is seen as too different to quality as a match

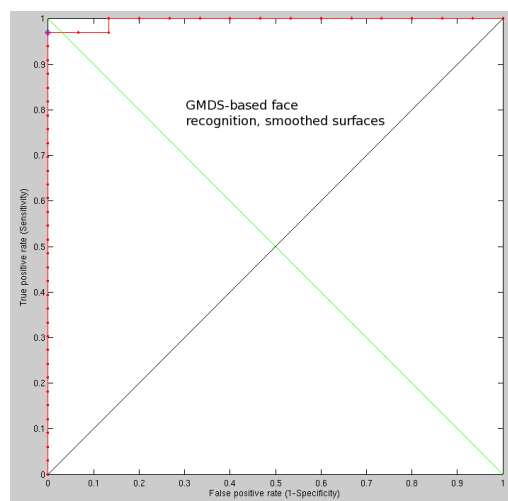


Figure 227: ROC curve based on the smoothed surfaces variant of the algorithm

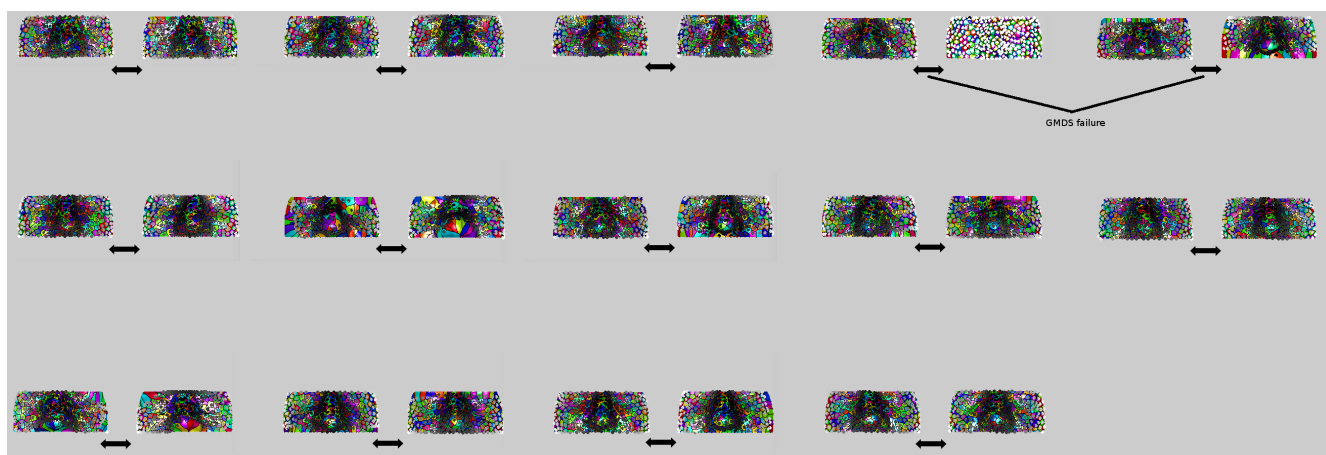


Figure 228: Example of some GMDS (mis)matches in the initial experiments

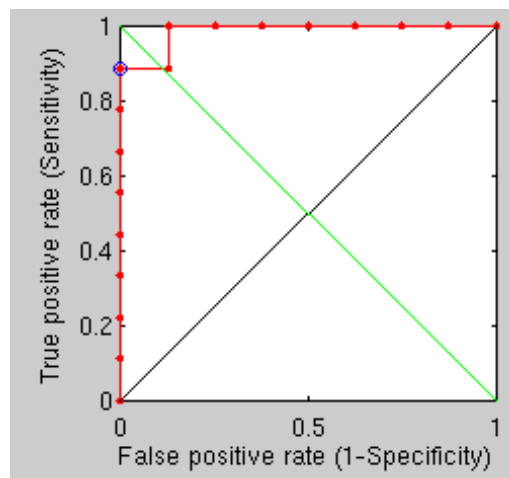


Figure 229: ROC curve based on the improved smoothed surfaces and somewhat better resolution

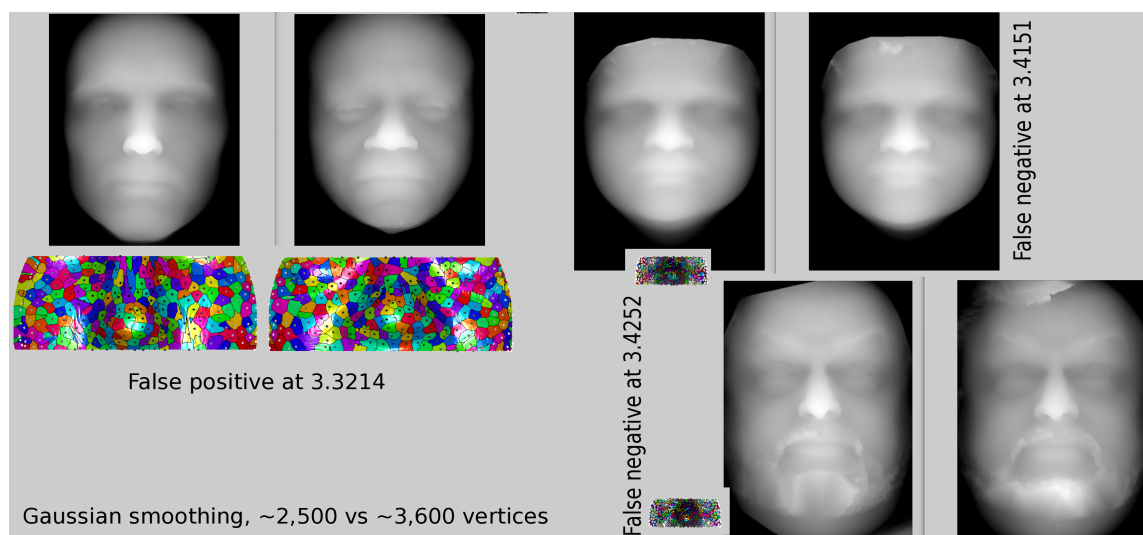


Figure 230: 3 problematic image pairs

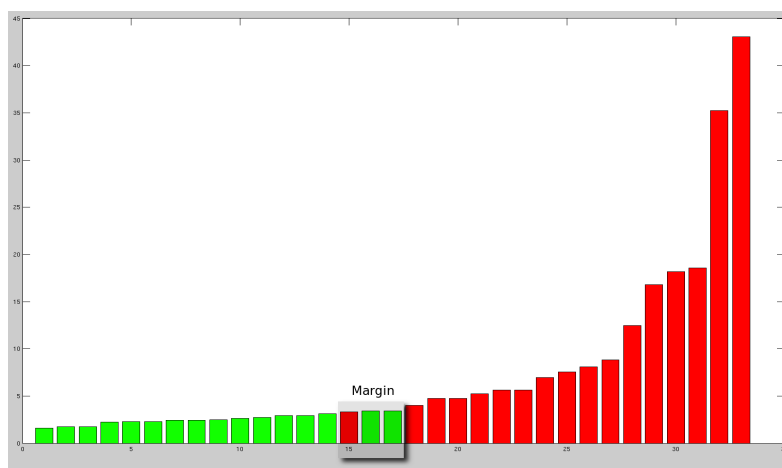


Figure 231: The area of collision in GMDS-based face detection

It is hard to understand then how the false positive came to be. The geometries look substantially different. Dealing with this problematic image in isolation, the relative values of GMDS (relative to other of its kind) were studied to better understand when and why they diverge to the point where GMDS matches quite well faces that are from different people (and inherently dissimilar). At 3600 vertices it enters the margin like no other pair does. At 2500 this problem persists and at 1000 this problem is gone (the score is well within the true negatives territory). But at 600 it becomes a tad problematic again.

With smoothing and other parameters kept invariable, it should be possible to run these experiments at multiple scales and then learn by observation how to later conduct an uncontrolled, unsupervised and blind experiment where more than just one single GMDS operation is used to assess similarity. An-

other possibility would be to 'hybridise' methods, e.g. use a simpler method of comparing pairs in conjunction with GMDS, or as a sort of regularisation term in a more compound objective function (maybe normalised as well).

7.12.8 Residuals

On their own, simple surface-to-surface metrics seem to be weak as classifiers, but in cases of GMDS similarity, values falling around the margins (i.e. close to threshold of ambiguity) can be made more reliable by enhancing and increasing the amount of data. Current work refines methods of detecting and modifying GMDS/stress scores that are low despite inherent differences that might be non-isometric. This essentially combines geodesic metrics on the surface with Euclidean ones, ruling out what would otherwise be false positives.

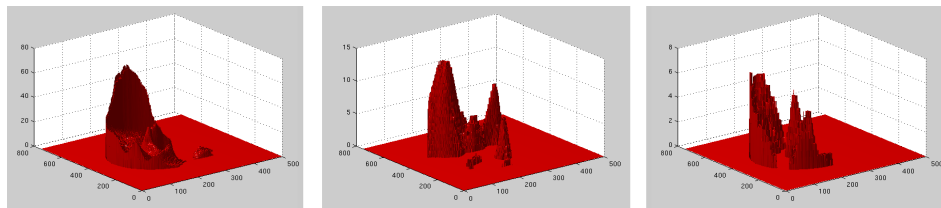


Figure 232: Examples of shape pair residuals and the corresponding ROC curve

The previous results demonstrated the great weakness of purely Euclidean measures that use the residual, where every small bit of misalignment almost dominates the difference. The challenge has since then been to identify a Euclidean distances-based measure which is robust to this type of variation

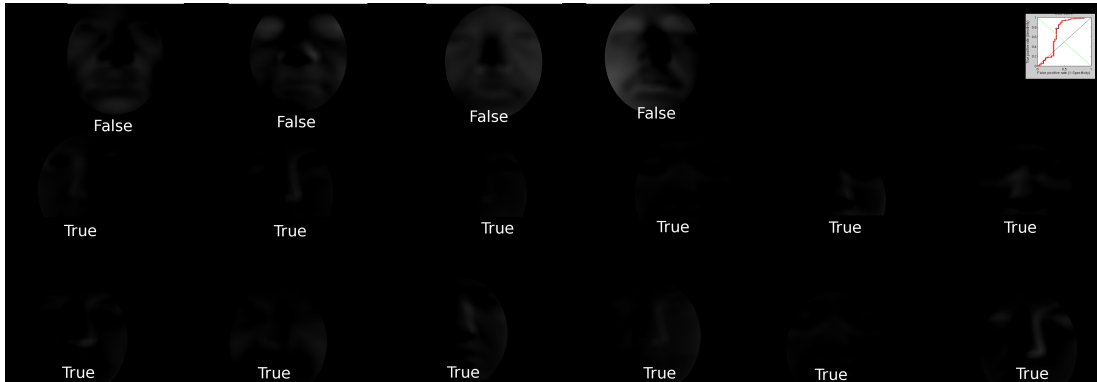


Figure 233: Residual difference and the problem of localised high signal (which makes this a weak similarity measure)

and then complement the purely geodesic distances-based measure (notably GMDS). In this first batch of experiments, a volumetric-type Euclidean distance (gap between the surfaces put on top of each other) gets measured. In order to demonstrate the great variation, even within pairs of the same individual imaged, a figure was produced (see Figure 235, showing areas of very high contrast, e.g. at the sides of faces. The ROC curve in Figure 236 shows the problem. By aligning around the nose and then considering just the nose area we can possibly get better results (although still rather poor, as shown in Figure 237 and Figure 238) that are based on Euclidean properties. Another Euclidean-based measure worth exploring might be distances between particular points of interest, e.g. eye corners and nose tip. The goal is to eliminate cases where two images are identified as belonging to the same person based on geodesic properties alone, even though based on other criteria this is clearly not always the reliable thing to do.

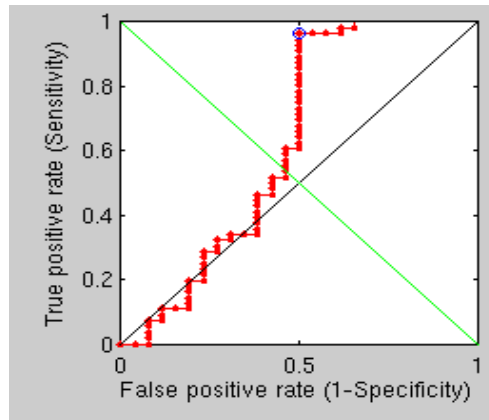


Figure 234: ROC curve obtained by using a residuals of just a particular image region (nose and eyes)

To make it more robust to movement around the nose tip, the surfaces are shifted a controlled amount in X and Y in search of an optimal match [238](#).

A good couple of matches are shown in [239](#).

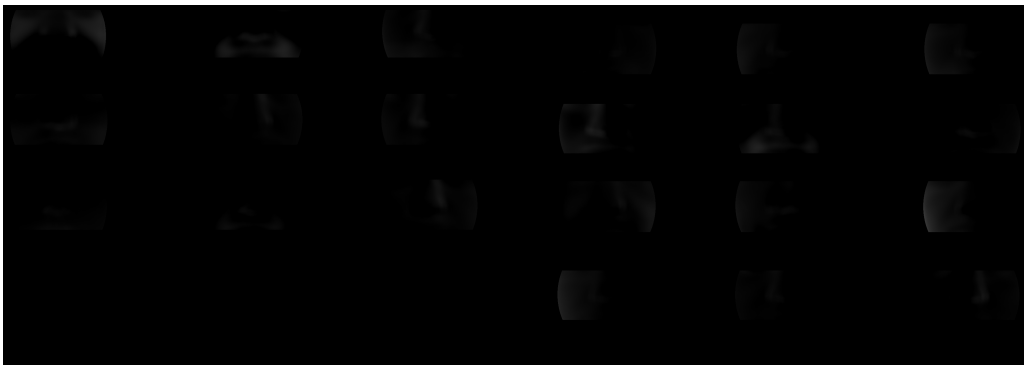


Figure 235: Examples of pixel differences for pairs of the same people

For recognition based on surface sum-of-squared-differences, the best achieved recognition rate is currently around 80%, which gives it vastly inferior discriminative power compared to GMDS (as expected). In order to make a

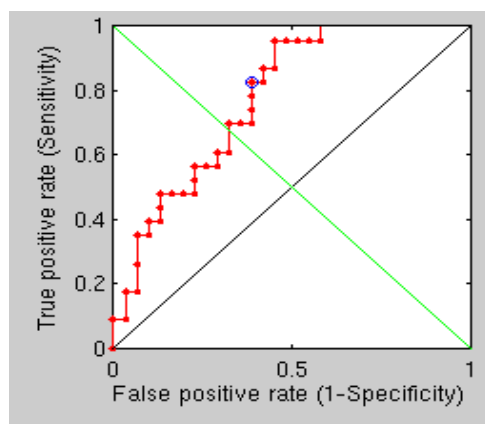


Figure 236: ROC curve corresponding to pixel differences for the whole middle section of the face

fusion of these two, e.g. using the weaker one as a mere regulariser, careful thought is needed because one can degrade from the usefulness of the other. One idea which was tested earlier is the invocation of a more complex classifier only in cases where classification is on the margin, i.e. GMDS is unable to comfortably discern real pairs from false ones. For the small test set used so far this can yield perfect recognition, but it requires further testing to be generalisable.

By applying a similarity test that falls back onto Euclidean measures when GMDS is unable to make a clear distinction (score between 3 and 4), the algorithm is now able to classify all image pairs (72 images in total) correctly. Increasing the number of those pairs might present new issues and, shall any such issues arise, we can design a workaround. To claim 100% recognition based on just 72 images does not make sense, so I will increase the number of images.

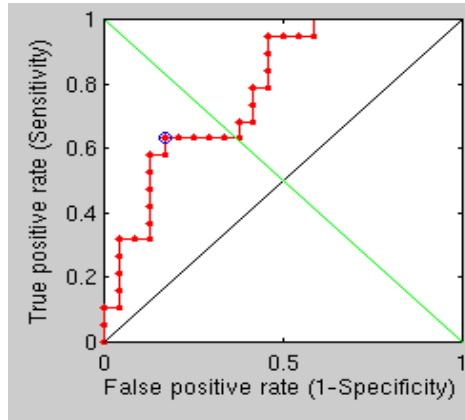


Figure 237: ROC curve corresponding to pixel differences for the nose area alone

7.12.9 Higher Resolution

We have increased the resolution (for GMDS) a little further and based on comparisons involving 44 images (no fallbacks for cases of ambiguity/uncertainty yet), there is one mistake and an unusual number of 3 cases where GMDS does not identify the structure correctly despite the high resolution (which is OK because it is not a false classification).

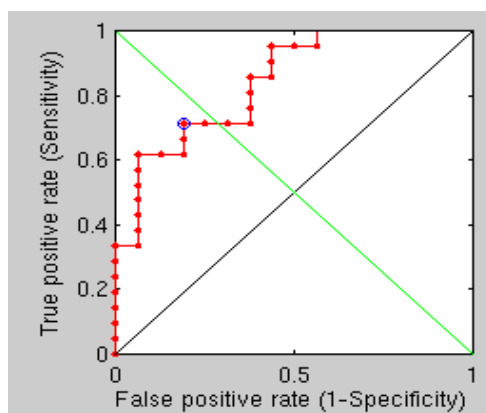


Figure 238: ROC curve corresponding to sum of squared differences for the nose area alone

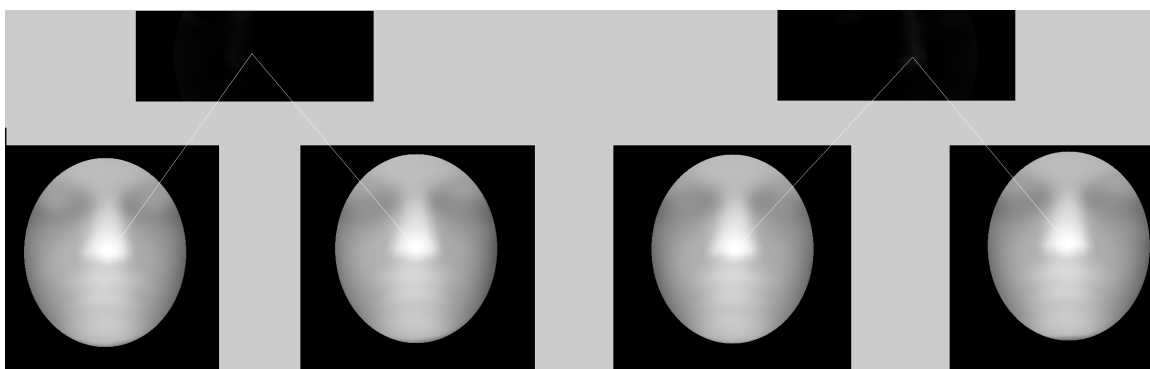


Figure 239: Example of 2 pairs from which the difference image is produced (shown at the top)

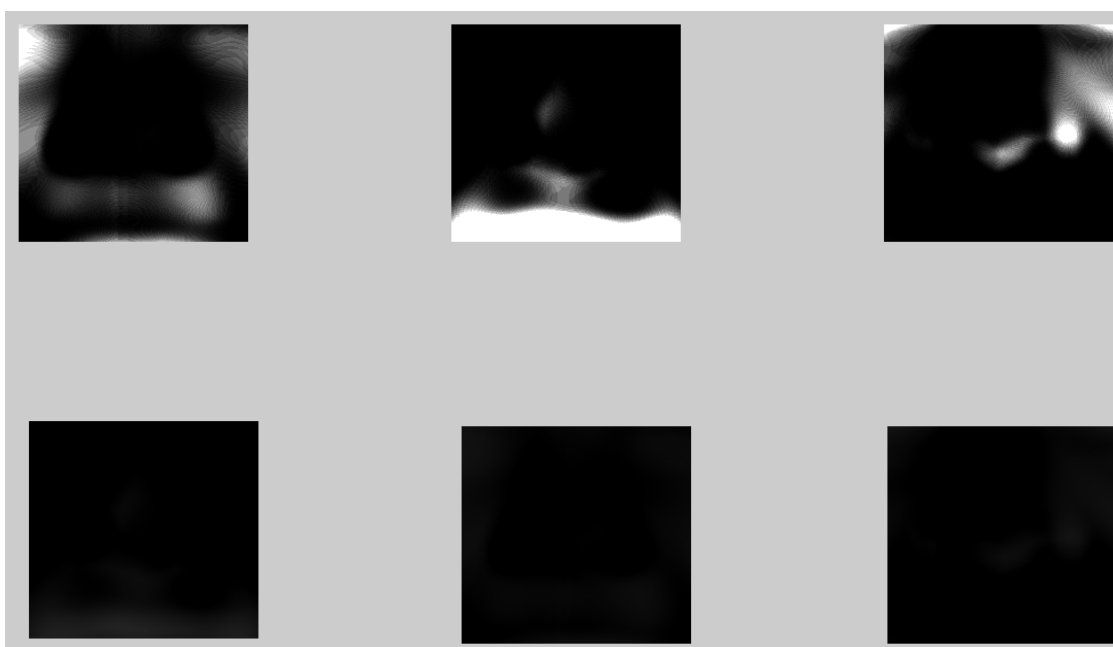


Figure 240: Top images show the sum-of-squared-differences of the first 3 true pairs, with the mere difference shown at the bottom

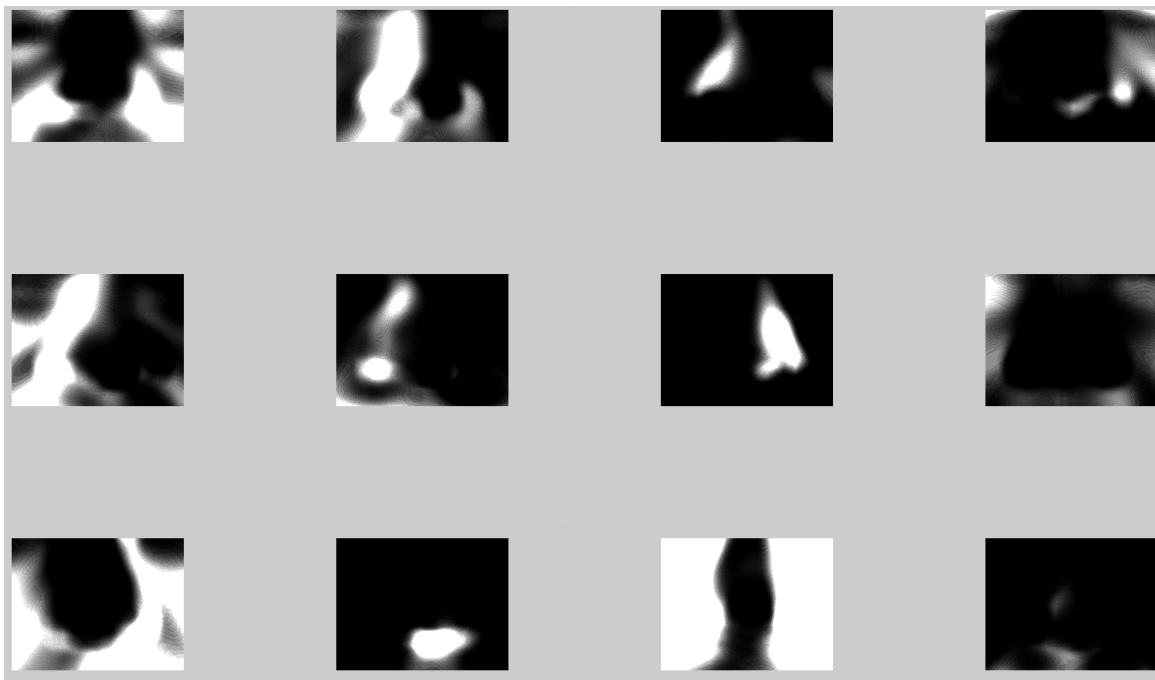


Figure 241: Examples of the first 12 false pairs (sum-of-squared-differences)

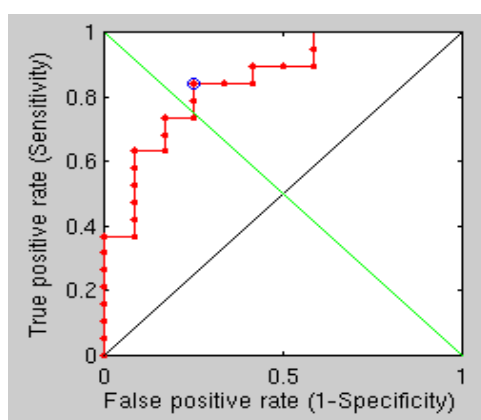


Figure 242: ROC curve generated by a sum-of-squared-differences-based similarity measure

7.12.10 2006 Experiments and Geodesic Masks

Further experiments – in particular ones with increased resolution (as in number of sampled vertices) – did give some decent results, but these did not necessarily supersede or consistently outpace the performance previously seen (at about 3,500 vertices).

In order to make further improvements by harnessing a fundamental rethink, the FMM code from 2006 (IEEE publication) was studied as it already thoroughly addressed/studied/justified the problem of facial recognition as applied by measuring geodesic distances between fiducial points with locally-acquired data (see Figure 243. Geodesic masks, such as those that we tried exploiting before (in earlier GMDS experiments), had been used back then as well.

Returning to the problem we are tackling and applying various forms of masks (also with a small buffer to latch onto) has not yet produced superior results. The main limitation does not appear to be resolution, especially not once a certain threshold is approached. There is some inherent variation there and a piecewise process is what we work on implementing at the moment (Figure 244. This clearly works a lot better than the Euclidean approach as it is robust to simple geometric changes. But even upon closer inspection it seems clear that GMDS can be too 'permissive' in the sense that it matches different noses very well, without a great enough penalty in the stress sense. The trick is making GMDS tests more stubborn and rigid.

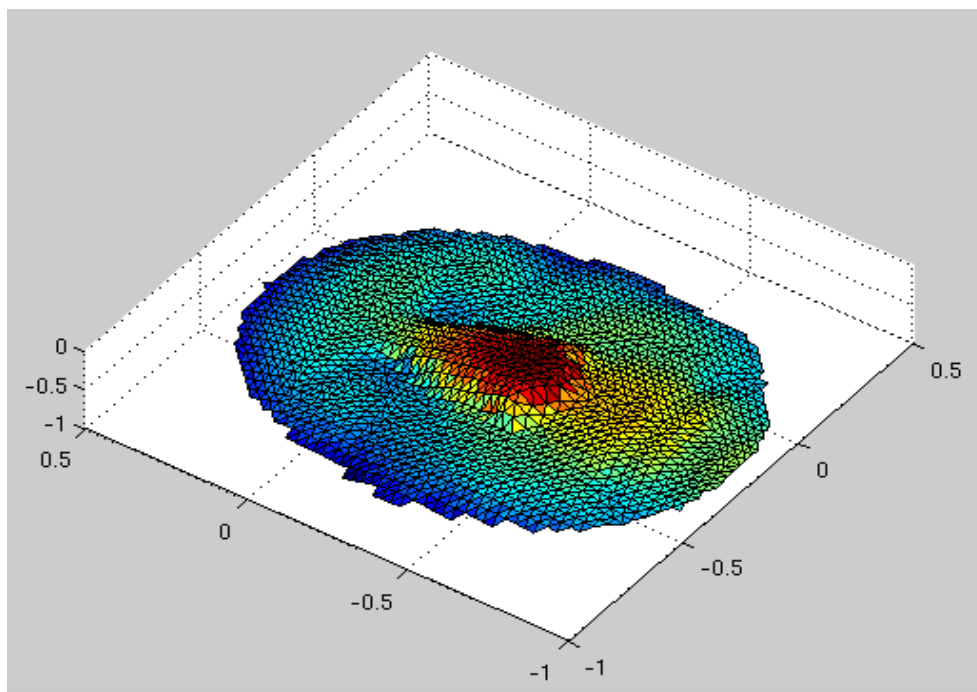


Figure 243: ROC curve generated by a sum-of-squared-differences-based similarity measure

Purely geodesic comparison with no errors can be demonstrated in small experiments. We spent a long time running and tweaking the more valuable among the experiments to examine the effect of various parameters in the similarity measure, e.g. by raising the number of points from 50 to 250, and 350 (other parameters helped differently).

With boundaries that are Euclidean altogether removed, we are no longer limiting ourselves to any criteria either than geodesic and then, combining it with a Euclidean measure as before (for borderline cases), perfect classification can be attained for the smaller experiments conducted to test the

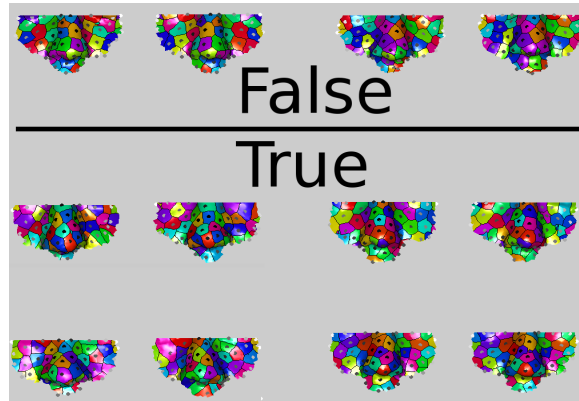


Figure 244: Examples of matches between "true" pairs and other matches between "false" pairs (different people). The separation is not yet profound enough to get state-of-the-art recognition performance.

surface, so to speak (with 60 images). For ROC curves, bigger experiments will be designed.

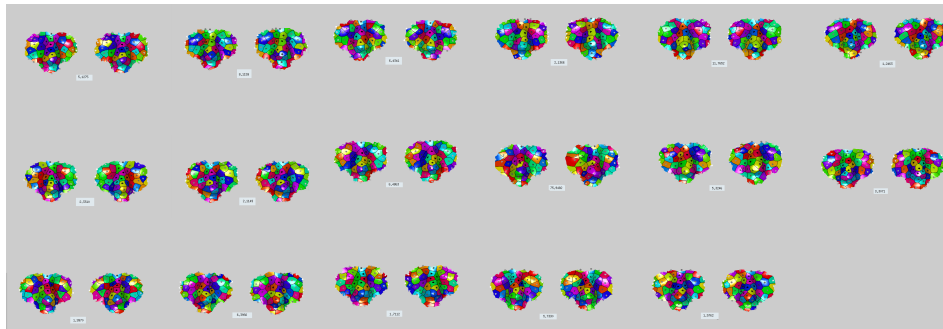


Figure 245: Example of similarity values after a Euclidean delimiter (above the eyes) was removed

7.12.11 Preparing Larger Experiments

We've debugged and resolved many of the recurring crashing patterns, so experiments can now be run a lot faster, without starting a new session

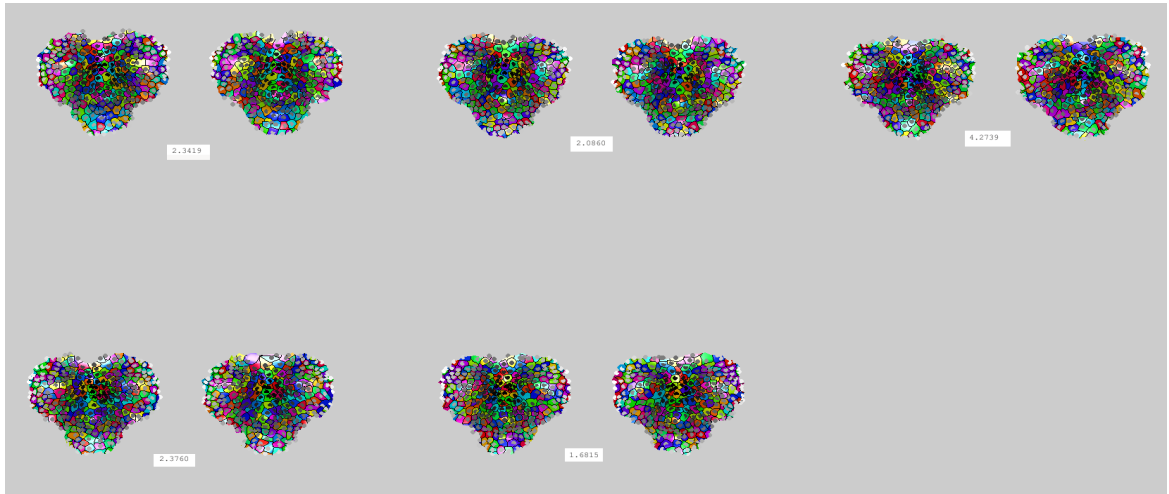


Figure 246: Example of similarity values with more points

following each crash. A lot of parameters have also been set based on trial and error, with results being quite satisfactory even at a low resolution (which gives stability and speed).

We have run extensive experiments where the number of vertices and smoothing kernel vary so as to be more strict about local variation yet attain better spatial information and thus latch onto similar structures, respectively. It's about striking the balance between being too stringent or lenient (or false positives versus false negatives).

We will start running large experiments and share ROC curves. Comparisons of ROC curves were part of the tweaking process, but these ROC curves are neither too interesting, nor do they use more than a fixed test set on which meaningful comparisons could be made very rapidly, on a case-by-case basis too. A lot of these experiments were not important enough to merit sharing

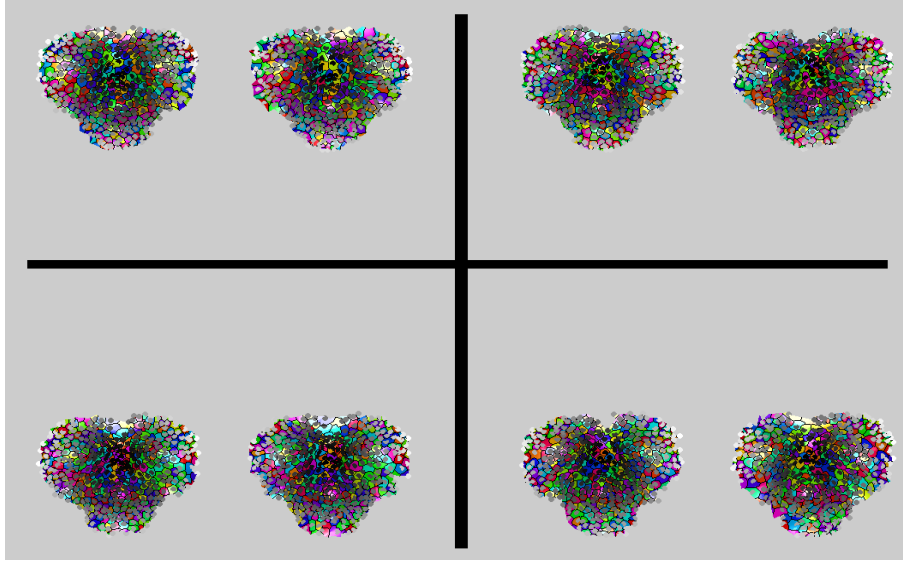


Figure 247: Number of points pushed higher towards 350 (near the maximal allowed value)

of interim results.

The experiments with 2000 vertices (and other properly set variables currently give good recognition rates. They also embrace a hybrid approach by using Euclidean measures to resolve cases of uncertainty.

The experimental set is being expanded at the moment. Having run it on a toy example with "hard" cases²⁰ to see the effect of using a hybrid approach, we get the triplet of curves shown in Figure 250. There is a lot of room for improvement, but the purpose of this experiment was to show the hybrid approach bringing recognition levels above 90%, which GMDS on its own can only ever achieve with simpler data (this data is deliberately difficult).

²⁰To accentuate discriminative properties, e.g. for comparative purposes and debugging.

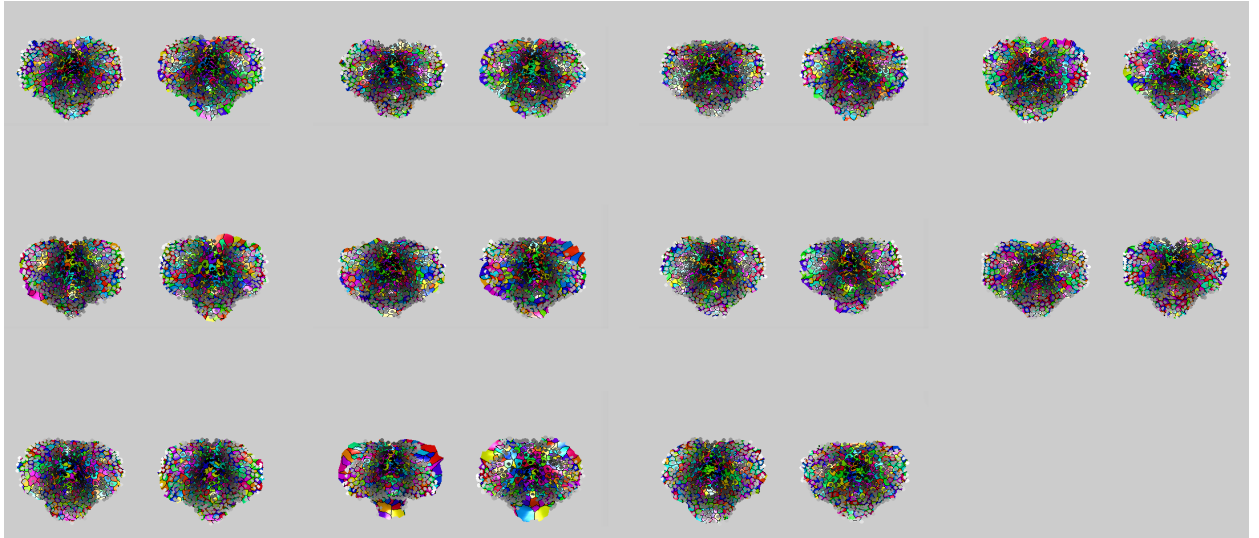


Figure 248: Pairing examples with false pairs, 1000 vertices on each

Things will improve when proper experiments are conducted.

7.12.12 False Positive - a Dilemma

Whilst in the process of programming a way around false positives (different people detected as identical) the following image was produced to provide insight into the process. For reasons of speed/pace of progress, not many such figures were previously prepared.

Shown in the images (see Figure 251) are 4 problematic image pairs in a test set where one pair is wrongly said to be similar (by GMDS alone) and 3 others are close to the margin, which still causes false positives to be encountered. On the surface, it should be easy to tell them apart, but based on a geodetic test around the eyes and nose alone, it is hard to draw the line

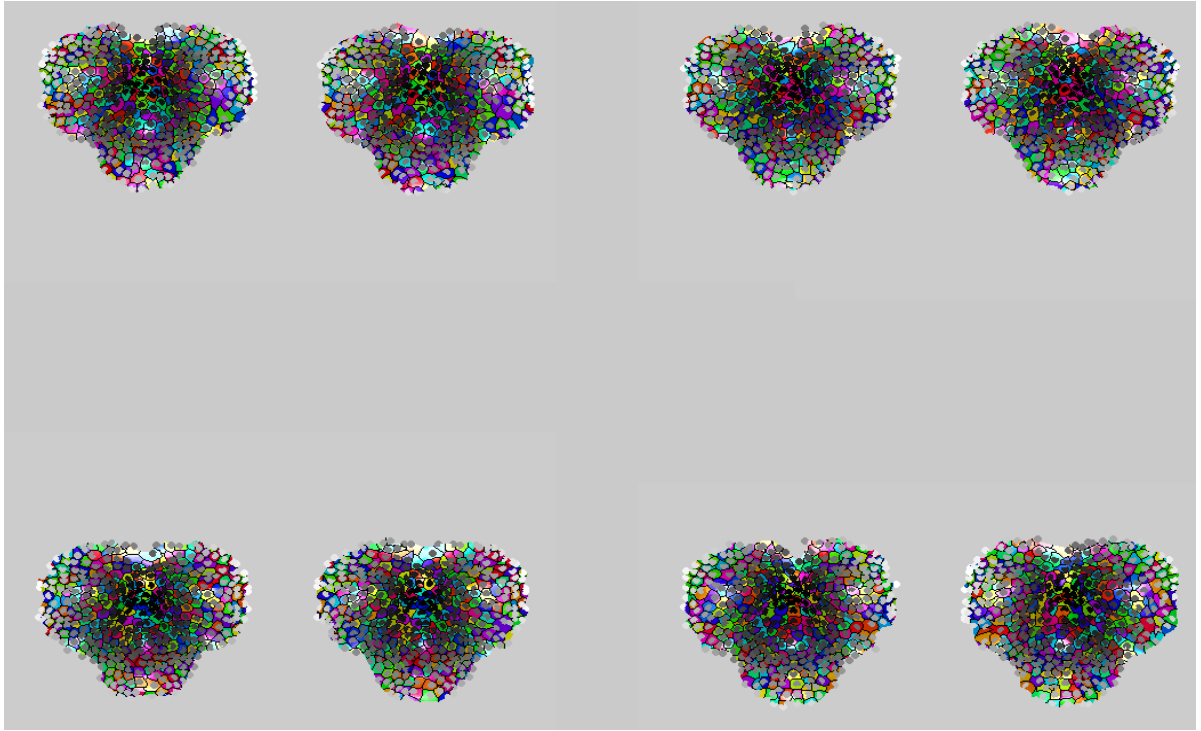


Figure 249: Pairing examples with false pairs, 2000 vertices on each

between identical and non-identical (need leniency as well as stringency). Measurement of distances between landmarks would work well, but it would miss the whole point of using an intrinsic geometry-based method, merely emulating popular/mainstream methods instead.

We now hope latch onto similar parts of faces. We could marry the classics and GMDS, extracting the best from both worlds.

This might defeat the purpose of the original plan/intent, but then again, it would not be research if we knew all along that GMDS can yield 99% recognition rate without – say – LDA applied to Euclidean-geodesics conjunction.

I shall come up with a way of intertwining those two logically such that they are not treated as entirely separate and independent (but texture will still be discarded, only the surface will be used).

7.12.13 Fallback Discriminant

To augment GMDS with another discriminant that is more anatomically aware and can be invoked in cases of uncertainty, a rather Euclidean-esque measure was sought.

Fiducial points are assumed to be unavailable as in reality, for example, manual markup won't be provided because it is impractical for fast assessment/comparison. Texture makes it easy to identify areas like eye corners, whereas the nose is easy to accurately locate based on geometric distinction. So assuming absence of texture and fiducial points, in order to use Euclidean distances between different points on the planes (3D only, not 2D+3D) it might be required to decide on geometrically distinct properties, such as the point at which the area above the eye socket flattens and becomes part of the forehead. This is not robust to eyebrow movement however, which brings into play emotional or expression-imposed variation. One natural substitute for this would be the steepness of the nose, which is quite immune to variation by expression and can help discern individuals. However range of change there is minuscule.

Another possibility is to explore alternative ways of measuring geodesic prop-

erties, for example placing a fixed point, carving around it a geodesic circle and then measuring the Euclidean distance to the edge, based for example on the sum of absolute differences in 3 dimensions. A similar pair of surface should be carved at similar positions. Exploratory experiments were to take a look at the potential of this property, based on a case-by-case assessment. This exploratory experiment should give a rough idea of the potential of Euclidean combined with geodesic means – a bit like measuring their volume in space along each dimension, at least when confined to lie inside a bounding box.

Having run some experiments manually (with somewhat encouraging results, as shown in Figure 252 and Figure 253), it seemed reasonable to carry on and implementing this in code, seeing what kind of ROC curve would result from it (see Figure 254). By incorporating further refinements we might get a discriminant more effective than the fallback currently in place.

7.12.14 Making a More Stable Classifier

With 3-D only (no texture) methods in mind, the pursuit for a stable classifier might be a combination of geodesic, geodesic-Euclidean, and purely Euclidean criteria (not photometric, but surface-based only). In addition, several PCA-based criteria are available, but they have not been combined yet as they measure range images or their derivatives very sparsely (the curse of dimensionality in PCA limits this considerably and makes it less

pragmatic). We are combining these different criteria, excepting the weaker ones that do not help much in determining the outcome. But one that is explored today, following preliminary experiments that showed some merit, is one that gradually expands the surface around the eyes and the nose, then measuring Euclidean distances on these gradually-expanding boundaries. In a sense, this is the gradual measurement – in a 10-step process right about now – which accumulates distances by traversing the triangulated surface with the expectation that identical faces will give similar distance differences as the geodesic circles grow bigger and bigger (or conversely, smaller and smaller as it is currently implemented).

Results will be shown in terms of some ROC curves. This is slow enough to take hours for one ROC curve, especially because the code is a lot more inefficient than it can be (if optimised and polished a little).

7.12.15 Occlusion Based on FMM for Matching

Rather than explore GMDS as a self-contained solution for the computation of face-to-face similarity, FMM is being used in a level sets-esque approach where we expand the surface and then do matching hinged on Euclidean measurements, carried out upon the resultant sub-surface. Based on the overnight experiment which yielded the ROC curve, there seems to be potential here for a measure at least complementary to GMDS. This was a rather shallow implementation designed to just test the waters, so a bet-

ter experimental design and algorithm will now be put in place to see how much performance can be improved. This current trick is a better substitute/fallback for GMDS than before.

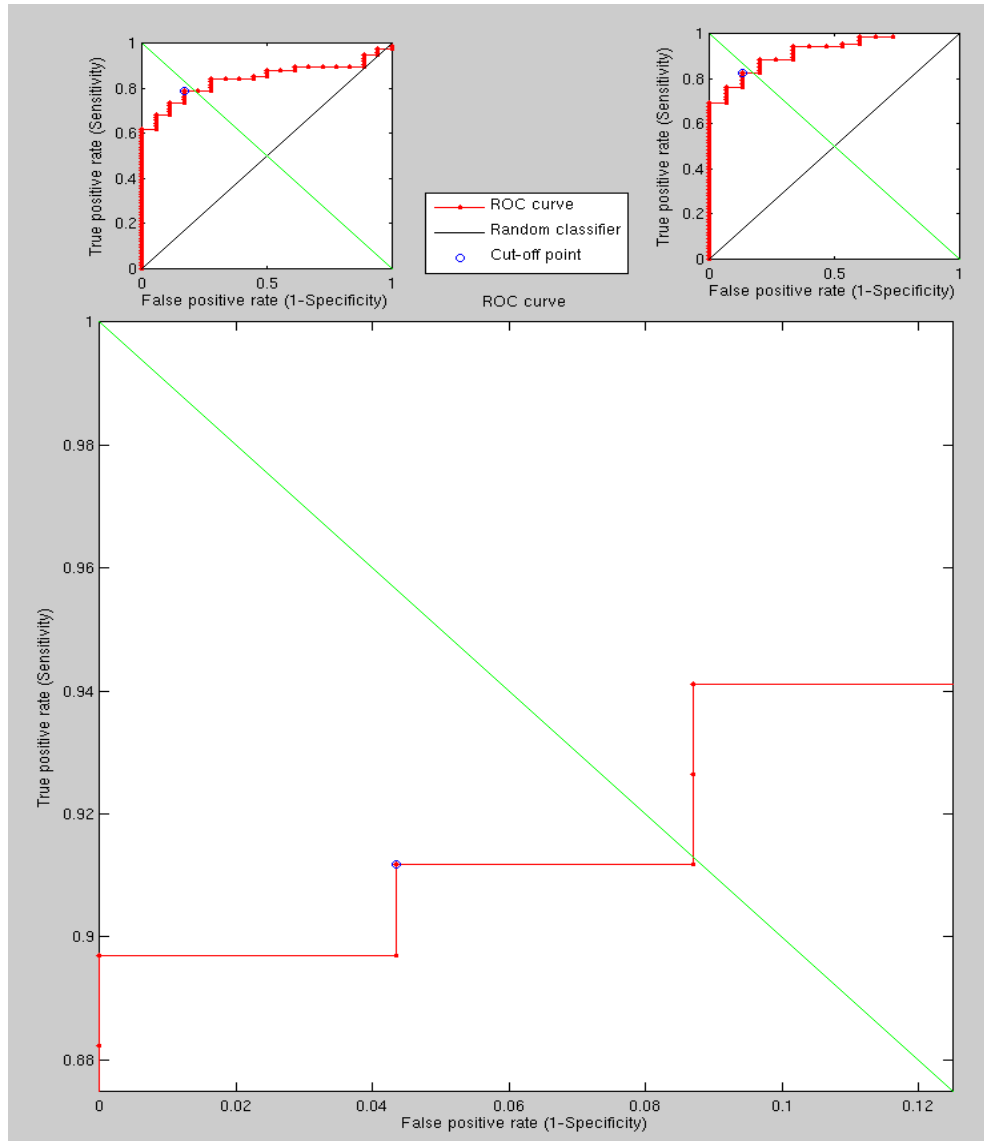


Figure 250: At the top left is just a naive implementation, the top right shows what happens when GMDS failures get detected and removed, and the large plot shows what happens when Euclidean measures are factored into this toy example.

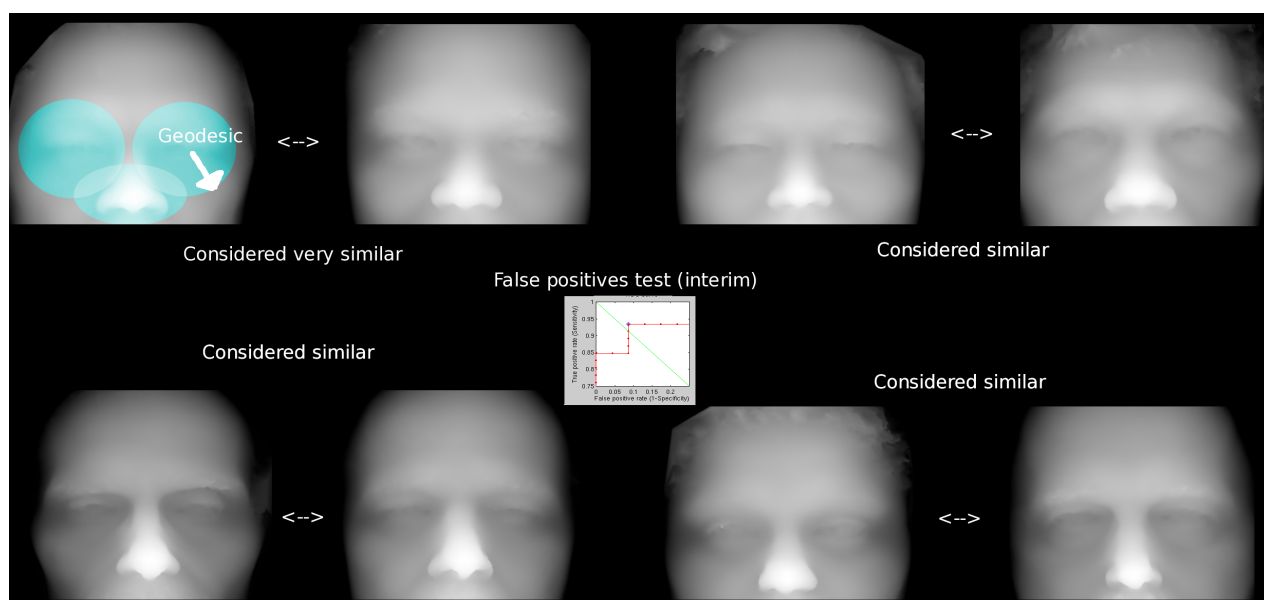


Figure 251: 4 problematic image pairs

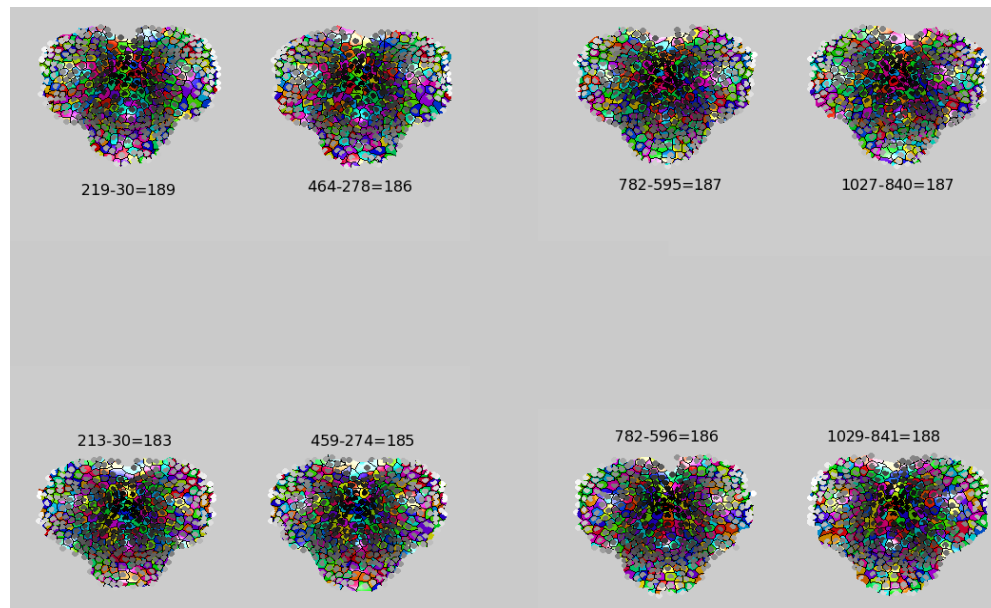


Figure 252: Manually-measured width values for pairs of faces corresponding to different people

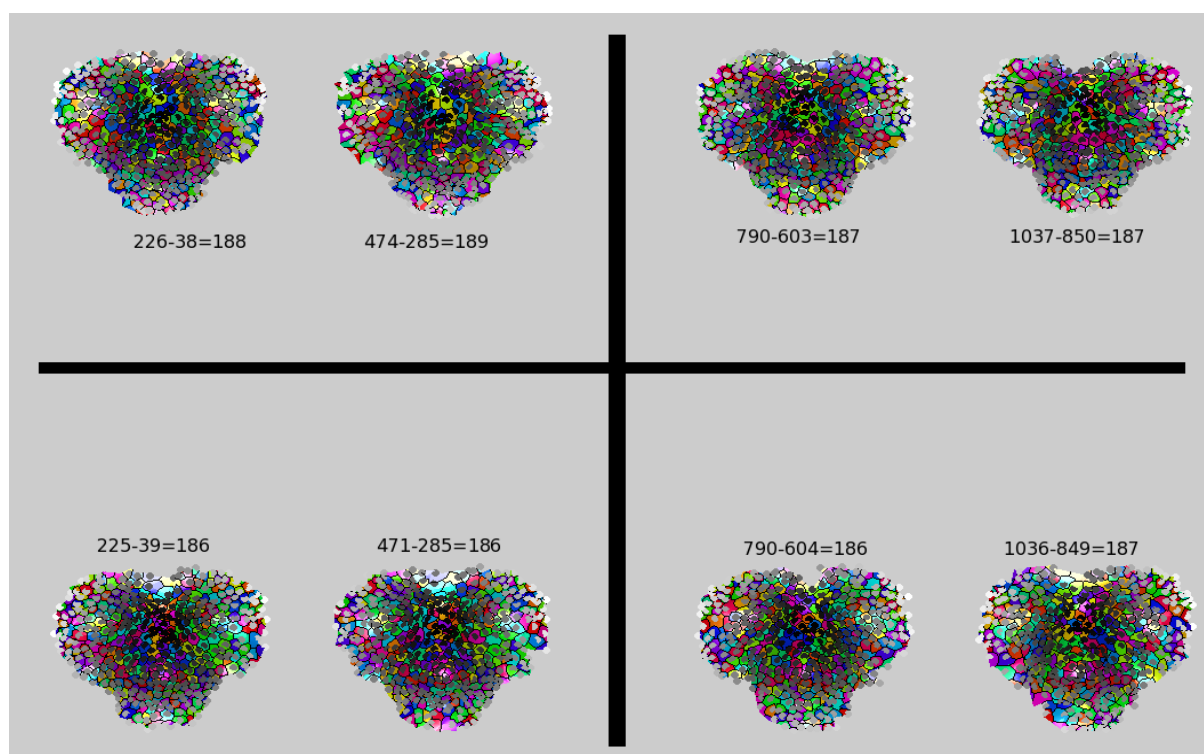


Figure 253: Manually-measured width values for pairs of faces corresponding to the same person

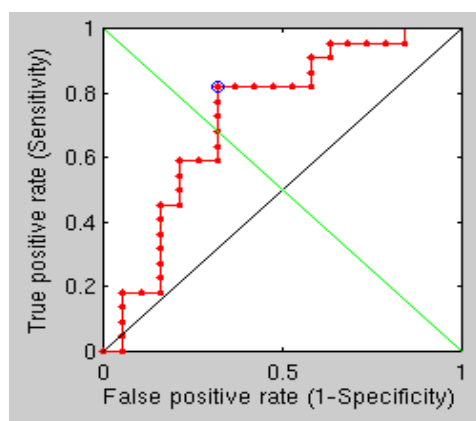


Figure 254: A geodesic ring/circle-based measurement as applied to tell apart anatomical equivalents from inequivalents

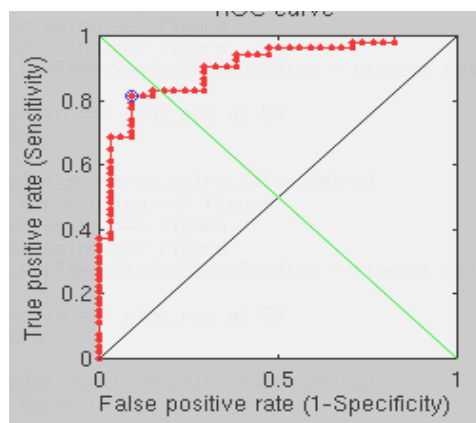


Figure 255: FMM is being used in a level sets-esque approach

7.12.16 More FMM Results

Further experiments which look at Euclidean measurements upon geodesic circles (no GMDS) were able to smoothen the previous curve a bit and change in the range of circles improved recognition rates somewhat. This is not a substitute but a complement for GMDS – specifically for cases where GMDS does not provide a satisfactory classification (high uncertainty). This newer approach can be further refined although it takes overnight experiments to autonomously ‘manufacture’ a decent ROC curve that provides sufficient comparative insight. I shall place more markers on the image to apply FMM to as it ought to amplify the signal and cancel out some of the noise (e.g. beards and other acquisition errors). No fiducial points have been used thus far, obviating the need for any human intervention in this process. No texture data is being used, either, just the raw surface.

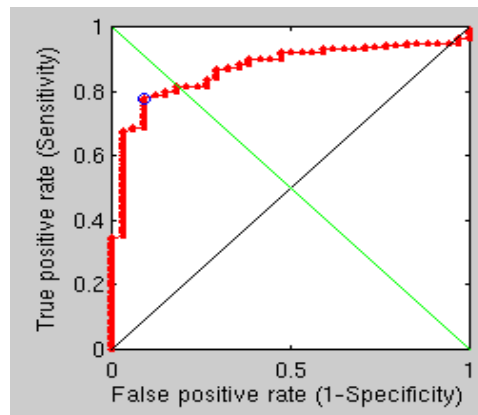


Figure 256: An extension of the original (first) experiment which explored FMM (with Euclidean measures) as a classifier

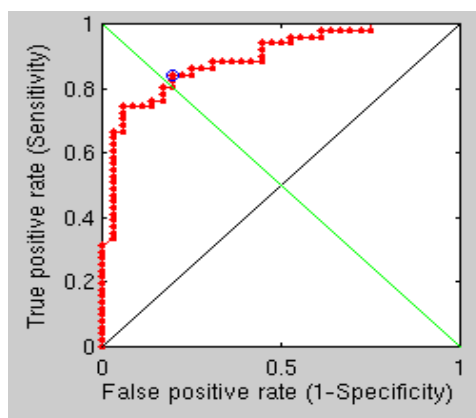


Figure 257: Results from an extension of the range of radii/distances traversed from 20 to 50

7.12.17 FMM-based Dissimilarity

In existing (ongoing) experiments, rather than vary the geodesic boundaries, the locations of the points get altered, under the assumption that this can provide a greater source of variability, covering a greater extent of the surface being probed (in isolation for separability of regions). This is not stochastic yet, but it can be made so.

The results are interesting so far (no mis-detections), but more of them are required to draw some meaningful conclusions. GMDS might not be ideal for measuring FMM-dependent similarity, so composing a substitute or complement for this task might make sense, improving it one step of complexity at a time (assessing what improves it and what does not). Ultimately, perhaps a problem-specific or similarity-optimised method can be devised as a substitute rather than a fallback for GMDS and/or PCA (where scale and thus

speed/memory are an issue). The sensitivity of GMDS was at times also a weakness, matching things that oughtn't be matched without a penalty large enough.

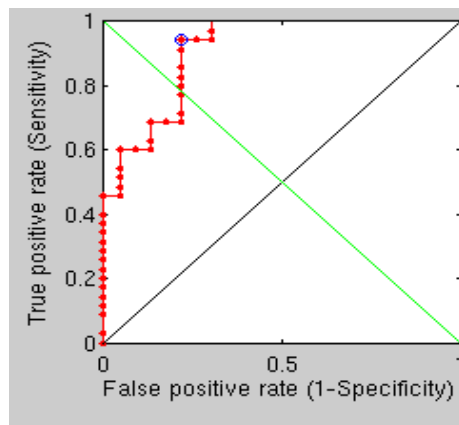


Figure 258: Interim results (70 images) show 95% recognition rate with FMM-only (no GMDS) utility, but this tends to degrade as more difficult images are presented. Two good recognisers (classifiers), one of which is a Euclidean-geodesic hybrid, might give pretty good and mutually-independent results without using texture or fiducial points.

Various images that GMDS deals with just fine are not handled as easily by this other method I gradually refine (a hybrid of FMM and a level sets-inspired technique), so they can correct one another and make a better joint recogniser. One problematic pair, just for the sake of an example, is shown in Figure 259, which is basically detected as almost belonging to the same person (it is actually on the margin of uncertainty), so the new method ought to be made more sensitive and less permissive. Currently, the results it yields can be seen in Figure 260 and Figure 261.

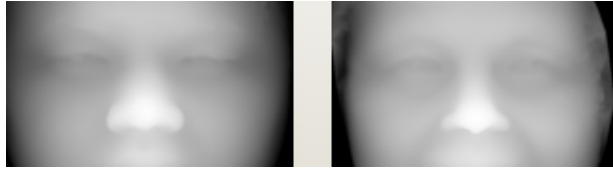


Figure 259: An example of two images from two different people, which nonetheless the FMM-based recogniser cannot quite detect as being different

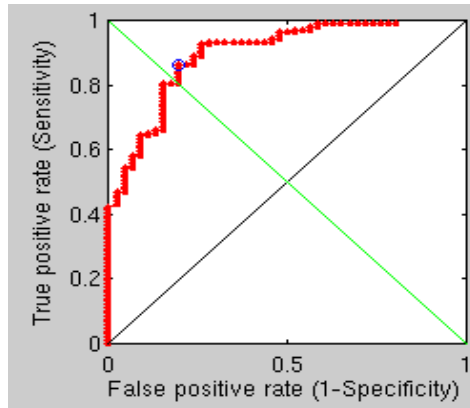


Figure 260: An FMM-based recogniser results in nearly 90% recognition rate now (without GMDS)

7.12.18 Rotation

An experiment was run on the two 8-core servers for 5 hours in order to test the range for geodesic masks with the effect on results. Next, the implementation is improved by adding an element of rotation that helps measure more distinct distances.

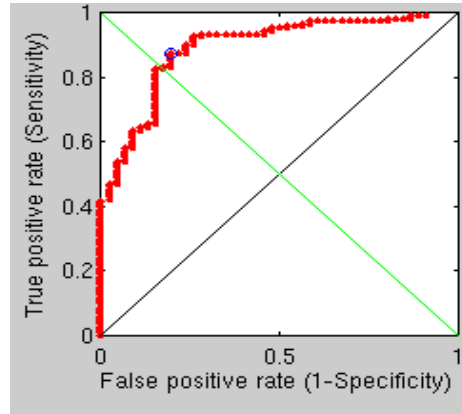


Figure 261: The FMM recogniser ROC curve after increasing the number of true pairs

7.12.19 More Data Points

By measuring distances whilst also rotating surfaces for more data points, there seems to be better ability to detect difference and more improvements to be implemented (still testing one step at a time, judging by ROC curves whether to backtrack or not). An improved version will be run overnight on the two computational servers.

7.12.20 Geodesic Slices

The Euclidean (edge-to-edge) distances around geodesic rings tend to correlate quite well with the identity of a person, which following some further refinements show a recognition rate flirting with 90%, based on this overnight experiment. Further improvements can and will be made. Combining this with GMDS can give much higher recognition rates.

The reason why rotation helps is implementation related. Geodesics are intrinsic measures. The Euclidean distances are measured along axes in three dimensions and by rotating around the axes we can conveniently measure more separable distances. The next step will be taking a sample of 60 Euclidean distances and modeling them, encoding each face as a parameterised model with PCA, then comparing faces in Eigenspace, measuring the distances between them in a clever high-dimensional way.

7.12.21 Surface Signatures

The currently-worked-on approach will try to measure distances between images in hyperspace based on their parameterised version, where these parameters are basically a small set of distances, each (hopefully) encompassing a sort of concise digital signature corresponding to a person's facial surface alone. Currently, 60 such parameters are planned for use, centered around points in the eye and nose regions (less prone to change due to expression).

The sketch shows the approach tested so far. It is a brute force implementation that measures many geodesic distances and then compares surfaces based on distance-to-distance subtractions. It is not particularly clever, but the results of recognition tests are not too bad, either. They help validate the premise that by measuring Euclidean distances in XY, YZ, and XZ (based upon geodesic operators like FMM) we are able to carve out the surfaces and extract meaningful measures from the sub-surfaces.

Another figure, Figure 266, shows the next step.

7.12.22 Vectorised Signatures

Experiments are already being run to assess common methods of separability testing in hyperspace. Given a vectorised signature encompassing the distances we deem meaningful in a given set, it remains to be determined how exactly to measure that clustering of them in a high-dimensional space, determining whether or not they fit within a particular cloud of one person or simply lying outside of it (thresholds and decision points can be appropriately adjusted, even by adding GMDS as a separate discriminant). Mahalanobis distance, Hotelling's T-square distribution for multivariate statistical testing and Fisher's Linear Discriminant Analysis can help here, but simpler units of distance are being tested first for some insight.

Scripts were produced to turn a sequence of images into an animation of small size that provides insight into variability of surfaces on which geodesic measures are taken. For particular cases, detection is made harder by motion around the eyes (including closure), but this is one of the caveats of dealing with surfaces of these kinds.

Taking the first imaged individual vs different imaged individuals (92 different individuals), the following results are obtained using the new method, which is still being refined and adjusted to the task at hand (Figure 270).

The animations show the already-aligned surfaces, which my method needs

to deal with (as shown in the animations) and then detect as "identical". It is not always easy, but I apply a lot of smoothing to annul the effect of variation inside the eyes, for instance. Perhaps selective smoothing (localised) would yield better results and it is definitely something worth studying in overnight experiments. In very large experiments there are some rotten apples that are clearly outliers (wrt to other images in the same set), weighing an order of magnitude more than the rest. I don't remove them from the results. Ideally, using multiple classifiers would help just eliminate this issue (this has not been attempted or tested yet). There is a need to code 'around' them because they stand out like a sore thumb.

Taking half a dozen random people and applying to them the same method applied previously (to one one versus 92), the ROC curve is not too bad, but there is plenty of room for improvement, especially by addressing the characteristics of outliers. Maybe a multiple classier approach would also come handy here, essentially utilising two separate methods each of which giving a high degree of accuracy.

By running analogous experiments with all data, seeds and methods in complete alignment with the exception of smoothing (moving, 13 pixels wide) we are able to see slight improvement incurred by the use of smoothing in the new FMM-based method. It makes sense to do this around the eyes, but currently the filter is applied uniformly to the entire image.

We found an error in our metadata, which caused the reported recognition

rates to be slightly worse than they ought to be. This came up when investigating those aforementioned outliers. The algorithm did just what it was supposed to do; it was the expectation which was erroneous. Tessellation density will now be increased somewhat to test its ability to help discern identities. This is generally progressing at an encouraging pace with a solution based on FMM which is tailored to the task at hand (rather than something more general-purpose like GMDS, where stringency and leniently are generally hard to balance against each other).

One caveat of this approach is that by measuring Euclidean distances upon something when moves anatomically in non-rigid parts of the face this approach will become sensitive to narrowing and expanding parts. For instance, when smiling one's cheek may move up and down a bit. From a geodesic point of view this may not be a problem, but when this is then measured in a Euclidean way the distances will change although it is the same person imaged. Raising of one's eyebrows might cause similar issues and some of the hardest images (which are most helpful for meaningful comparison testing where errors need to be common) have this sort of variation in them. This is perhaps why combining this approach with GMDS would be useful. This has not been attempted yet.

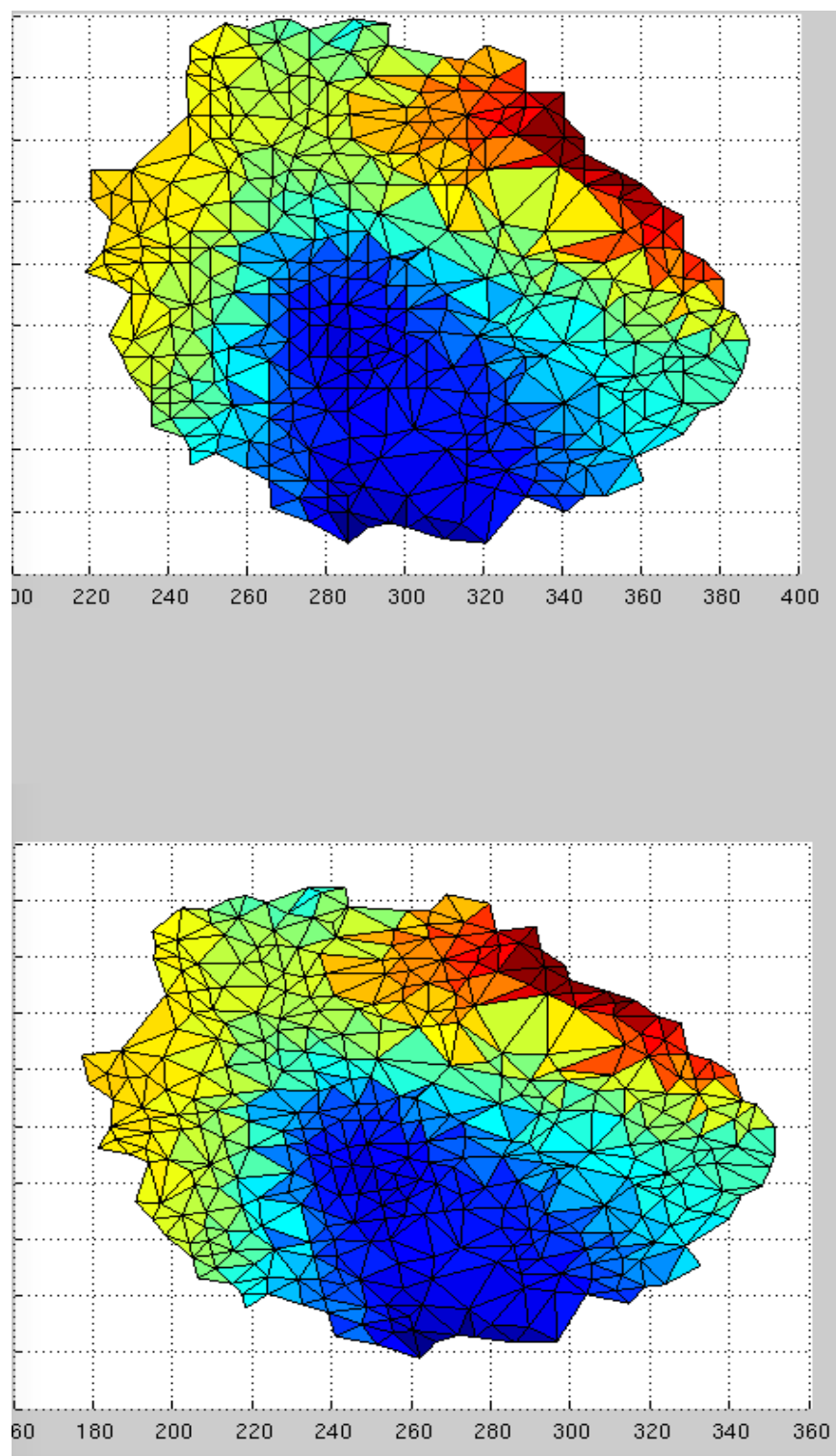


Figure 262: Example of a 10 degrees tilt

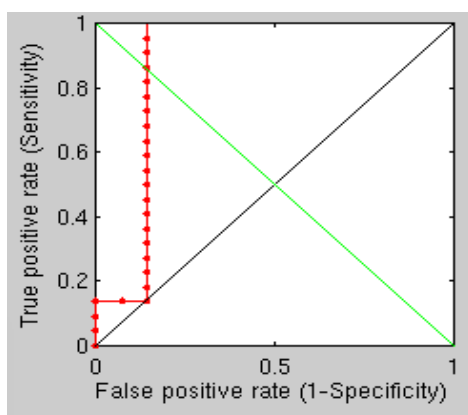


Figure 263: Recognition results from tilting one corresponding eye 360 degrees, then measuring distances on the geodesic boundaries

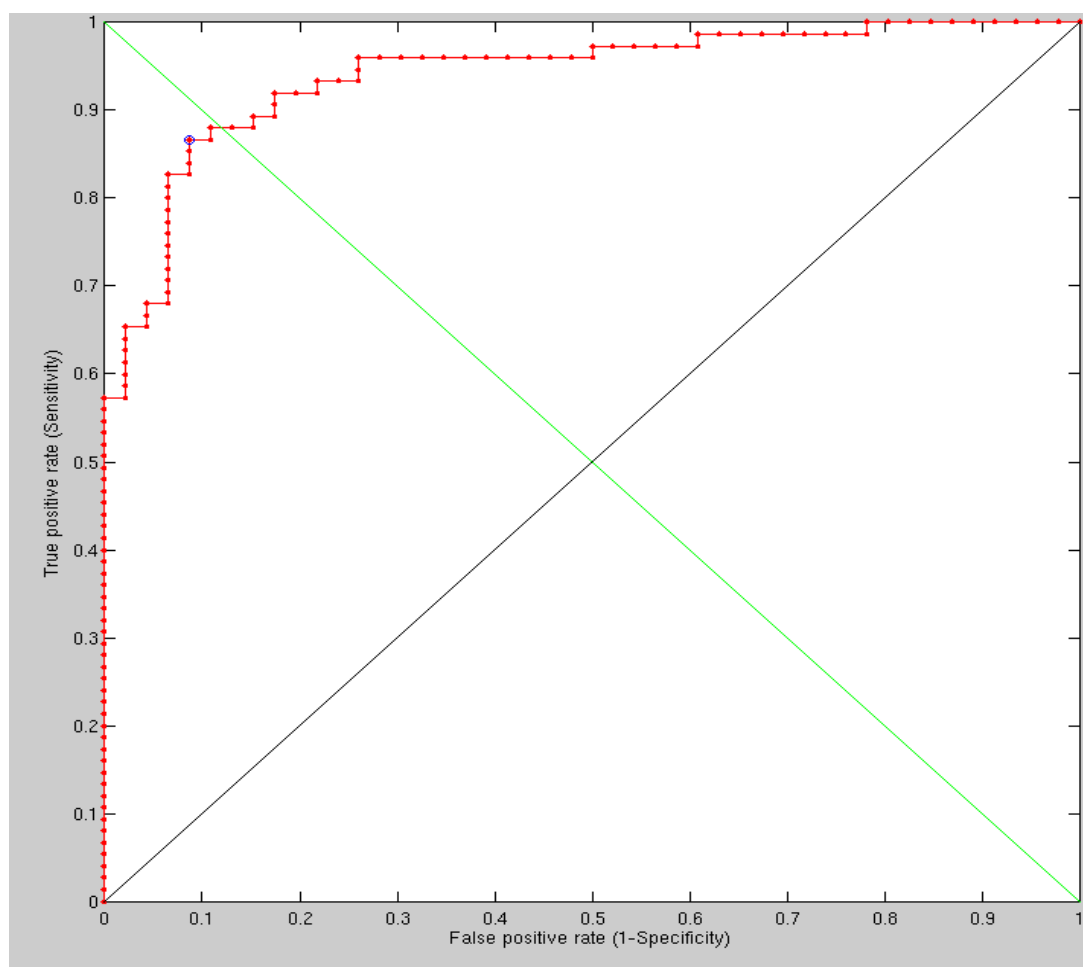


Figure 264: ROC curve based on comparing 220 images, where their Euclidean properties are measured upon geodetic slices

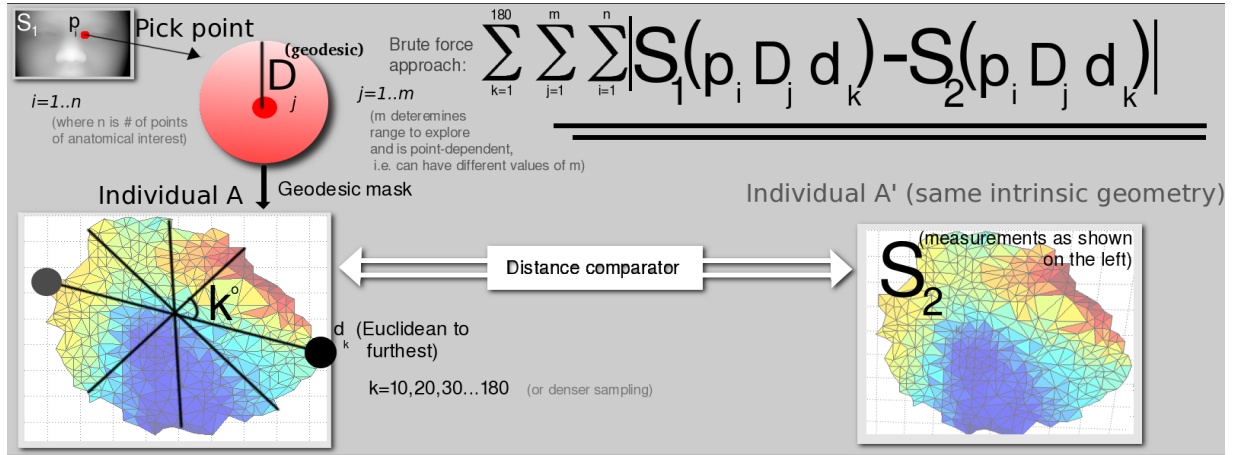


Figure 265: Brute force implementation that measures many geodesic distances

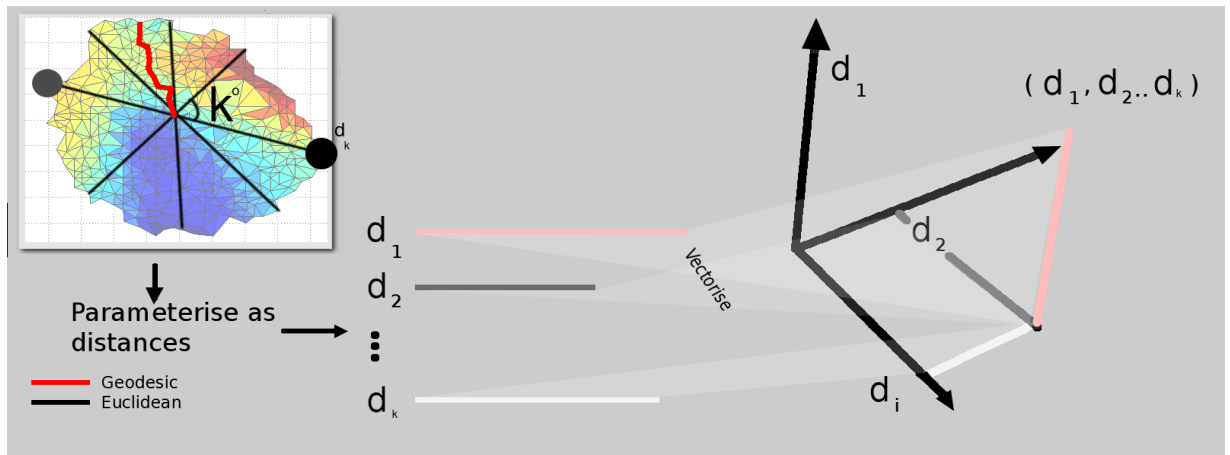


Figure 266: This figure visualises the idea of encoding surfaces as a vector not of surface vertices but an ordered list of Euclidean-upon-geodesic distances, which are fast to compute and sensitive to isometric/mildly detectable alterations

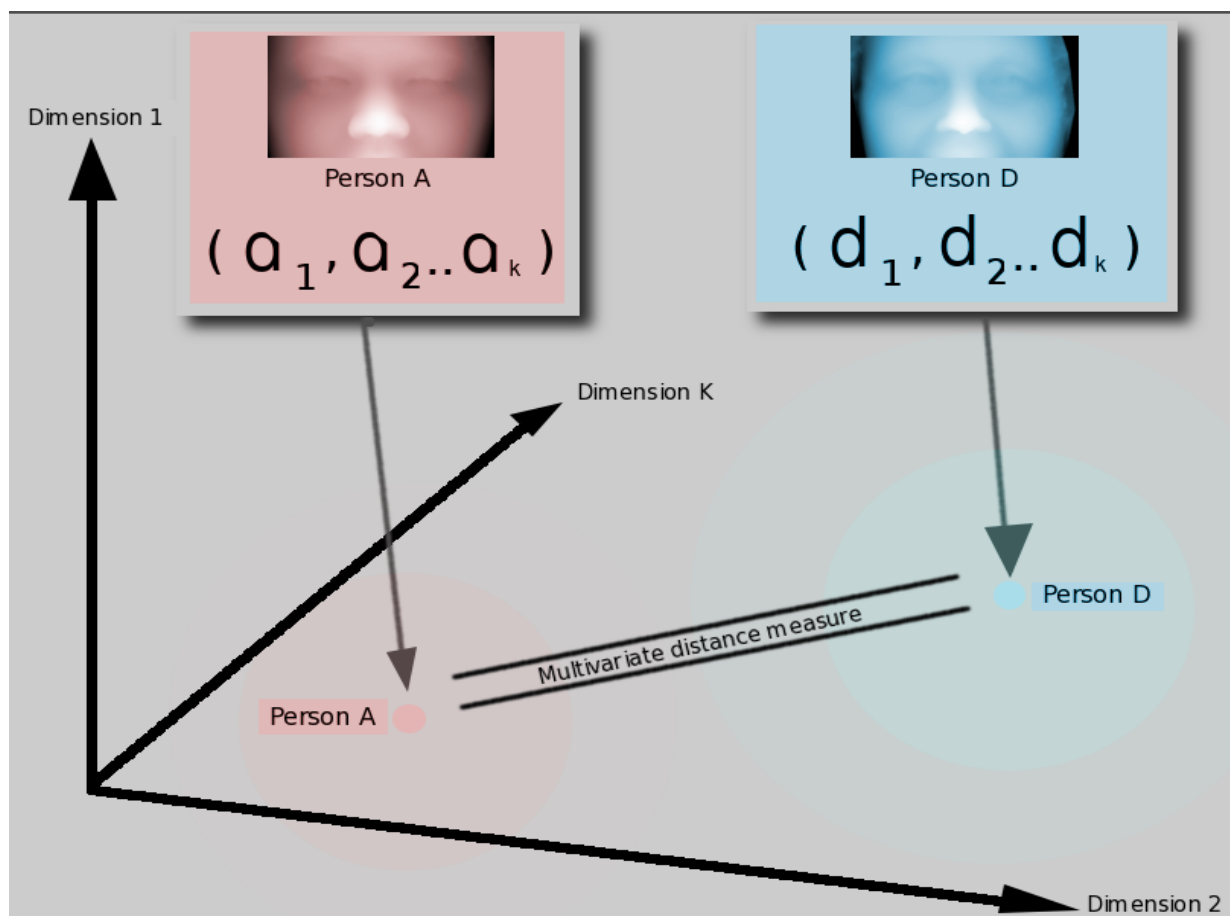


Figure 267: Separability testing in hyperspace

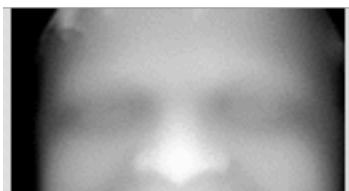


Figure 268: The image set of the first imaged individual in then test set, as an animation. The animation of the data from the 95th person is originally a GIF file.



Figure 269: Animation of the data from the 96th person

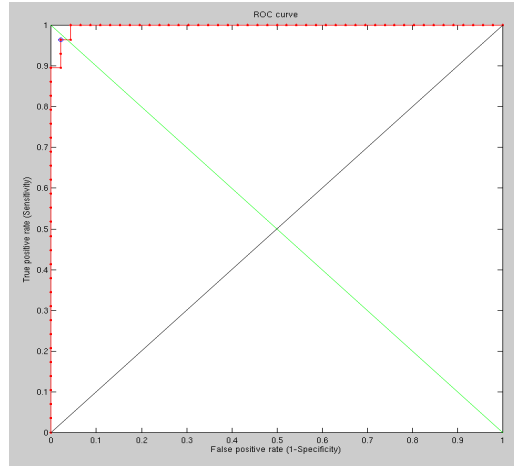


Figure 270: ROC curve obtained by measuring geodesic-Euclidean distances on the first imaged individual vs the same on different individuals

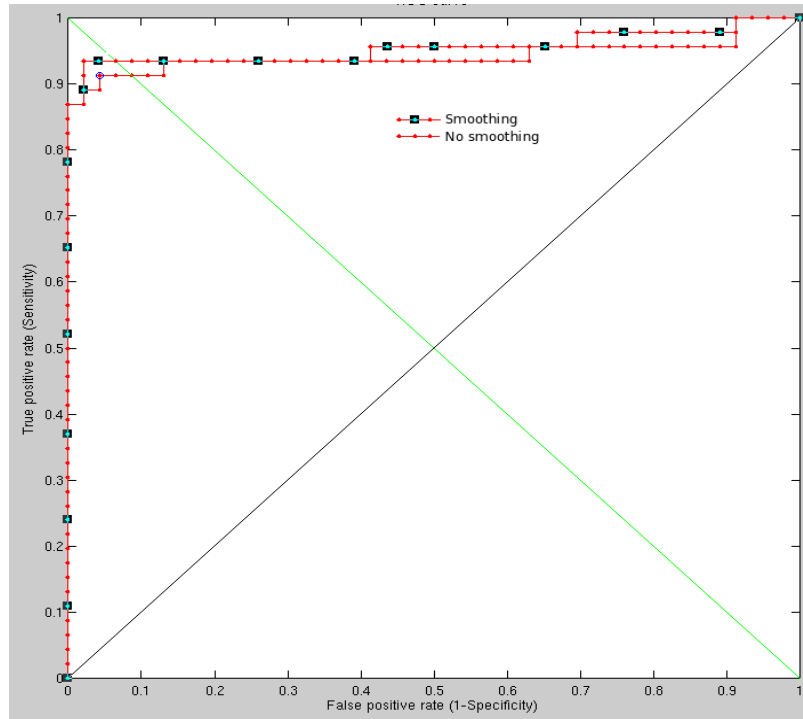


Figure 271: Smoothing versus *no* smoothing before measuring distances for identification purposes

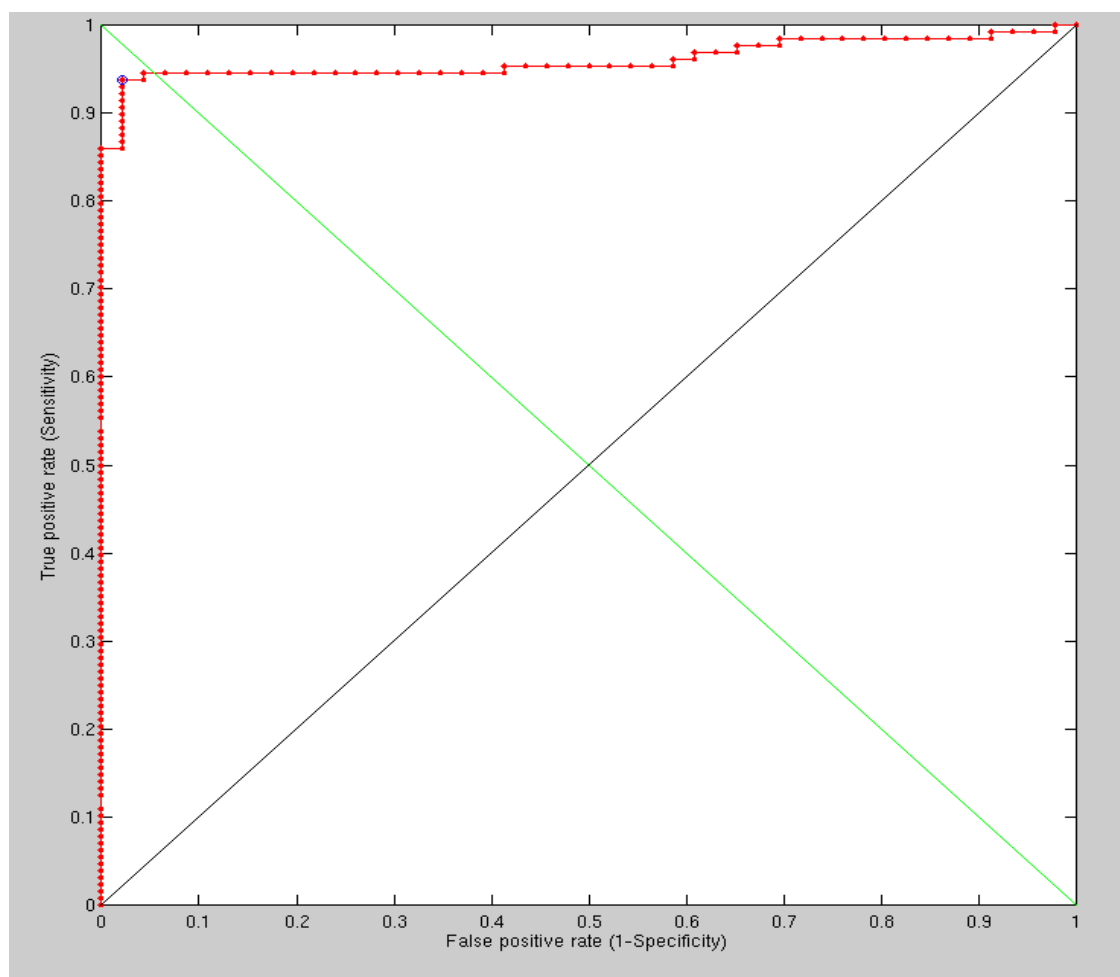


Figure 272: The result of running the test set further (not for comparative purposes)

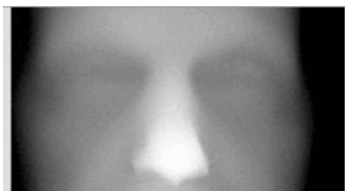


Figure 273: Animation of the data from the 103rd person. It is based on a set of images from the same person (numbered 103), without particularly challenging variation

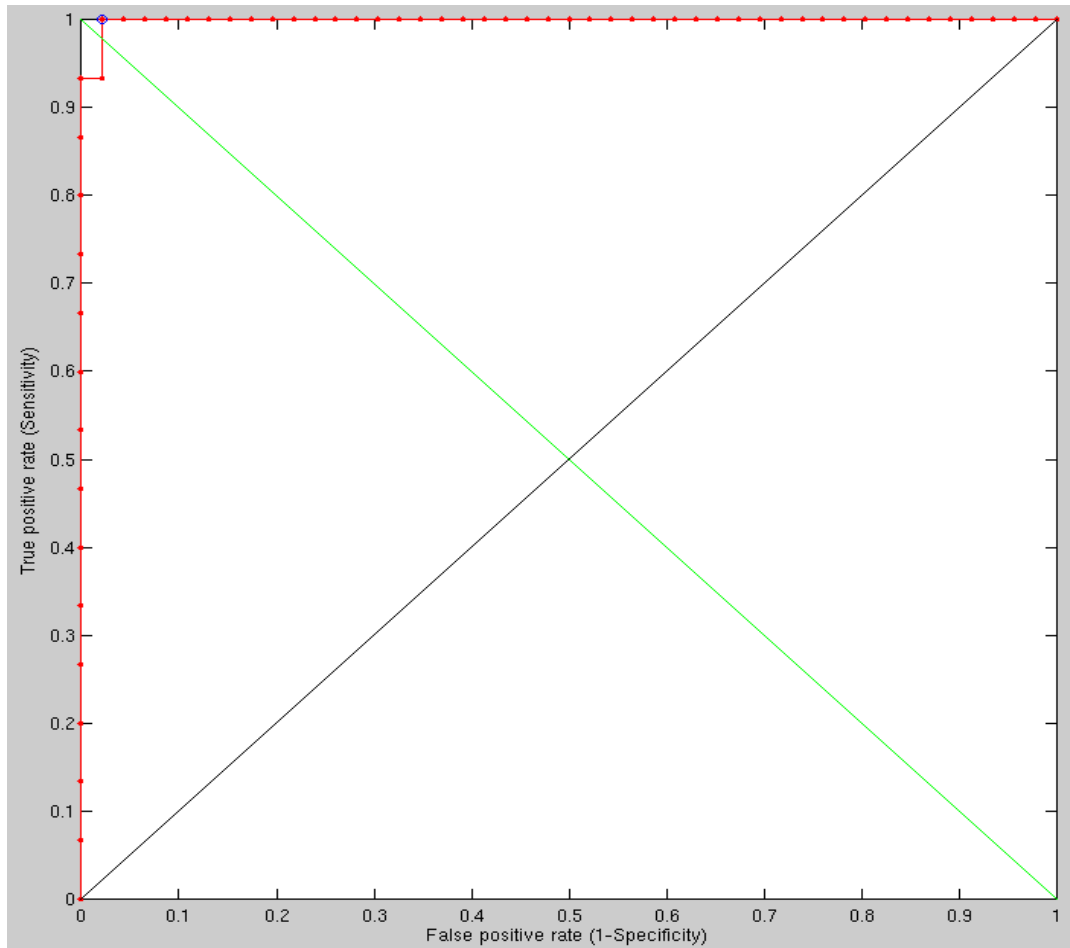


Figure 274: The ROC curves based on a comparison between arbitrary (non-identical) pairs and pairs of images from the 103th person

Having increased the 2-D smoothing scale from 11 to 19 (to verify that excessive smoothing eliminates the signal) we get the expected outcome. The results shown in the ROC curves were obtained within hours. It will be interesting to try a few more values (smoothing-oriented) to test what is optimal given other choices of other free parameters (everything is inter-dependent).

The increase in smoothing window dimensions (from 13x13 to 19x19) was not chosen just to see the effect on performance but also to understand to what degree we can eliminate the artifact of eyes closing, opening, and moving in ways that affect the geodesic distance measures in that area.

Drawing a colour map to visualise areas of distance mis-correspondence might be the next logical thing to implement as coding this might help see where and why there are errors. It is possible to do more work on this if it's seen as vital at this stage or at a later stage.

Attempting to show how it extends to a larger database would be interesting. It would help smoothen the existing ROC curves and potentially integrate with another classifier. One option would be running an experiment that simultaneously uses both GMDS stress and Euclidean distances to do the scoring, then make a decision based on multiple scoring criteria. 1,000, 10,000, or even more pairs than that can be put together for the purpose. This can take days to run, so careful attention is required throughout experimental design (getting more results would possibly require a rerun). The upside is, the results can be shown cumulatively, refined over time as the experiment goes along.

GMDS and my new method tend to produce values that are not so far apart, but they do occasionally deviate and diverge. There is a close tie and a strong correlation between the different measures where the images are, in fact, prone to be difficult. 3 classifiers (or more) can be used in a fashion

that makes expanding the set size trivial.

Experiments have begun which expand the size of our gallery of false pairs, where images are also being understood along the way if errors arise. Some of these errors cannot quite be discerned by the human eye based on mere observation, i.e. looking at 3-D data alone. By limiting the scope to rigid parts of the face (mostly eye and nose) we rid ourselves from challenges of deformity but at the same time discard much of the signal. The following image provides a real example of the dilemma being faced.

Both my method and GMDS struggle to tell apart those 3 images to a sufficient extent and it is usually the source of the only error seen in some ROC curves shown priorly.

in Texas University, based on their recent papers, they have classifiers with similar recognition rates being put together to attain about 98% recognition rate (overall, as the pertinent classifiers hover around 95%).

To demonstrate edge cases that produce some errors, the selective gallery shows 3 images that are as mutually intrinsically similar and the same three which are seen as exceptionally different based on intrinsic measure criteria. A lot of time was spent reassuring that there is no bug there or an easy tweak to parameters which would overcome this without also negatively affecting other pairs. Any tweak made may take a lot of work to understand because regenerating ROC curves takes a considerable amount of time and effort.

In the figure of cropped faces, the top line looks similar to an observer and

different from the bottom, which is what we expect, i.e. one would have classified it as such. These are really boundary cases that probably have to do with missing (occluded) parts of the nose that we interpolate for, while others apply more sophisticated tricks.

With an expanded list of image pairs (about 600, added semi-manually earlier on), both servers are running and producing results of comparisons, where so far the method is: use GMDS as primary, FMM+Euclidean distances as a fallback for borderline cases (those first two could be reversed), and other Euclidean methods (e.g. simply surface-to-surface comparators) can be used as secondary fallback in case the first two cannot return a definite answer (adjusting the threshold may be observation-based), i.e. the answer lies within the borders of uncertainty, based primarily on geodesic methods. Results will be shown cumulatively and the method revised along the way if it is found to be inferior to another. There are more principled and formal ways of combining these, but they can be explored later (optimising weights for performance). The resolution currently used is not high and the smoothing filter has a window of size 13x13. This could be improved at the expense of speed.

Experiments were run on about 1,000 images (using two servers) but unfortunately with a coding error in the simulation (it's because GMDS measured distances upon just a small region in all cases). For the sake of testing, everything will be kept coarse in the rerun as it is possible to increase the resolutions to improve performance at the expense of speed at a later stage.

Salvaging some results from the error is possible. There are lessons to be learned also from the bug, for instance the fact that sub-parts of the image are not good enough classifiers, which means that we depend on combining several.

The settings at the moment are, there is a 0.3-0.9 range that invokes fallback for my FMM-based method, with another threshold for exceptional errors (I also tried a 0.5-0.9 range for fallback), but the analysis given by GMDS would not be handy until the experiments are rerun correctly. If GMDS also fails to produce a definite answer, then we use a poorer measures or simply announce that no decision could be made reliably (passing rather than making an erroneous guess).

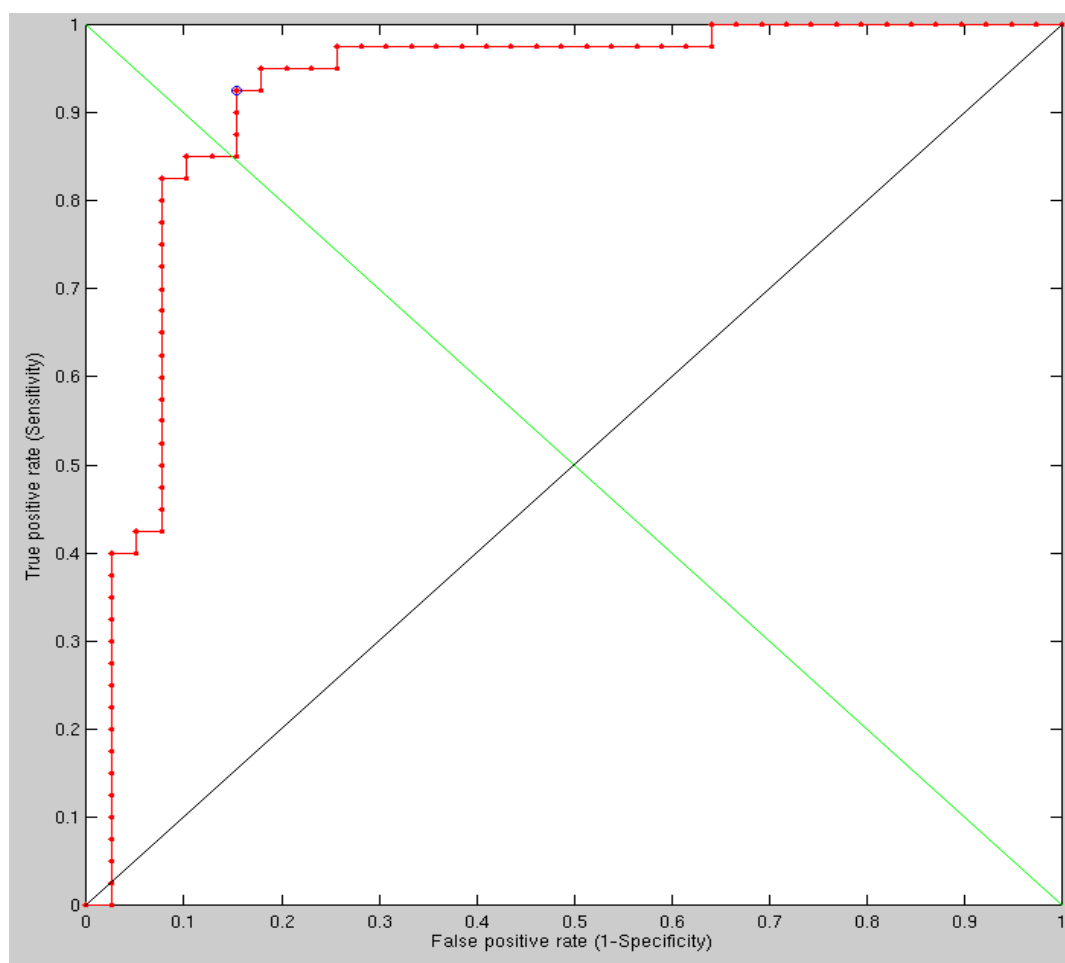


Figure 275: The results following an increase in the smoothing range, demonstrating significantly degraded performance

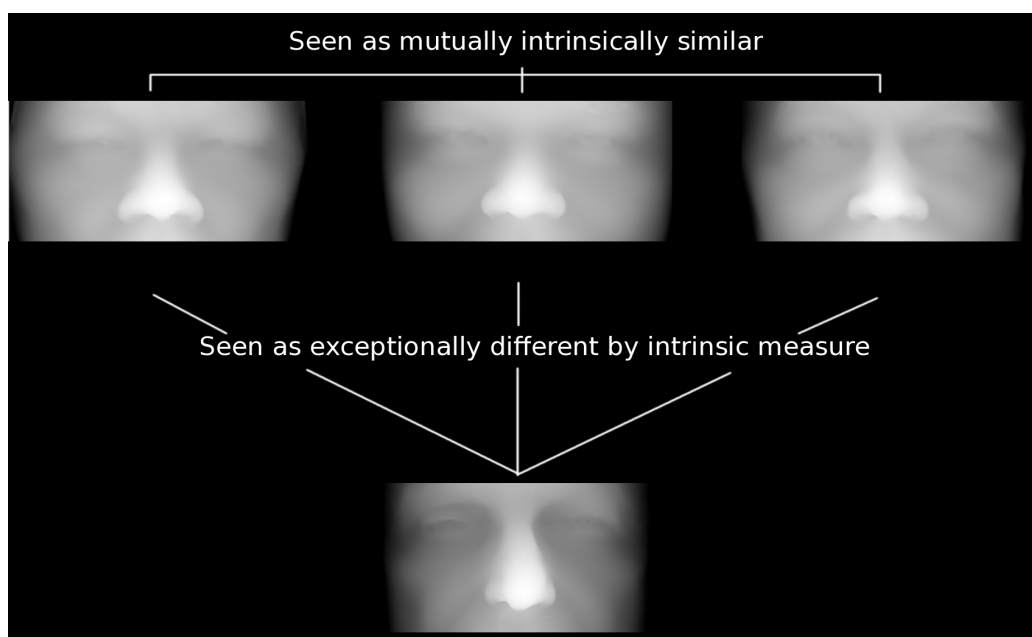


Figure 276: An illustration by example of some images that prove to be challenging in the sense that their intrinsic properties are so similar that they almost get classified as being the same person (depending on how the threshold gets set)

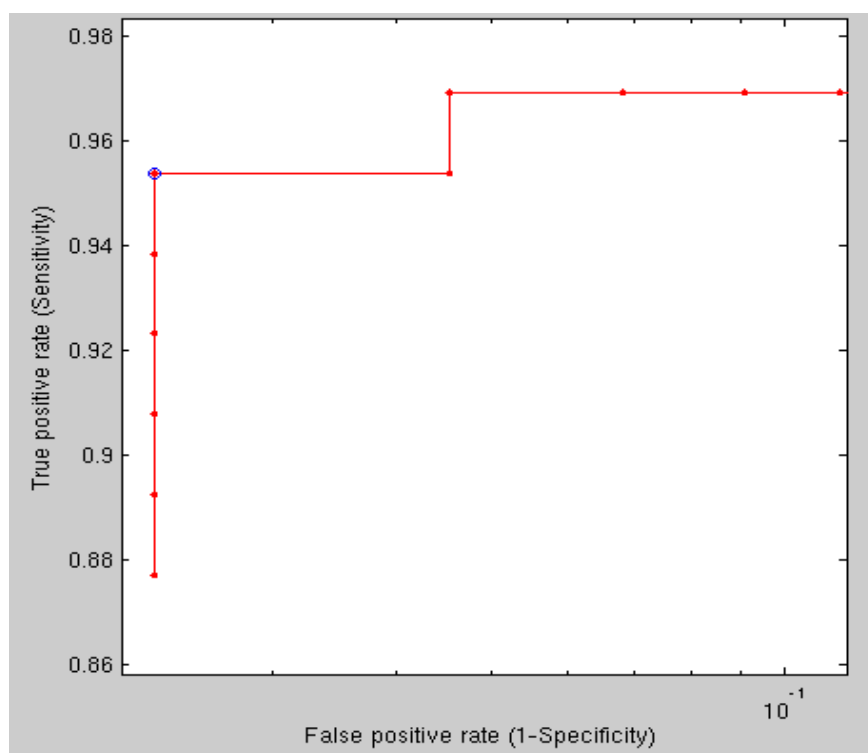


Figure 277: Detection rate of my FMM-based method without GMDS as fallback (just annulling cases where fallback is invoked). X is log-scaled.

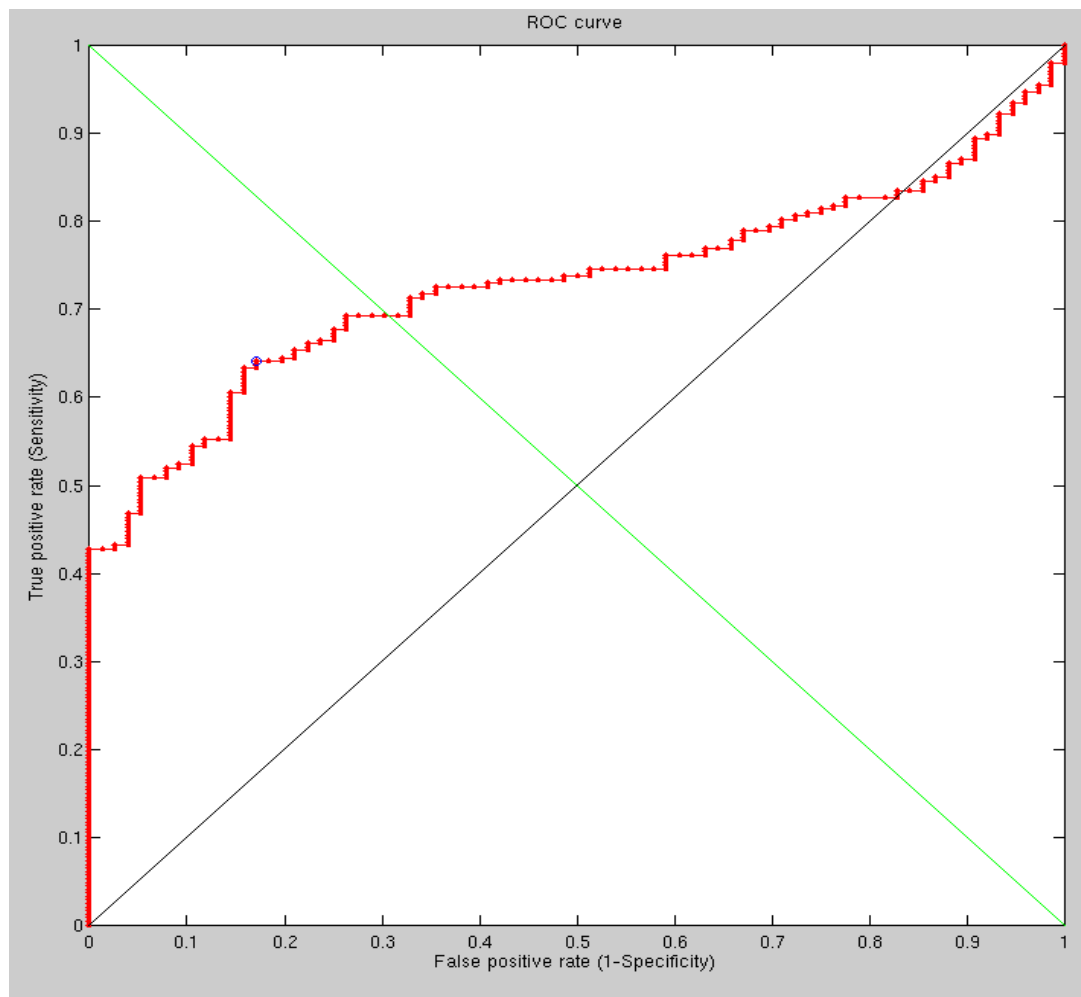


Figure 278: Results of GMDS applied to one single region rather than many, demonstrating the importance of having enough samples

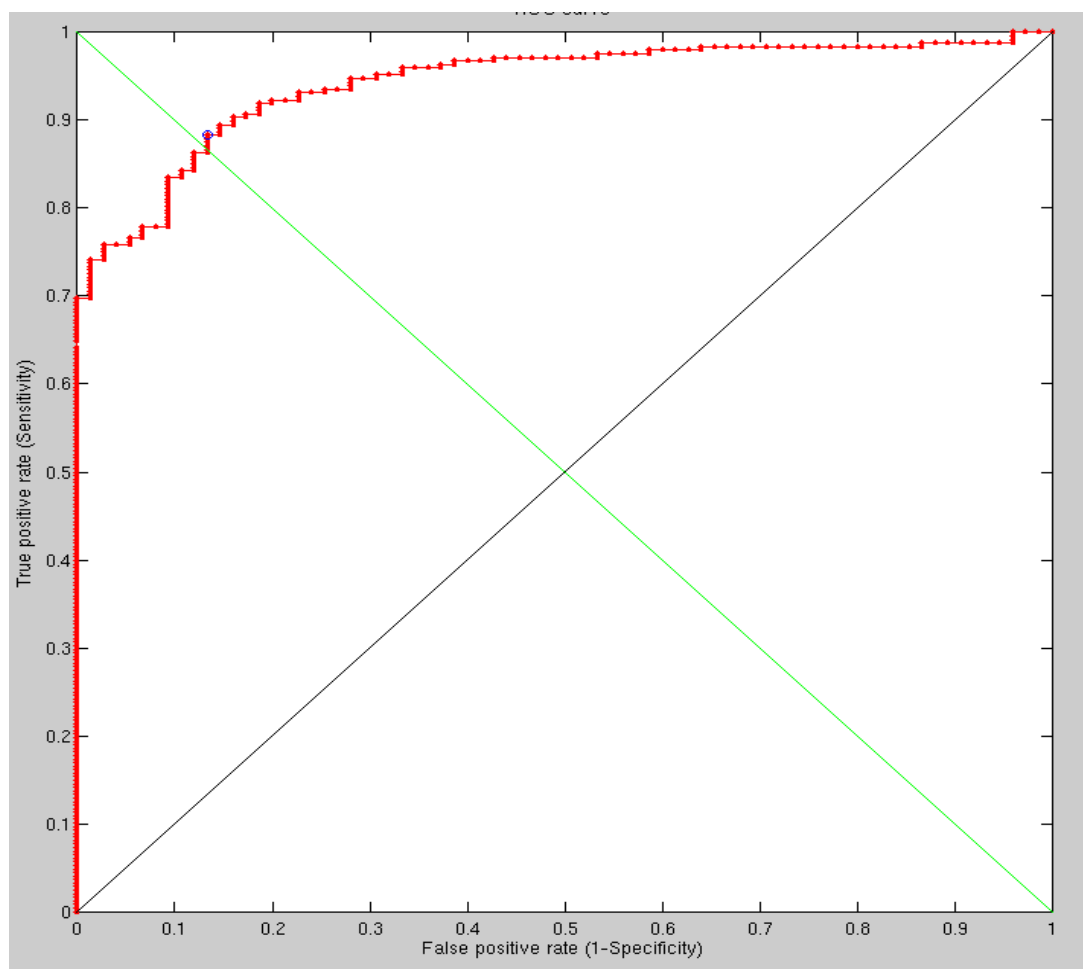


Figure 279: FMM-based method without the use of ranges for fallback (and with some errors in pairs, which degrades the quality)

7.12.23 Hybrid and Bugs

A mixture/hybrid approach does not work quite so well just yet. That would be because there is an impact on what GMDS is measuring when the resolution gets increased. We need squash a bug related to how GMDS behaves at high resolutions sometimes (which is the cause for the previous large experiments getting GMDS results wrong, and thus rendered unusable).

The method is now in a state where, if a difference is noticeable enough, then it is easy to correctly classify in almost all cases, but if the difference is not very small either, then the decision is left to be made in another way.

As expected, lowering the resolution gives poor results, so it misses the point somewhat except when the effect of different parameters gets learned.

7.12.24 Alternations to the Algorithm

At risk of being inconsistent wrt older results (which were comparable), I have made alternations to the algorithm, which, based on my observations, have the potential of boosting recognition accuracy. Preliminary results might be ready soon.

7.12.25 Eyes vs Nose

All the prior experiments weighed eyes and nose regions equally, not quite permitting one to discern the impact of each components (no visualisation

tools for the task). As a systematic experiment that aids understanding, someone could isolate these two. So, as a little exploratory experiment, hard cases were piled together to just see what areas of the faces – if taken in isolation – provide better recognition rate at borderline cases. The conclusion drawn from it is that the eyes region contains valuable discriminatory properties.

Work on exact geodesics had begun before the above experiments were concluded.

Work then involved Michael Zibulevsky who was working on exact geodesics. One is looking forward to the boosting. Regarding GMDS and high resolution. The issue is that currently the GMDS is using fast marching (FMM) to compute geodesics, which is a first order accurate scheme. There are alternatives which include using 2nd order method, also called exact, for computing geodesics. Note that with exact geodesics, the interpolation phase should also be replaced with an "exact" distance computation. It could be exact at least on the polyhedron approximating the surface.

These "exact" 2nd order methods have been studied. We I suspended other experiments for the time being.

The MATLAB code is an enclosure of existing code. The code is an implementation of geodesic methods as per MMP (Mitchell, Mount and Papadimitriou). It is an improved implementation with some C code that scales gracefully. It proceeds by identifying shortest paths on a graph in a way that

is similar to FMM.

The algorithms/code are at <http://code.google.com/p/geodesic/>.

This is an implementation of exact geodesics algorithm for a triangular mesh (first described by MMP) with some minor improvements, extensions, and simplifications. The algorithm has $O(n^2 \log n)$ worst-case time complexity, as opposes to $O(n \log n)$ in FMM.

In searching for exact geodesic it follows an approach similar to [73].

In our own experiments (thus far), we have been running experiments with just several hundreds of triangles that are coarse equivalents of face parts such as the nose and eyes. There is a limitation due to scale and accuracy, affecting whatever assesses similarity based on FMM-derived distances, which discrete nature (and no interpolation) makes the measurements an approximation.

We still suffer from dependence on the degree of information loss, resulting from pixels-to-meshes conversion. "Fast Exact and Approximate Geodesics on Meshes" describes in some level of detail the approaches of others who tackled the same problem on dense meshes composed of triangles and explains why operations like these are commonly applied to polygons in general (for computer vision, computer graphics, and more). It says that shortest paths "typically cut across faces in the mesh and are therefore not found by the traditional graph-based Dijkstra algorithm for shortest paths." Contrariwise, Surazhsky proposes an exact algorithm for an efficient implementation of the method from Mitchell, Mount, and Papadimitriou (MMP). He can do

better than $O(n^2 \log n)$ in practice, nearly removing the quadratic complexity (reduced to linear) and claiming to be able to deal with half a million triangles in less than a minute, giving all-to-all distances. For the one-to-one case, the claimed performance is a few seconds for a mesh with approximately 1 million triangles.

Another suggested paper is [\[2\]](#).

The paper deals with partly missing data too. It cites Kimmel and Sethian as "present[ing] an approach called Fast Marching Method (FMM) on triangular domains that computes approximations to the geodesic distances from one source point on S to all other points of S by solving the Eikonal equation on a triangular grid. The algorithm's running time is $O(n \log n)$ and therefore optimal. The accuracy of the approach depends on the quality of the underlying triangulation; namely on the longest edge and the widest angle in the triangular mesh."

The paper proposes a heuristic solution based on multi-dimensional scaling, which is similar to classical PCA or PCO. They have very few experimental results, with holes added to a human body mesh with 20002 vertices

We shall see how different measurements – "exact" and approximations – affect recognition performance when the number of triangles is kept intentionally low (for speed).

7.12.26 Exact geodesic_library

I exchanged some messages with someone who was vaguely familiar with the code. It seems as though this implementation is very stubborn on using DLLs, so compiling for 64-bit Ubuntu servers requires some work, including the change of MATLAB code such as "loadlibrary([geodesic_library '.dll'], hfile);" and the addition of libraries to compilation, including Boost which is 86MB (compressed). Some of the C++ code has been changed a bit and since non-Windows platforms are listed as not supported, there is no guarantee that this is going to work. For exact geodesics, many searches around the Web brought nothing up, not even links from within peer-reviewed papers.

After several hours hacking around the code to make it suitable for compilation on Linux and 64-bit platforms it still seems rather elusive because the Boost library is quite a major dependency and notorious in this regard. In general, many searches around the Web (quite exhaustive) reveal almost no other public implementation of exact geodesics for MATLAB (except the one which in turn latches onto peripheral binaries for Windows, requiring Windows servers and perhaps forking). No MakeFile is provided, so making this code cross-platform would possibly require a great deal of deviation from the original (last updated in early 2008 by Danil Kirsanov).

7.12.27 Removing Cases of Uncertainty

Culling out cases of high uncertainty, the newly-created method can make accurate assessments (when it makes them), but current experiments already have listed down areas of imminent improvement, such as increased resolution, increased penalties for errors (the black art of adjusting regularisation terms), etc.

7.12.28 Recognition Results

A couple more overnight experiments were run with the penalty term elevated somewhat, so as to better account for cases of mismatch. This improved the recognition results, as had been hoped right from the start.

There are some technical issues associated with increasing the number of vertices because this either surpasses memory caps in MATLAB (large matrices) or it gets stuck with no debugging information, requiring a restart each time. It is reasonable to assume that a lot of useful information gets lost due to this sampling limit and the smoothing, which in some sense does aid performance, does not always help so much, either. There are inherent advantages and disadvantages to this approach and the experiments help recognise them, as well as assess the performance attainable taking all the drawbacks into account. It is generally understood, for instance, that either PCA or GMDS rely on calculating everything on large matrices, which only ever subsample the original data. They can only be as accurate as the quantisation, unless

of course more sophisticated approaches are devised.

The next experiment will continue to add improvements that, based on empirical evidence, ought to entail further improvements. For 50% of the data (where there is greater certainty) it is possible to classify correctly at a rate of about 99%. Examples from that other 1% or so can help show what remains to be 'hacked' around in a way that generalises to the entire dataset.

It is worth noting how large the data base you experimented with to get to these figures actually is. The data pool is one of 1000 images and the aforementioned experiments used about half of those, not in any particular order. It is possible to add more, but this will require further manual work.

The bad 50% of the data on which we do not get 99% is simply undecided on. Taking every single pair, including those where a decision is somewhere at the boundary, gives about 93% recognition rate (in the last experiment). The goal now is to further improve that so that rather than make the classification "inconclusive" (then pass) there will be a correct classification returned almost every time. In order to understand the effect of resolution on performance, systematic experiment will be run and the results then plotted overlaid for comparison. One serious limitation right now is that the surfaces are shrunk by about a factor of 5 along each dimension, i.e. 25 times for XY.

7.12.29 Number of Vertices vs Recognition Rates

To provide insight into classification of all pairs (including ones where a decision can hardly be made), 3 ROC curves were produced.

In these experiments, one identical dataset was used so as to ensure the comparison is valid at this scale (apples to apples, not random). 300 images in total were used.

The experimental design was simple and the flow serial. This uses only one classifier – that which was originally designed to be a fallback for GMDS. It is in fact performing better than GMDS by now, as measured in terms of recognition rates.

Upping penalty terms and then moving down to 1000 vertices and 500 vertices (down from 2000), we are able to see the effect that the density of points on which distances are calculated has on overall performance. This ought to help learn whether or not adding more and more points will necessarily be beneficial or just time-consuming.

The conclusion is that even with as little as 500 vertices the results can be reasonably OK. "First order" distances are calculated on those. Adding more vertices improves things, but not by a considerable amount. This is quite consistent with the observation made following similar experiments with GMDS.

7.12.30 Trial and Error in Parallel

The adjustment of the code continues based on observation of tough cases. And based on early numbers, the results do not improve much by extending the range of search (to more parts of the face) or increasing the number of vertices to 2,700. The reason is errors. The experiments were aborted within hours because the numbers were unsatisfactory. It is currently being explored just how much of the face surface can be used without causing trouble, just adding entropy. The error is basically a purely technical one which relates to FMM and triangulation.

The general strategy is currently to run false pairs on one 8-core server and true pairs on another, expanding the size of the sample and providing output as soon as any is made available, at the very least for supervision. For more comprehensive comparative experiments that help guide the development, it may be handy to have remote access to more machines with spare CPU cycles, e.g. machines at the lab that are unused overnight and have MATLAB installed. This would speed things up by parallelising experiments (they are serialised at this stage). Results generally continue to improve based on trials (and errors) that either show improvement or degradation. The currently ongoing experiment expands the scope explored on the surface and it also fixed the penalty terms based on careful observation of previous experiments. It is ideal never to have to use penalty, but for practical purpose it is necessary (although it is a hack).

Expanding the scope of surface matching leads to a higher frequency of errors, which in turn seem to exacerbate performance, based on an overnight experiment. Some of the remaining errors are due to poor ICP, which fails to properly align some basic structures like nose and eyes, in which case, the method basically does what it's intended to do. The problem is – at least sometimes – registration prior to geodesic analysis. This can probably be corrected without much difficulty.

7.12.31 Geodesic lenses

The entire image/data set has been stacked up inside loader functions for large experiments. Some special cases were then studied in order to work around them not by detecting but passing based on borderline scores. A bug in the penalties was found and corrected, even though these penalties depend on how the algorithm is varied (must be normalised wrt other variables).

I am working on nice ways of visualising the localised geodesic errors between pairs, densely. This ought to help indicate, e.g. using colour maps/contours, where two individuals differ (if at all), at the very least helping a human assessment which uses Euclidean (human-visible) by providing 'geodesic lenses'.

For debugging or general analysis that helps understand why the same imaged person can be intrinsically different across images, a tool was made to highlight localised differences such that for each pair of any objects (not just faces), the discrepancy will be visually identifiable and therefore possible to

factor out, based on observation. Algorithms can be adjusted accordingly to avert false negatives. To be more effective, it will need to be remapped more like a compass and overlaid with some colours.

The overlay of choice is a spiral where lines represent D sampled varying (increasing) distances away from the fiducial/key point, wherein degrees are represented in a way that can relate to the original images. Overlaying the images in a way which cannot obscure anything may require colour, though.

Overlays of geodesic distance indicators are not easy to make visible, even by 'redifying' an intensity-scaled indicator of distance. At the moment, the output looks something like in:

We will improve this further in a moment.

At the point where detection rates are improving it is usually the subtle localised errors (with very high values) that put the whole classifier at peril. Using these maps ought to help judge – on a case-by-case basis – the composition of the overall score. I am currently examining the image pairs with those charts apart in order to better understand what to tweak for improved performance, especially fewer false negatives. The points have been dilated someone to improve visibility, as shown in Figure 293.

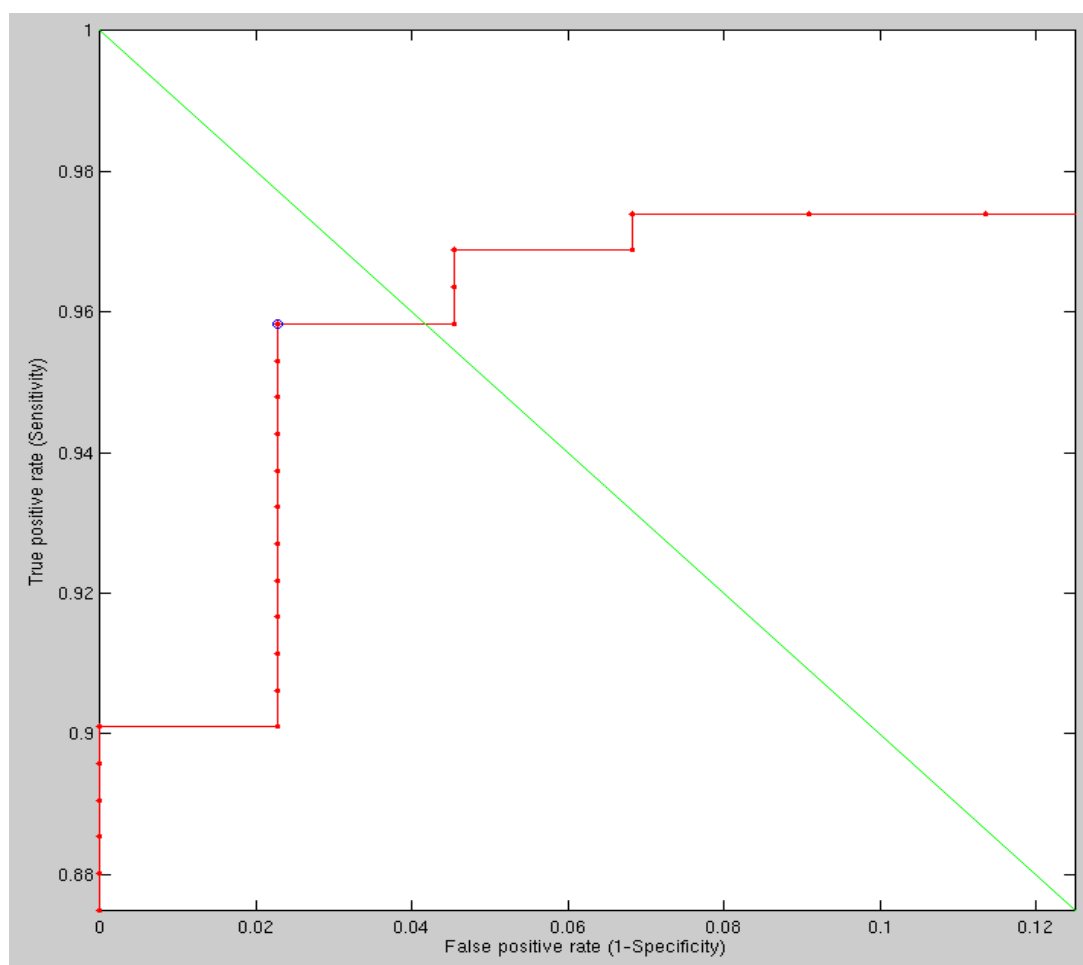


Figure 280: The result of applying the new method to pairs it feels confident enough comparing, based on pre-supplied thresholds

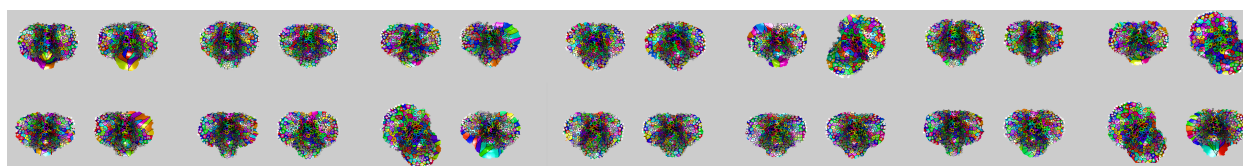


Figure 281: The problem with GMDS not finding a path through the graph in some cases, where eye regions get altogether cropped out as a result

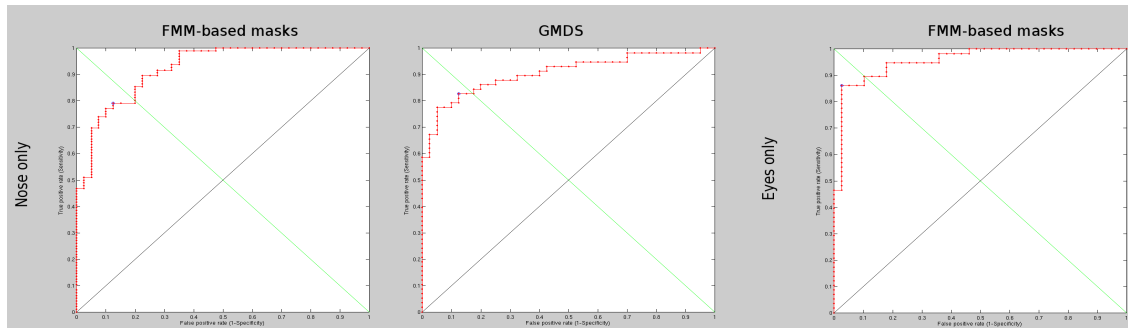


Figure 282: The performance one gets by handling difficult cases based on nose alone or eyes alone. The results from GMDS are similar to the results attained using the other method which is still undergoing development and gradual improvements, maybe with exact geodesics.

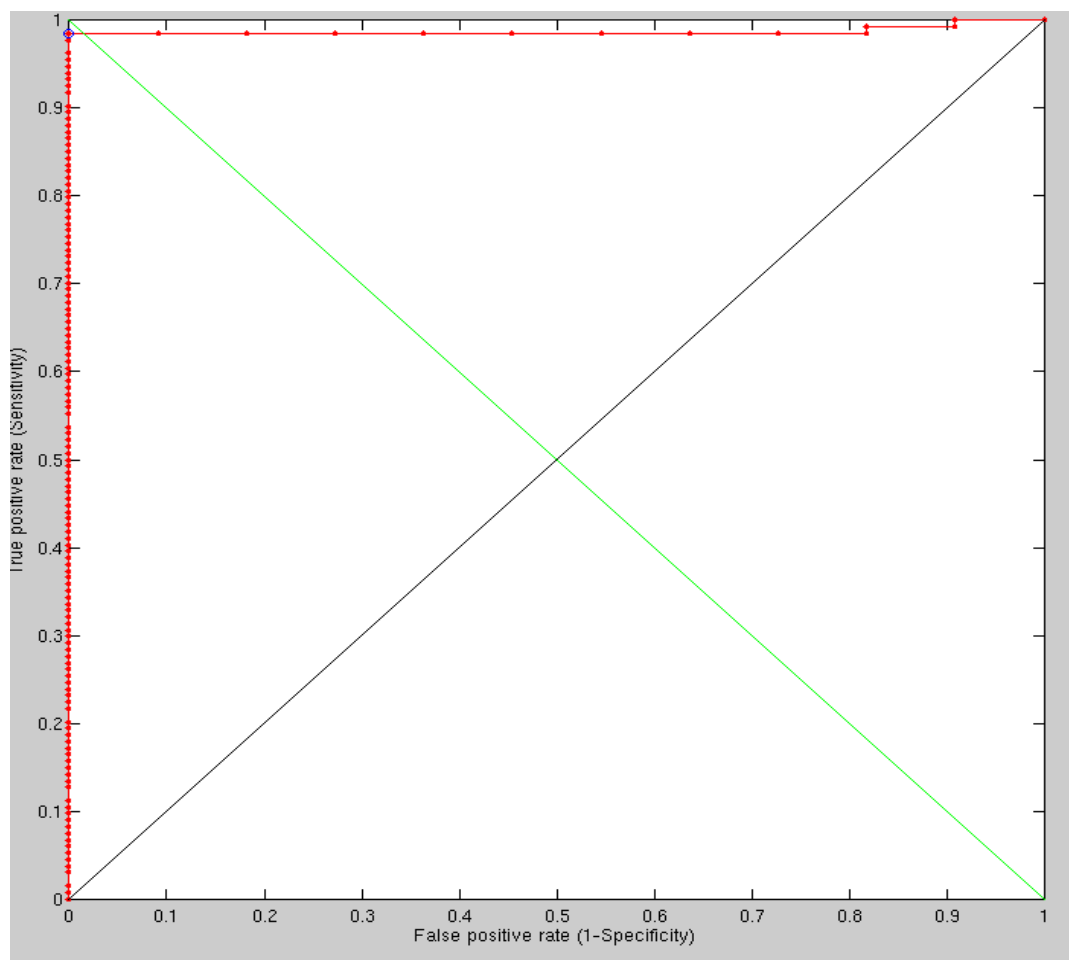


Figure 283: The performance attained by removing hard cases

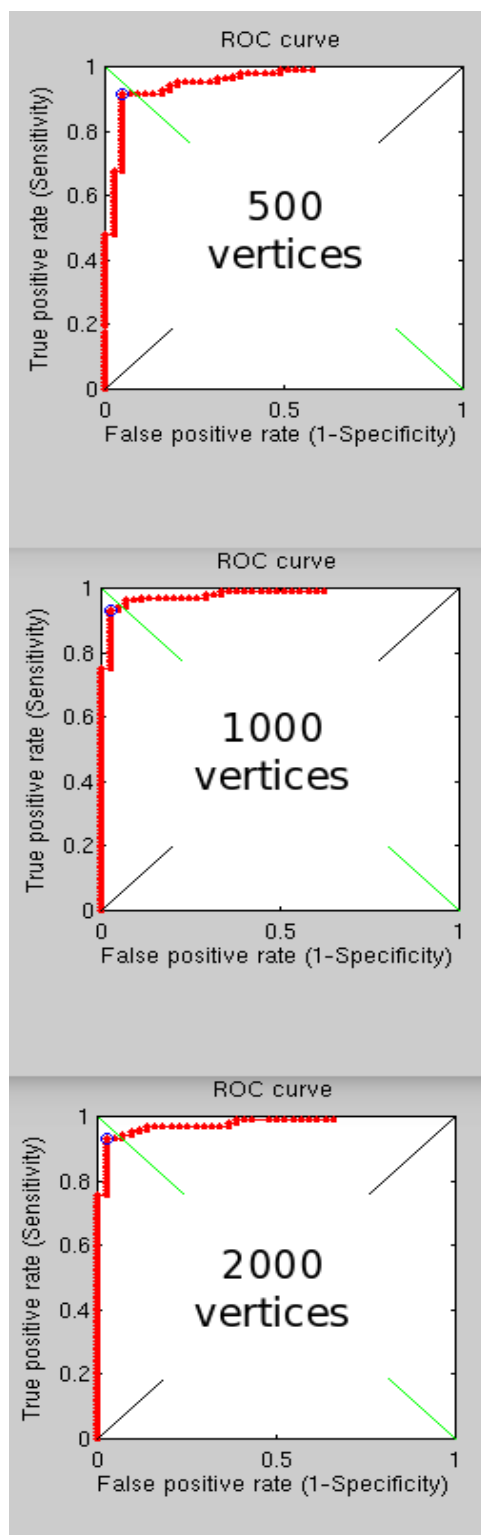


Figure 284: The performance attained by changing the number of vertices and keeping all pair examples to be judged for similarity



Figure 285: Example of poor alignment in the original set

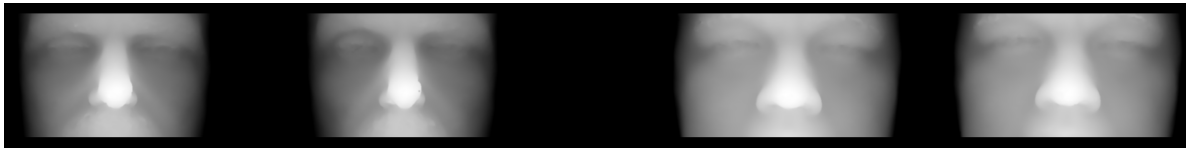


Figure 286: Two examples of easy matches from the remainder of the dataset (which was enrolled in its entirety into the experiment)

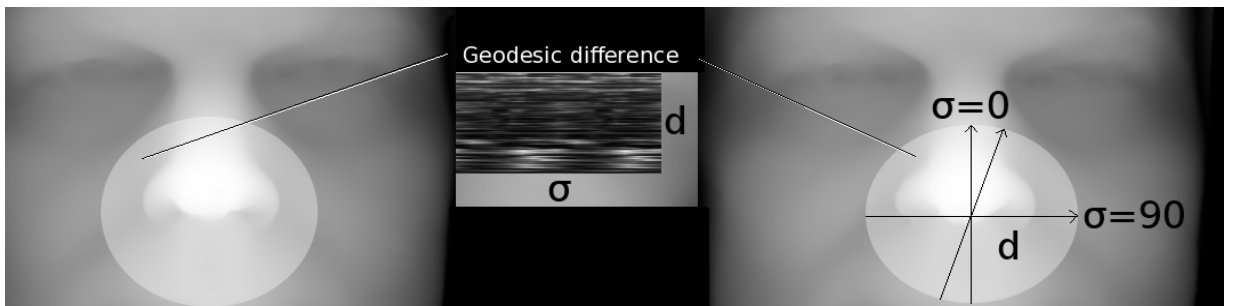


Figure 287: Example of geodesic differences map around the nose (to be improved)

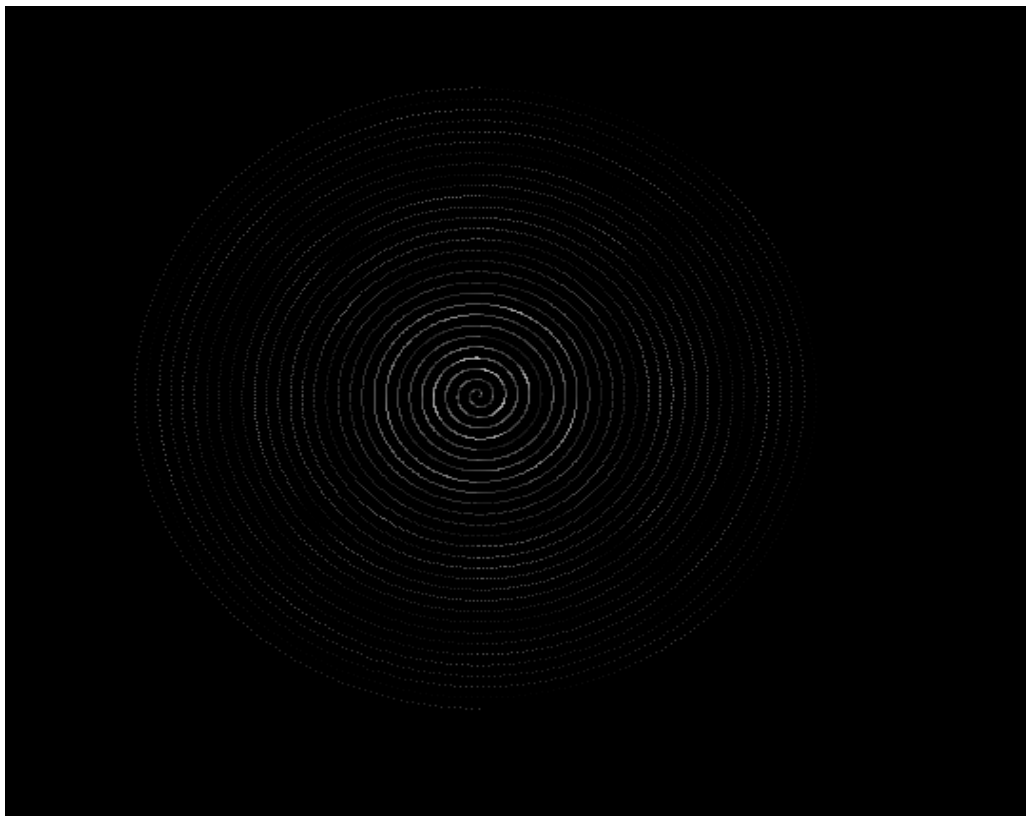


Figure 288: Example of a thin FMM spiral

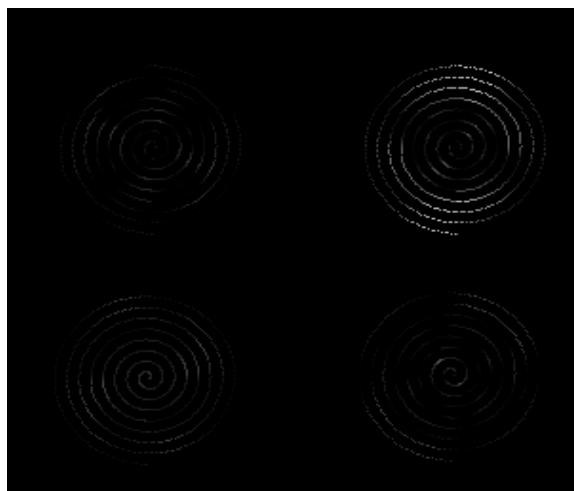


Figure 289: Four small examples of distances spiral in isolation

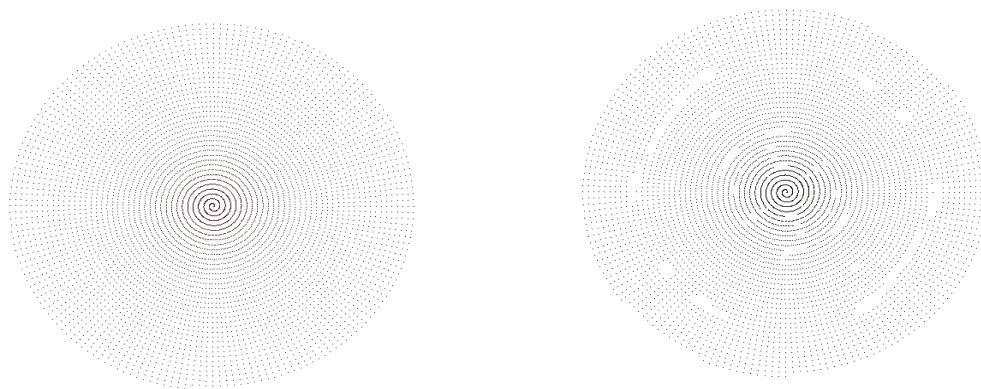


Figure 290: 2 larger examples of distances derived from pairs of images of the same people



Figure 291: The distances spiral overlaid on the images it corresponds to (9th person in the set)

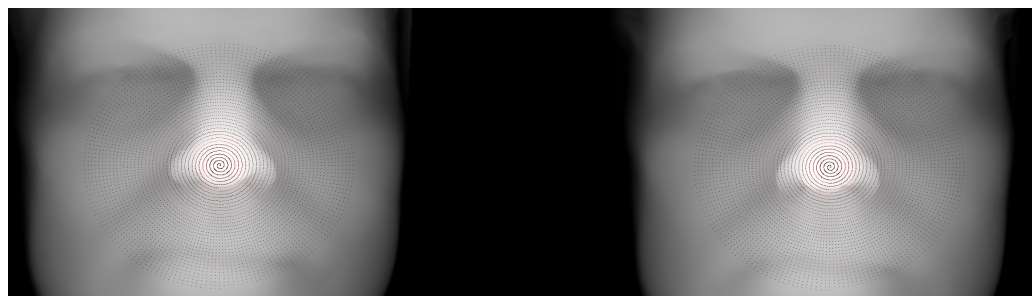


Figure 292: The distances spiral overlaid on the images it corresponds to (13th person in the set)

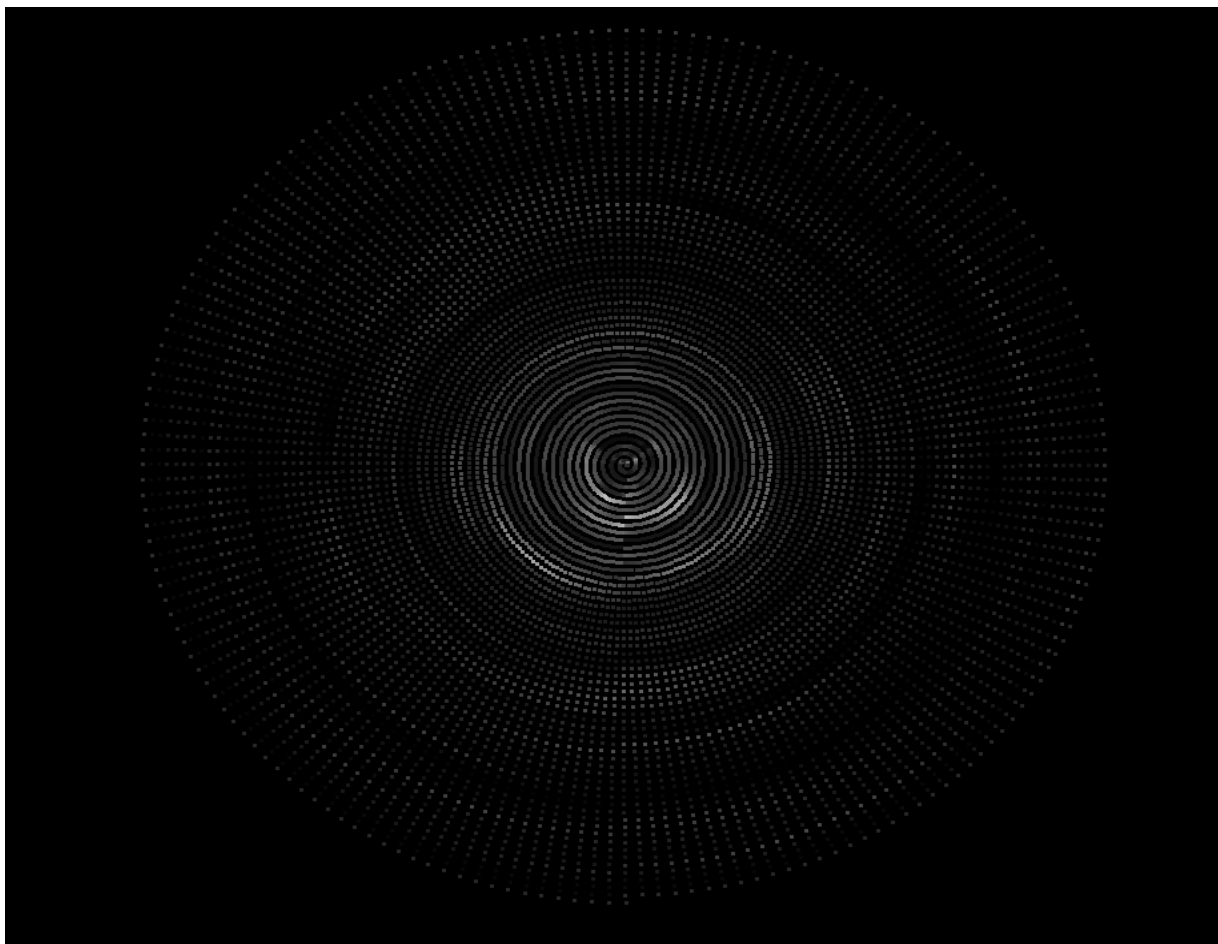


Figure 293: The distances spiral with larger, clearer points

7.12.32 Weighting for Source Points

It has become clearer where the similarity measure is picking up penalties that should be avoided for optimal performance. If the objective function was changed to apply weighting to the different points, then a lot of problems would be averted. For instance, earlier systematic experiments showed that there is more discriminative power around the nose, yet with two eyes equally accounted for, only a third (or less) of the overall similarity measure is based on outward geodesic dilation from the centre of the nose.



Figure 294: Example of a true pair (same person) with a simplified representation of distances around each source point (FMM)



Figure 295: Another example of a true pair with simplified representation of distances around each source point (FMM)

7.12.33 Weighted Similarity Measure

There is an experiment underway where weighting is applied based on the lessons taken from an analysis with the new tool. So far there are no classification mistakes, but it ought to run further and compare more pairs before meaningful conclusions can be drawn and the weighing tested to a point where ideal values are found.

7.12.34 Early Results With Weighting

Taking the first imaged person (58 images) and enrolling it for an experiment (versus 92 random people), the ROC curve shows good performance. Some other people remain more problematic because they deliberately change expressions and complicate things around the eyes (further smoothing might be needed there), which leads to greater necessity for cunning weighting schemes. This will require some further exploration to work around.

With borderline cases removed (detected and rejected), the results are without an error in detection. Without automated filtering of borderline cases (declination to classify), the ROC curve now looks as shown in Figure 301.

How many subjects get included may be important here. How many instances is important as well. The number of subjects is 82 for the false pairs and 5 for true pairs. I now work on expanding the latter.

Upon closer inspection of the problematic cases, improved initial alignment would help eliminate some of the borderline cases. Additional experiments

have been run where increased emphasis is put in the nose for alignment, as the nose region is also weighted most heavily by the similarity measure.

7.12.35 Nose Tip Revisited

One of the remaining issues – one that leads to errors where alignment leaves much to be desired – was studied more closely with aims of overcoming problematic cases. The harder images were enrolled into a gallery, which then had applied to them the FMM-based algorithm with alignment redone. Basically, one of the challenges is that the nose tip, for instance, may be seen as existing at one of 10-50 different spots, as all share the same Z value (discrete and within range which is taking an integer value). To demonstrate this, showing a region of equivalence helps. There is basically a flat quantised surface where the choice of point can determine the accuracy of the FMM-based method. In the experiments, two alternative methods were studied. One looks for the first point which is closest to the camera and another averages the location of all such points and therefore takes the point roughly at the centre. In terms of performance, given the same difficult set, things did not improve much. The performance in both case is comparable, so other methods will be studied.

7.12.36 Similarity Measure Variants

Several more approaches – variants of the original algorithm – were tested and later aborted as they did not show improvements. It does not seem as though realignments help in all cases because although they can resolve some problematic cases, they also ruin perfectly fine ones at times. What complicates things is the duration required to test variants, as even by running false pairs on one 8-core computational server and true pairs on another would typically take about an hour (for useful conclusions to be drawn).

We are thinking of testing a couple of new measures for alignment, among which are scale invariant and equi-affine invariant distances on surfaces. It could be interesting to see if those could boost somewhat the alignment process.

7.12.37 ROC Curve - Without Smoothing

Upon curiosity to discover what disablement of smoothing would do to performance, it turned out – at least based on the first person versus all of the rest – that this does not have a noticeable effect on performance. This can perhaps be explained by the sparse triangulation and the "first order" geodesic measure which determines the surface boundaries.

7.12.38 Spectral Masks

There was a one-week period when several ideas were tested, including a mask that is spectral rather than geodesic. It was tested for potential improvements. Not much documentation was produced for it, due to technical limitations associated with distance from my main workstation.

We ended up spending a day building, testing, and running experiments with a spectral similarity measure for faces, but could not guarantee exciting results. The experiments from that day (scripted run over the weekend) indicated that there was merit in the approach tested, but in order for the ROC curve to look decent, a lot more work will be required.

7.13 Diffusion Distance

Spectral masks of the types used so far suffer from a particular weakness that was not foreseen and can possibly be overcome by adjusting some parameters. The short explanation of the weakness is that distances do not increase or energy degraded quite so linearly, meaning that a move along distances in search of a geometrically useful cut will either strip out too much or too little. It makes the mechanism impractical for the recognition purposes at hand. What probably remains needed is a better understanding of which parameters to change and how. We looked at 4 papers related to this, but none of these covered specifically the problem at hand.

After an exchange of code and documentation it was possible to produce some results that look like an improvement.

7.13.1 Spectral Rings

Experiments are being run to study the potential of multiple spectral 'rings' around which distances get measured for the intrinsic comparison of faces.

7.13.2 Values Reset

The tricky bit in the experiments (so far) is finding the range of valid signatures that do represent a useful comparator. Some long experiments have been run which combine useful ones with highly noisy ones, leading to poor results which necessitate some redesign and learning of parameter value to set next time. Unlike geodesic distances, herein all distances are highly sensitive and need to be properly adjusted.

7.13.3 Diffusion in Facial Features

With a single diffusion-based ring around the nose we can get a recognition rate of about 80%. This improves considerably when more rings and reference points are added, so the results so far should be treated as proof of concept or exploratory at best.

The figures show a similar approach with geodesic rings, which gave recognition rate of more than 95%. The challenging thing is adapting diffusion

masks to the task at hand. We are running experiments in order to learn which parameter values work best.

According to Wikipedia on [Diffusion wavelets](#), "[d]iffusion wavelets are a fast multiscale framework for the analysis of functions on discrete (or discretized continuous) structures like graphs, manifolds, and point clouds in Euclidean space. Diffusion wavelets are an extension of classical wavelet theory from harmonic analysis. Unlike classical wavelets whose basis functions are predetermined, diffusion wavelets are adapted to the geometry of a given diffusion operator T (e.g., a heat kernel or a random walk). Moreover, the diffusion wavelet basis functions are constructed by dilation using the dyadic powers (powers of two) of T . These dyadic powers of T diffusion over the space and propagate local relationships in the function throughout the space until they become global. And if the rank of higher powers of T decrease (i.e., its spectrum decays), then these higher powers become compressible. From these decaying dyadic powers of T comes a chain of decreasing subspaces. These subspaces are the scaling function approximation subspaces, and the differences in the subspace chain are the wavelet subspaces."

If we are using diffusion geometry, we should also compare descriptors and not only distances. A closer look at the code will hopefully make it clear what comprises the diffusion wavelet basis functions (or equivalent/s). So far it has been used through the interfaces merely for masking purposes.

We spent a couple of days looking at descriptors and how they can be used

to efficiently distinguish between surfaces representing faces. A dense, brute-force operation did not work well, so the path explored at the moment looks at taking just few interesting features like eyes and nose tips, then measuring the spectral distance between those. The ROC curve shows the result of a crude comparison – measuring the distance between two points only. As more such distances are aggregated performance (accuracy) ought to improve.

Taking two spectral distances between landmark points yields a similar performance, so a denser distances map will be implemented.

The number of anatomically meaningful points that can be reliably extracted from a 3-D image is limited, so even by using the spectral distance between all of those to measure intrinsic differences does not make up a powerful enough discriminant. In essence, more experiments were run where differences in spectral distances – to to speak – were measured, raised to the power of two and aggregated (summation) to give a figure of merit. Getting recognition rates at the rate of 90% or above is still extremely hard, no matter the adjustments made.

Arbitrarily aligning and drawing analogous points from a grid would not work well both for practical and theoretic limitations, such as the fact that we are not guaranteed to measure the same anatomical points while moving from one image to another. With fiducial it might be another matter altogether.

A different approach is going to be explored rather than time being spent under the premise that subtitling geodesics with wave or heat kernels will,

on its own, improve the results considerably. Exact geodesics seemed like a theoretically sounds substitution, but the existing implementation of them cannot be trivially run on the computational servers.

Many measurements on the surface do work, but they are not always accurate enough and robust enough to anatomical variation. More points could be obtained by running common edge detection operators on the photometric part, but then it becomes a 2D+3D problem.

7.13.4 Similar Work

The [Matlab toolbox for fast marching](#) does something relevant, but nothing that involves diffusion. Having browsed several recent papers that adopt an approach similar to ours, I found one paper from 5 years ago [1] where the idea was similar and the results inferior to ours. In other papers, Elad and Kimmel incidentally get cited, but there are no results, just analytical writing [57].

We spent a few hours browsing through anything which overlaps our lines of research. Along the way I also found and read/skimmed [56]. Its abstract says: "The performance of automatic 3-D face recognition can be significantly improved by coping with the nonrigidity of the facial surface. In this paper, we propose a geodesic polar parameterization of the face surface. With this parameterization, the intrinsic surface attributes do not change under isometric deformations and, therefore, the proposed representation is

appropriate for expression-invariant 3-D face recognition. We also consider the special case of an open mouth that violates the isometry assumption and propose a modified geodesic polar parameterization that also leads to invariant representation. Based on this parameterization, 3-D face recognition is reduced to the classification of expression-compensated 2-D images that can be classified with state-of-the-art algorithms. Experimental results verify theoretical assumptions and demonstrate the benefits of the geodesic polar parameterization on 3-D face recognition."

Also of relevance we have [94, 84, 58]. The latter says that "[f]ace recognition based on spatial features has been widely used for personal identity verification for security-related applications. Recently, near-infrared spectral reflectance properties of local facial regions have been shown to be sufficient discriminants for accurate face recognition. In this paper, we compare the performance of the spectral method with face recognition using the eigenface method on single-band images extracted from the same hyperspectral image set. We also consider methods that use multiple original and PCA-transformed bands. Lastly, an innovative spectral eigenface method which uses both spatial and spectral features is proposed to improve the quality of the spectral features and to reduce the expense of the computation. The algorithms are compared using a consistent framework."

Those last two are not so relevant, but they consider an approach other than geodesic metrics (ish).

7.13.5 Gabor Filtering In Combination With Diffusion Distance

Mauro explains that "Diffusion geometries refer to the large-scale geometry of a manifold or a graph representing a data set, which is determined by long-time heat flows on the manifold/graph/data set." Other consider diffusion in another context.

From the abstract of [93]: "In this paper, by incorporating spatially structured features into a histogram-based face-recognition framework, we intend to pursue consistent performance of face recognition. In our proposed approach, while diffusion distance is computed over a pair of human face images, the shape descriptions of these images are built using Gabor filters that consist of a number of scales and levels. It demonstrates that the use of perceptual features by Gabor filtering in combination with diffusion distance enables the system performance to be significantly improved, compared to several classical algorithms. The oriented Gabor filters lead to discriminative image representations that are then used to classify human faces in the database."

Gabor filter are also used by some leading algorithms for face recognition, so this approach might be worth exploring.

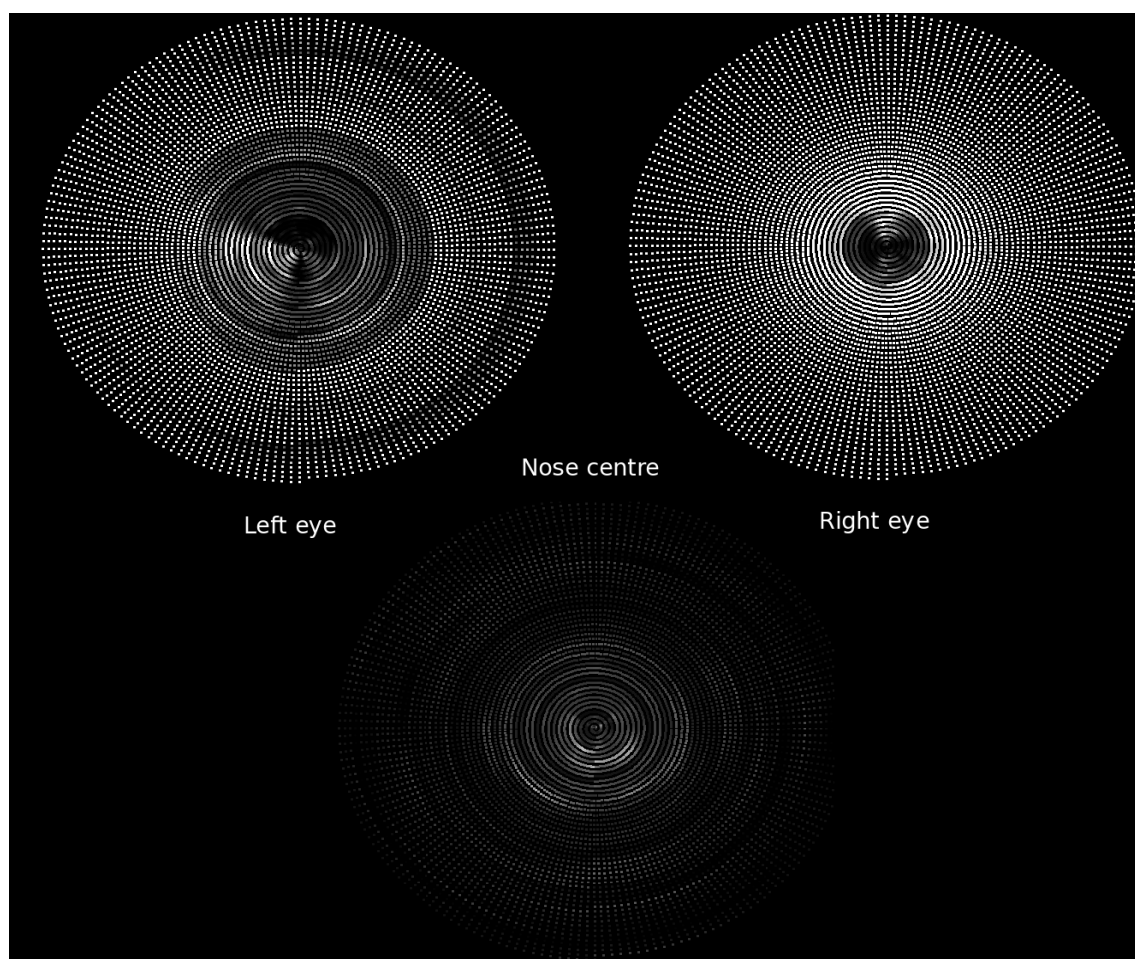


Figure 296: An expanded view on the mis-correspondence between regions, where brighter shades represent greater disagreement between the pair taken from the same person

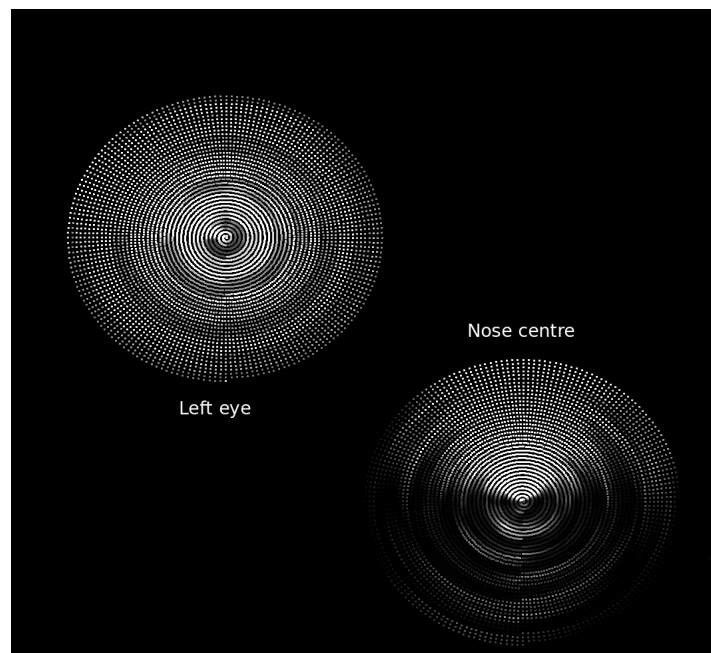


Figure 297: An expanded view on the mis-correspondence between regions, where the pair taken is from different people

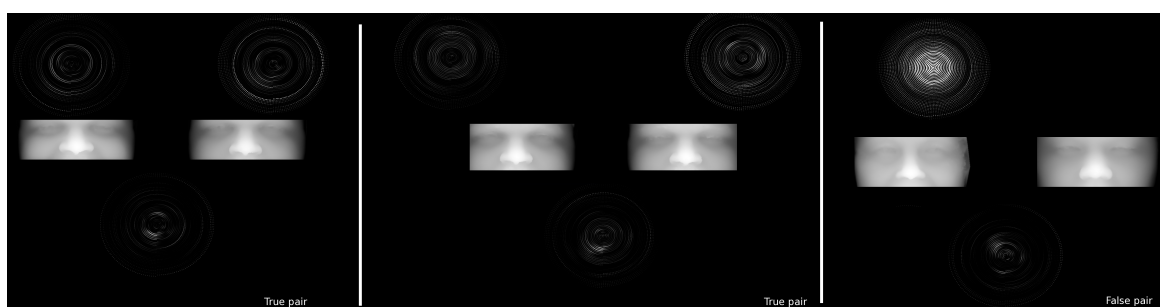


Figure 298: Overview of the debugging process with examples from two true pairs (same person) and one false pair (different people), with the eye component discrepancies shown at the top and the nose at the bottom

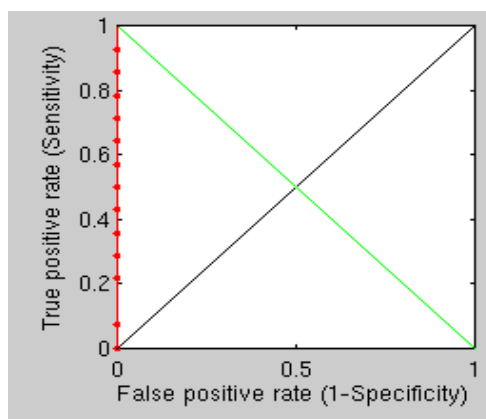


Figure 299: The ROC curve obtained by using a weighted form of the similarity measure

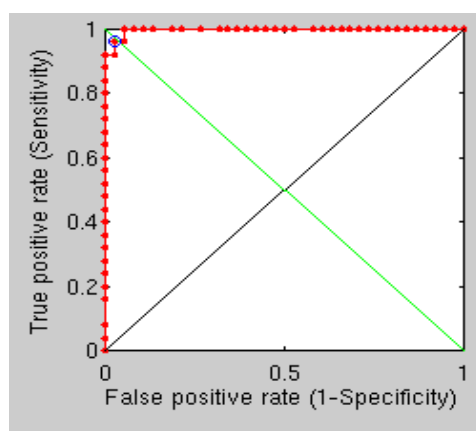


Figure 300: ROC curve for the first phase of the experiment, which compares one-to-one (same person) and many-to-many (different people excluding this person, except in one case)

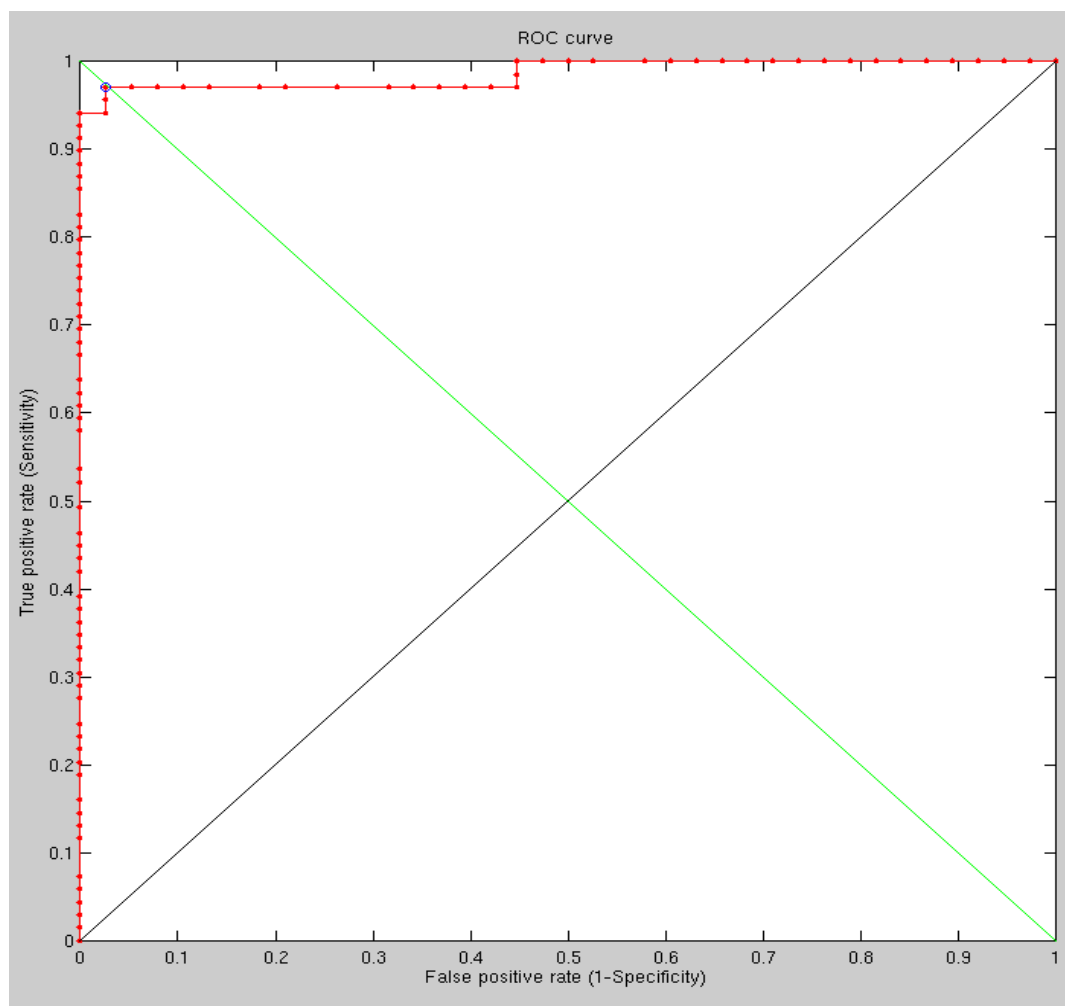


Figure 301: A broader scope curve for performance as in the previous figure

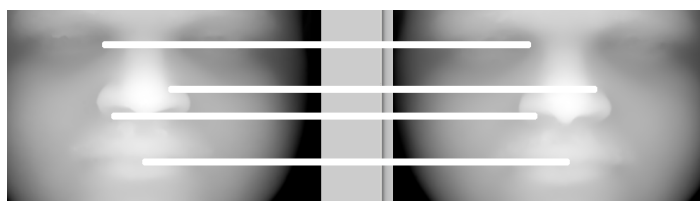


Figure 302: An example of misalignment in some parts of the nose in a true pair of images (same person), with the left nostril being a prime example

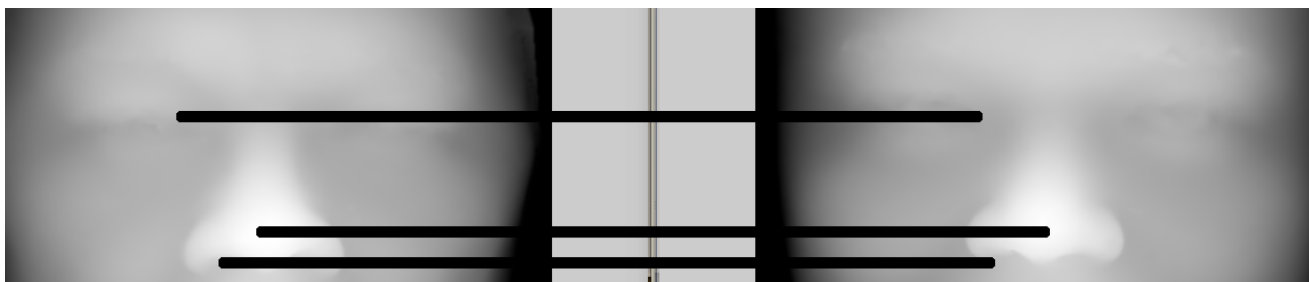


Figure 303: An example of a borderline case (leaning towards false positive)

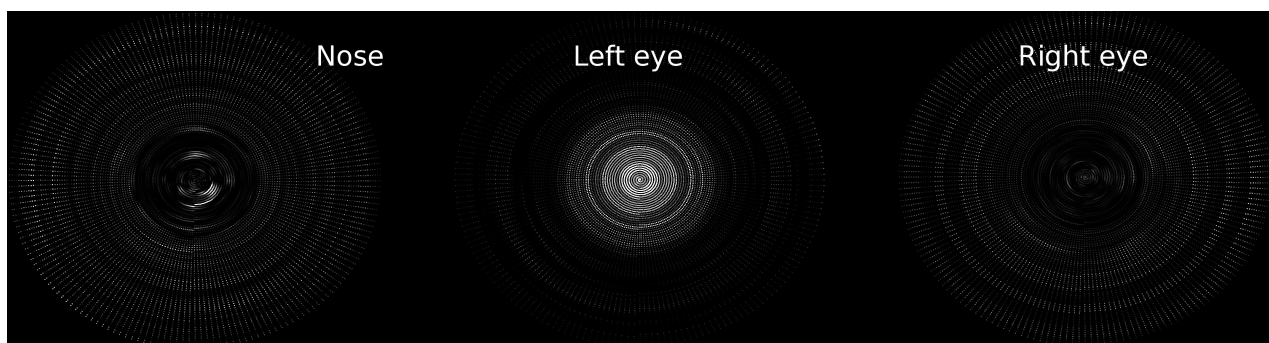


Figure 304: Debugging information for the problematic true pair shown before

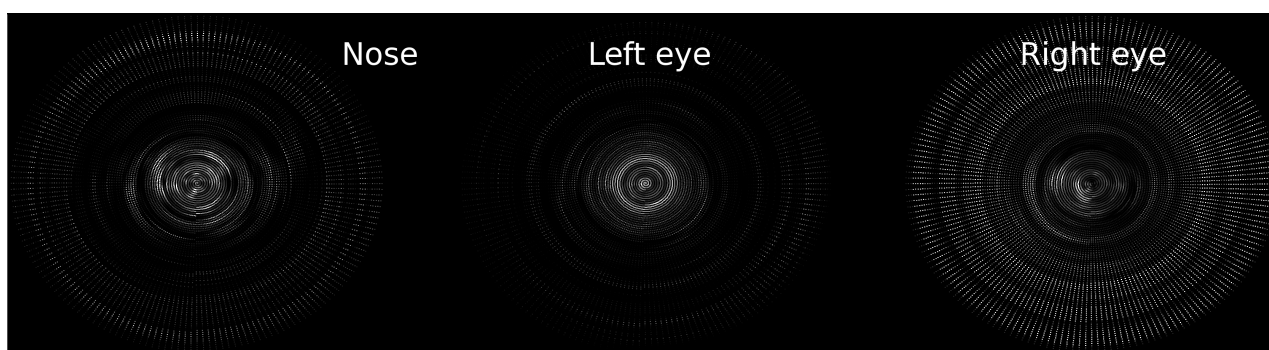


Figure 305: Debugging information (distance differences) for the aforementioned false positive

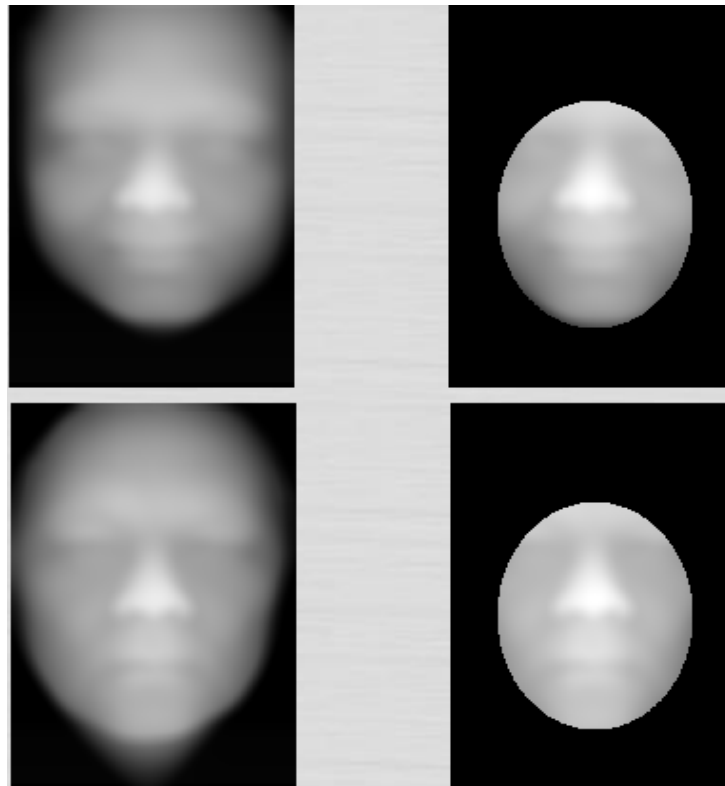


Figure 306: The problematic (borderline) false positive after the new alignment scheme gets applied

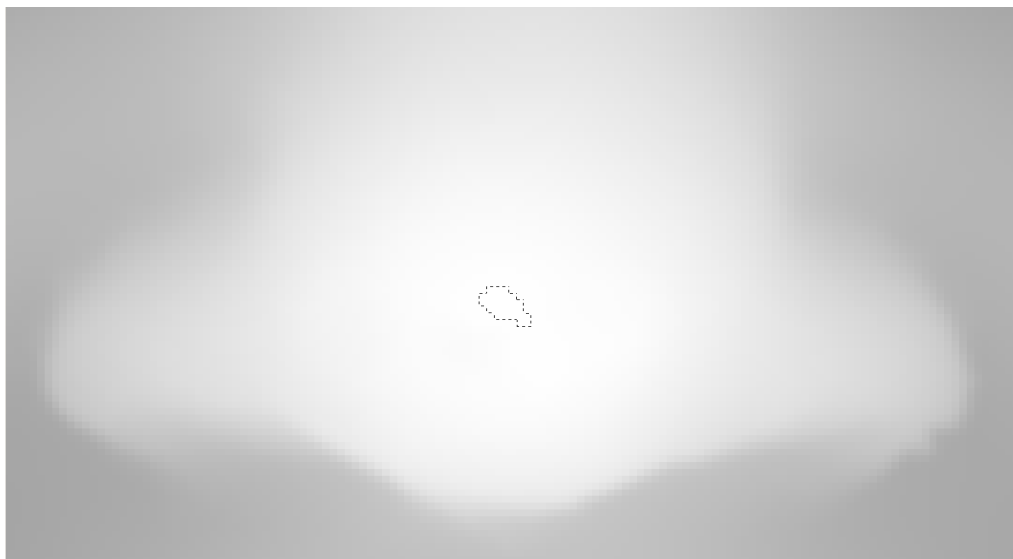


Figure 307: A contour around nose tip candidates all of which share the same (maximal) depth value, resulting in uncertainty

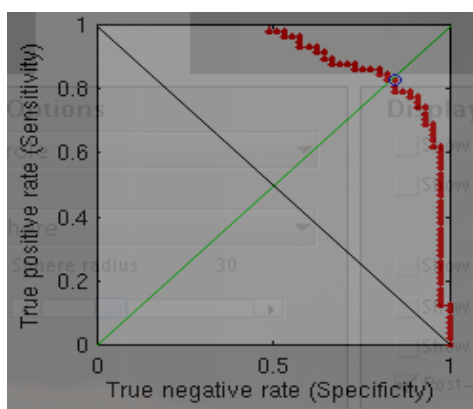


Figure 308: The Performance attained in hard cases where the tip is determined more arbitrarily than in a sophisticated fashion

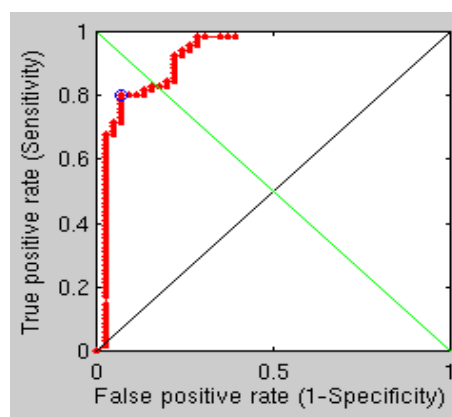


Figure 309: The Performance attained in hard cases where the tip is chosen based on the average location of tip candidates

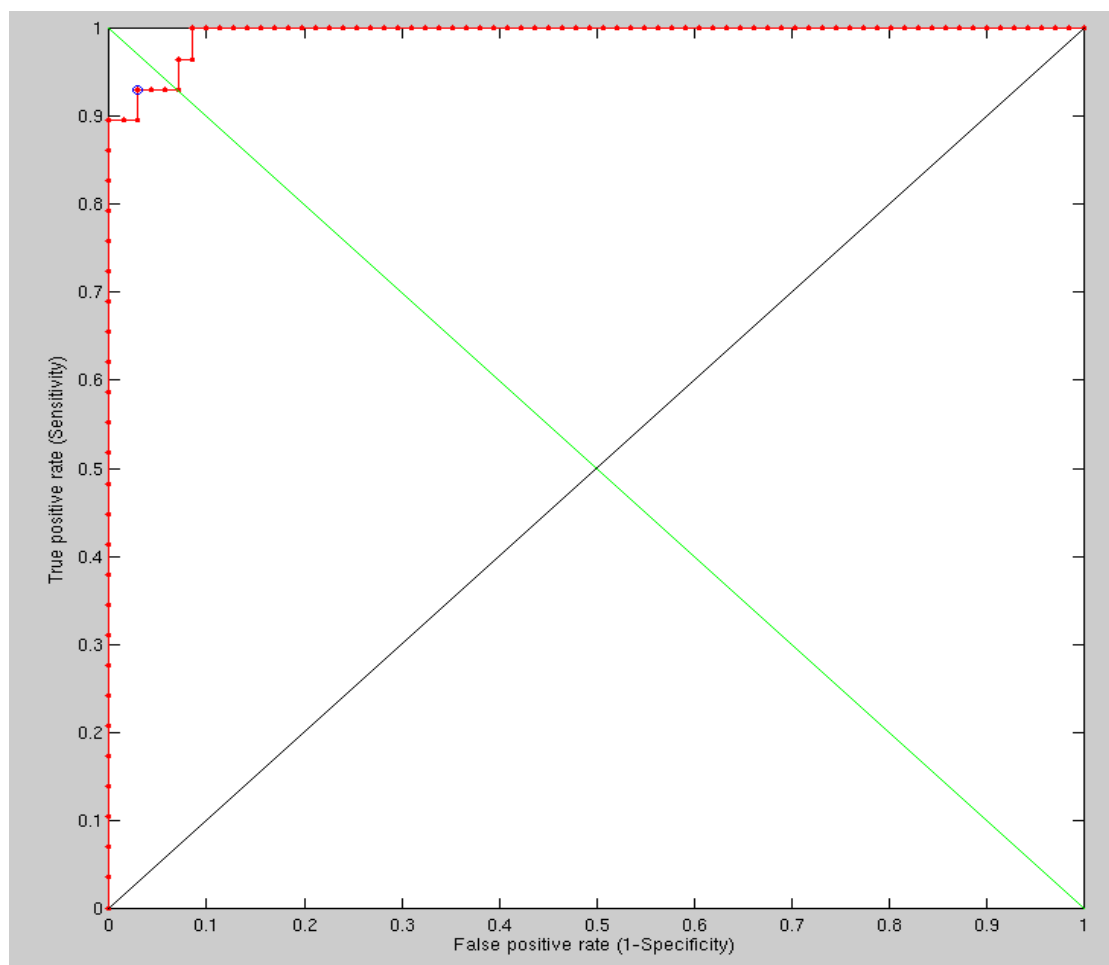


Figure 310: The result of applying a faster calculation of similarity, as applied to the first person against 90 different pairs from 90 different people

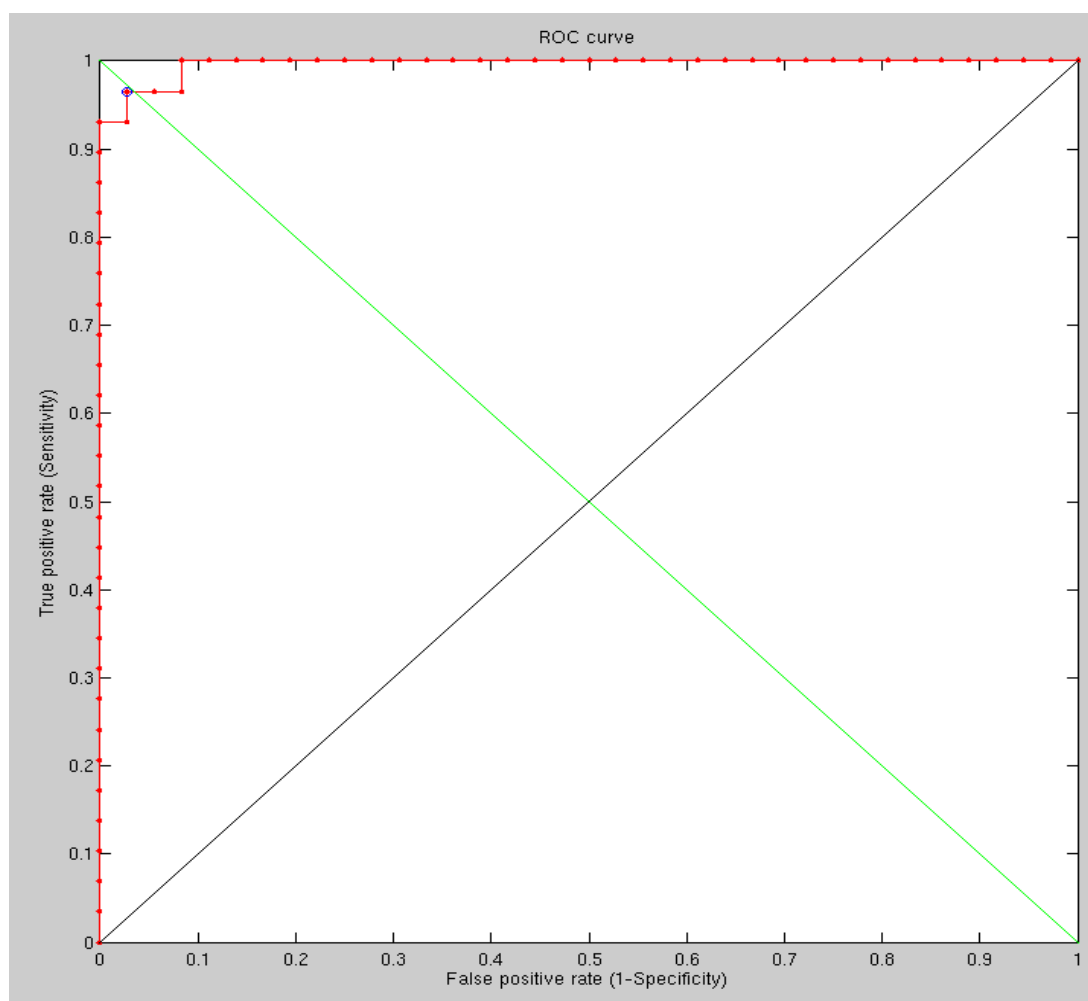


Figure 311: Performance when smoothing gets disabled, demonstrating little difference compared to prior results

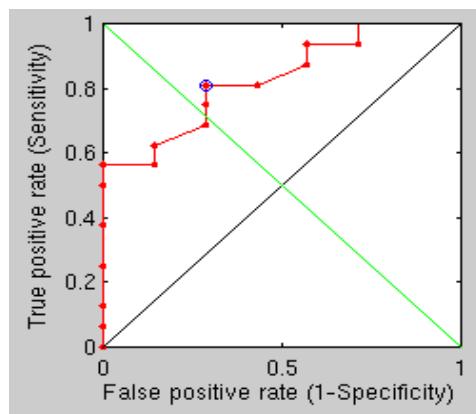


Figure 312: The sort of results we get by using spectral masks without proper adjustment to make the masks shrewd enough. There is some potential there.

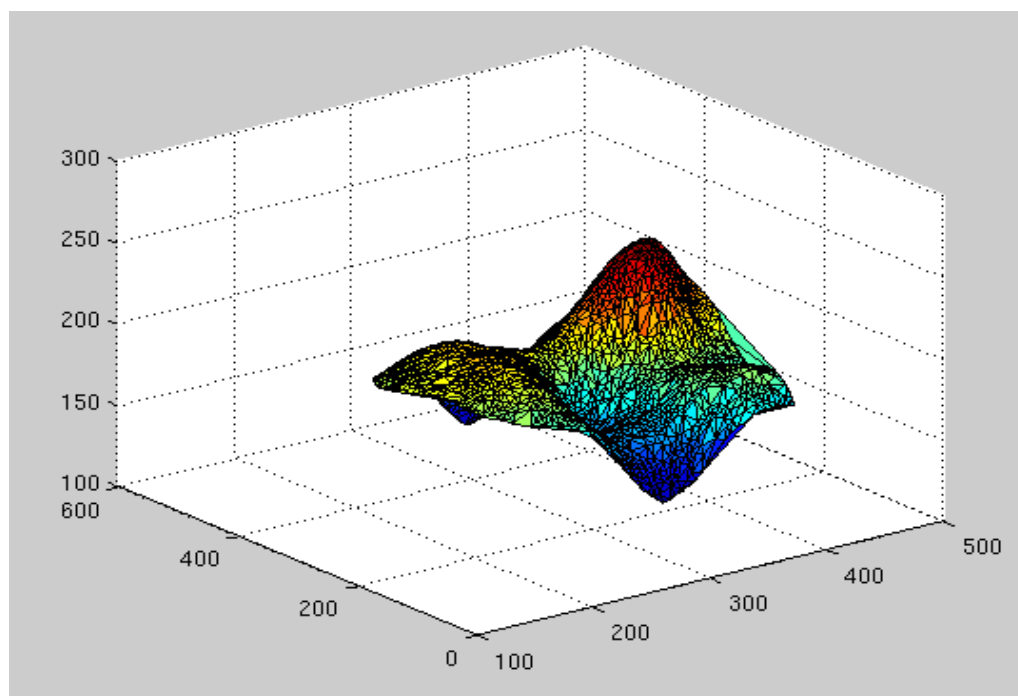


Figure 313: Example raw slice of the face of one subject

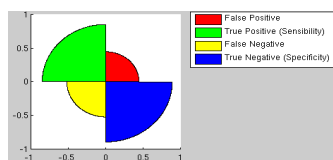


Figure 314: A test run with just one ring around the nose as a discriminant

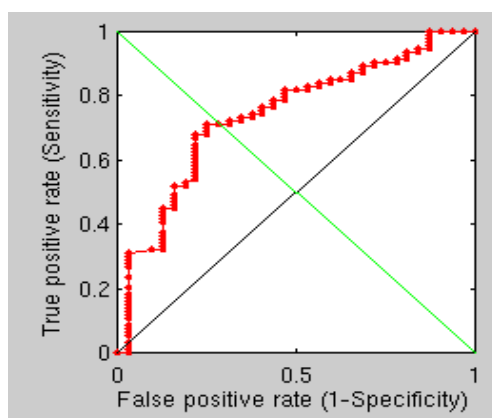


Figure 315: The ROC curve obtained by using one single spectral/diffusion ring

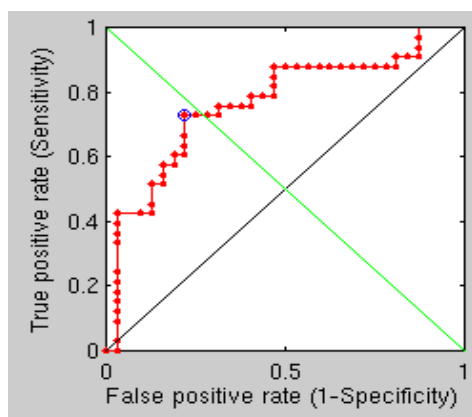


Figure 316: The ROC curve obtained by using one single spectral/diffusion ring, applied only to the true positive gallery in the set

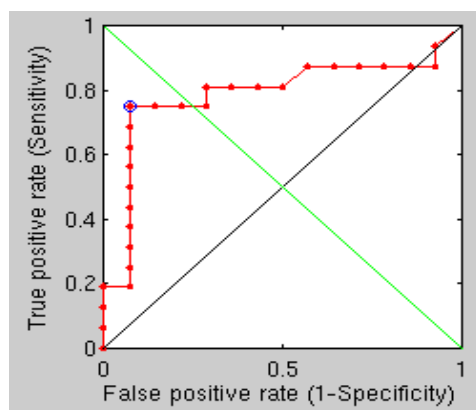


Figure 317: The ROC curve obtained by using just one diffusion distance as a discriminant

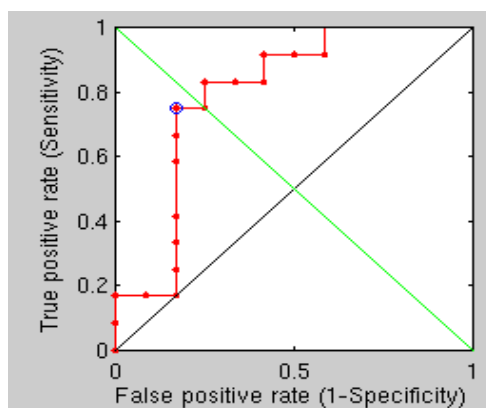


Figure 318: The ROC curve obtained by using two diffusion distances as discriminants

7.13.6 Mask Dilation Approach

Several attempts have been made (however unsuccessful) to follow an approach of triangle/polygon counting with diffusion distance-based mask dilation, where masks expand and add portions of the faces based on diffusion geometry. Rather than use Euclidean measures, one can just add up the differences, but as a discriminant it fails to work too well. Another approach will be explored instead, with the aim of using diffusion geometry for face recognition (so far the best we got is about 80%).

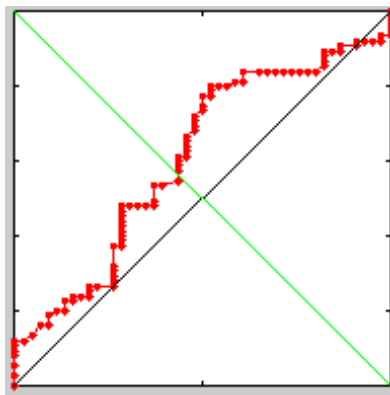


Figure 319: ROC curve for a rather disappointing approach of mask dilation based on diffusion distance

7.13.7 Dilation Range

Extending the range of dilation has not helped in getting better results, despite the fact that conceptually the approach made some sense and surely would have worked to some degree for geodesic distances (although this has

not been verified empirically). Surely there should be some way to exploit the descriptors for the sake of face-to-face comparison, but maybe the similarity in topology of faces weakens what is being measured (mainly around the nose tip). The two different approach which were tested could not exceed 80% recognition rate.

We will therefore try dilation at areas further away from the nose.

Expansion of the dilation range well beyond the nose (and accounting for the entire face in comparing images) leads to performance that is only slightly better than a random classifier. The problem as a whole is, we do not have the ground-truth correspondence in images, and being able to compare one image to another at the level of Eigen-decomposition requires this correspondence. Without some edge detection (photometric), it would be hard to obtain. As a geodesic distance alternative for Riemannian surfaces, this new method generally failed to work well enough in just about any experiment that had been set up (there were many). Maybe building a framework around a new Hausdorff distance-based measure will yield something that can usefully be applied to face recognition. Another option would be to revert back to geodesic distances that gave us the best performance so far (with FMM, not exact geodesics). A comprehensive search on the Web does not reveal many surface distance implementations written in Matlab syntax (ones we have not explored already).

While working on 3-D data in isolation we still require at the very least some

point of correspondence such as the nose tip. ICP can help get an approximation of other correspondences, but these are never accurate enough. Perhaps there are other known methods of accurately finding more correspondences (from 3-D only), so I will look for some. With more correspondences, far better performance can be assured. Faces would not just 'float' around a single point.

7.13.8 Automatic 3=D Correspondence Finding

For automated correspondence across images I have looked at David Lowe's SIFT (for 2-D) and a Bookstein-inspired approach to finding correspondence in 3-D, which was implemented and published as code one month ago. A quick look at the code reveals a fairly low quality and there is also a requirement that normals for each point get supplied.

There is an interesting recent paper on finding 3-D corresponded at

<http://www.waset.org/journals/ijeee/v3/v3-3-21.pdf>

Kaiser *et al.* combined this with texture [39].

“Ten people who speak make more noise than ten thousand who are silent.”

– *Napoleon Bonaparte.*

8 Summary and Conclusions

FOLLOWING a systematic process which looks at two analogous approaches was originally sought for the studying of PCA and GMDS, but we moved on well beyond that. This section embodies some results from early work.

The aim is to obtain the same ROC curves as reported in the IJCV paper so that we have a baseline for comparison. Once this goal is obtained, the next goal would be replacing the ICP/PCA with GMDS, Or maybe even refine the alignment with GMDS and hope for better recognition rates. When it is time to run GMDS there ought to be efficient implementation to work with.

In order to get the required results it is worth outlining exactly what is missing before getting the ROC curves of Mian *et al.* As the implementation stands at this stage, it sometimes misidentifies the nose tip, which is very unhelpful if it then feeds PCA (and obviously pollutes the signal). We could tackle this maybe by selecting a large subsets of images which we know can be handled in an acceptable fashion; otherwise it's back to going around in circles trying to tweak for particular cases and then botching the others. This

is the most frustrating part of this project, and clearly it became a distraction because it's the part which is not novel and we mustn't care about all that much. This took more time than the whole GUI.

As filling holes and smoothing is a source of difficulty we have reused some simple denoising/hole filling filter.

Regarding nose identification, if we take a generic nose (template) and try to ICP it to the estimated location, then there is room for improvement. That this is more or less what we were trying to achieve, but if we take a couple of such template noses, then there is also improved robustness, so that is definitely an idea worth implementing. Provided that one can remove all the points associated with the background, the cloudpoint we try to do fitting with should suit the template's surface (or one of several candidate surfaces corresponding to templates). In fact, we could take such a generic nose with high resolution (cloud of points) and lower resolution as we depart from a generic mask. Multi-scale and search window might be needed because in the database provided by the Grand Challenge project coordinates there are some odd cases where the imaged person is somewhere at the side and very much away from the aperture. If it is intentional, it begs to test one's ability to locate faces, not just recognise them.

Alternatively, we could train a Viola-Jones like detector for the tip of the nose (with enough support) that would work on the shading image created by, for example, the shading image. It would then serve as initial conditions

for the above ICP trick. We ended up implementing new means of reliably finding the nose in grainy images. ICP should be resistant to localised noise and phantoms that may resemble a nose, assuming that the background can be removed in a consistent fashion. This in its own right is an interesting approach. The accompanying image shows the GUI with the axes displaying post-filtering data [320](#)(can also show it as a surface given the tickboxes). Aligning video frames is a separate challenge that needs tackling.

ICP-based nose tip detection is implemented with various options, but the results thus far are unsatisfactory because the suggested alignments are incorrect. If smaller regions are chosen for the template (avoiding the mandibles and choosing just the nose region), more or less the same type of results are arrived at. [Viola's method](#) seems interesting but unavailable^{[21](#)}, so for the time being, further refining an ICP fit is worth exploring. In addition, I mailed Yaron to ask about frame calibration.

While digesting and processing the frames of the video sequences one at the time, noise gets removed using some filters and the face can then be seen more clearly. There is additional difficulty, however, especially when it comes to dealing with the lack of frame calibration, meaning that when a frame other than the first is opened there is basically a dividing area between two frames (or complementary parts of the same frame), so additional code which gives one clear picture is required. It was worthwhile checking if someone already

²¹ Although implemented in [OpenCV](#) which means we may have to code it from scratch, at least if moving in this direction. Upon closer inspection, there is an implementation we can reuse over at [MATLAB Central](#).

written such code, but it looks like an offset, so in subsequent we have an offset in writing to the buffer. If we try to read each frame separately and display it in MATLAB, there is still an odd effect. By using a function call like `readGipFile(expressions_fileName, image_n)` this behaviour is reproducible with, e.g.:

```
displayRaw_SingleFromVideo
```

taking paramaters:

```
fileName = '~/Facial-Expressions-Recognition/Smile.v3r'; frameToView
= 8;
```

There is a thick line in the middle of the face where cloudpoints do not appear at all²².

As for frame being less than calibrated, it turns out that the file loaders need some special offset for any frame other than the first in the sequence. Once issues such as this are out of the way, it should be possible to handle the interesting parts of the experiments. This data type in general, unlike the typical one that is easy to deal with [321](#), is very noisy and the face not so trivial to identify with great certainty [322](#).

Here is an image showing how we presently handle GC faces... almost always well enough. GIP data (typo in filename) is more complicated. We corre-

²²Among the available functions there are `displayFiltered_Single.m`, `displayFiltered_SingleFromVideo.m`, `displayRaw_Single.m`, `displayRaw_SingleFromVideo.m`, `readFrameProperties.m`, `readGipFile.m`, and `readGipFileHeader.m`.

spond about that. Having tried ICP and Viola's (*et al.*) method, there was not much progress because the ICP one needs decent initialisation and the latter just finds lots of faces all over the place, or none, depending on how it is used. We can try to train it on noses rather than whole faces (which in turn give good estimates of nose location using simple measurements). The problem with that is, the cruft all over the image space – and especially the sides – sometimes resembled small noses. Suffice to say, looking at the images, getting an initial estimate of where the nose typically is, then initialising there would be easy, but it would not generalise to other datasets; ad hoc methods are assumed to be another realm altogether – one where we use tricks to get things working on particular datasets (convenience/pragmatism) rather than develop robust computational methods (principled approach). It's tempting to just embrace the former approach, at least for now. Explaining and reasoning along the lines of, "we look for a nose somewhere in the middle" is just not compelling enough (sloppy even), not as much as leveraging of Paul Viola's recent work. These issues something crop up in peer reviews, so sometimes it's better off done right in the first place and not later on, some time in the foreseeable future when exceptions creep in. The same goes for hacking-like development where the code works but becomes unmaintainable and unusable to anyone but its author/s. People typically learn this the hard way, through arduous experiences and frustration. The GUI now has ICP and Viola-Jones as possible methods in the drop-down menu. We shall see what else we can find/do...

The program is centimeter-aware and also pixel-aware, so different measures are taken into consideration, e.g. when looping through pixels/cloudpoints and when performing a physical measurement (nose to forehead for instance).

A fifth nose-finding method is now implemented and it takes a small range within which to find objects resembling a nose (using the method from Mian's group, but limiting the search window). Dealing with frame offsets is another matter and working around it not at API level would be unwise. We lack an expressions-neutral GIP dataset, too. Meanwhile we press on with ICP/PCA code.

Here are some more images [323](#) which are examples of our 3-D registration, based on the matching of two cropped faces that are centred wrt the same axes. Rather than display just the cropped cloudpoints after registration, shown here are the difference images of the entire face surface before and after registration. This is just a sample of several such images which are generated with robust cropping having been applied (all hits, no misses) to mostly expression-neutral scans selected at random.

The plan was, at this stage, to see some ROC curves. We will get to that later.

OK, I will start running large experiments, but one might warn in advance that the expressions dataset has no neutrals in it. I will make something rudimentary to serve as a baseline and notes will be expanded to keep track of progress.

An implementation of offset correction will be required for GIP data (See (Figure 325). Otherwise, as the images show (Figure 324 and Figure 326), the face images which contain a lot of noise and move up/down depending on the frame, leading to mis-location of the nose tip marker. A way of visualising the shape residuals is now implemented too.

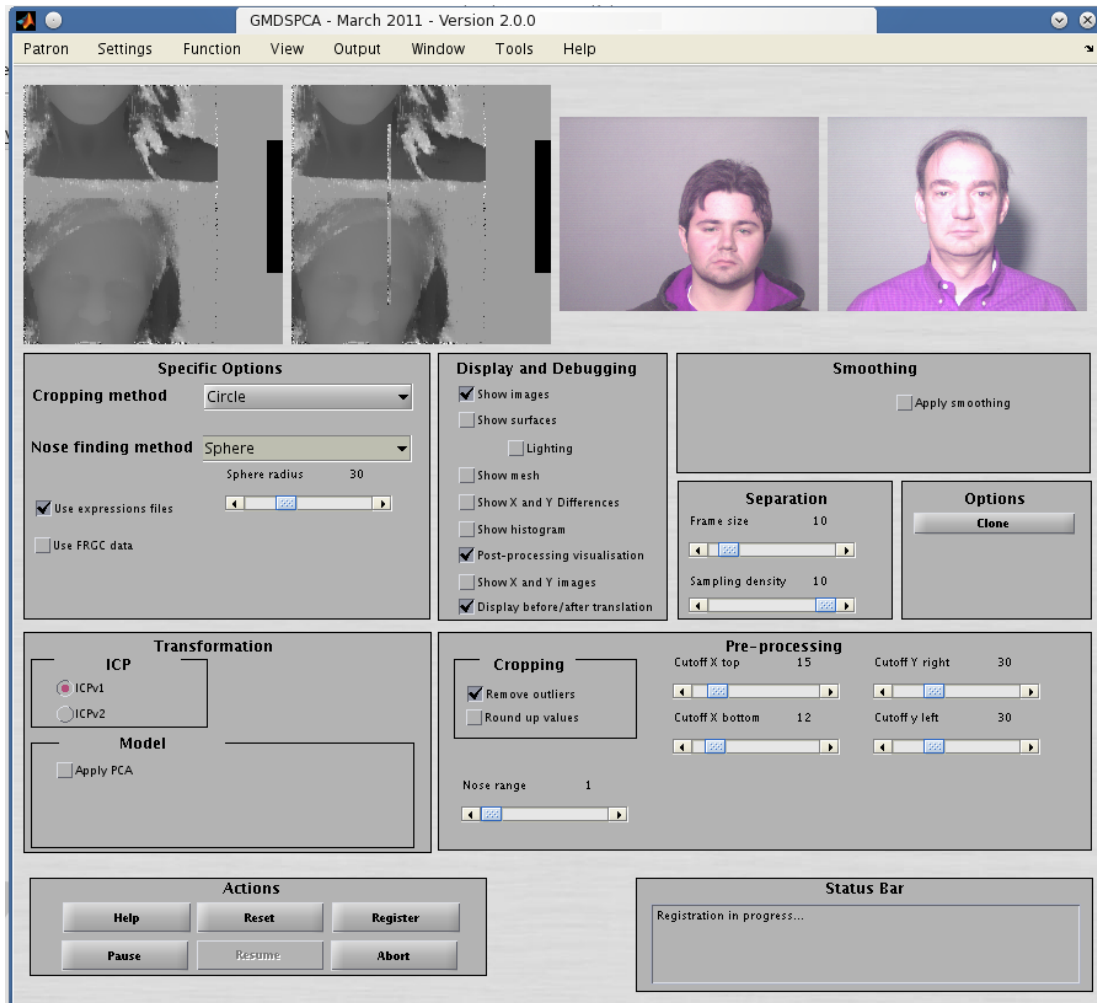


Figure 320: The same GUI in late March



Figure 321: Face cropping for standard experiments data

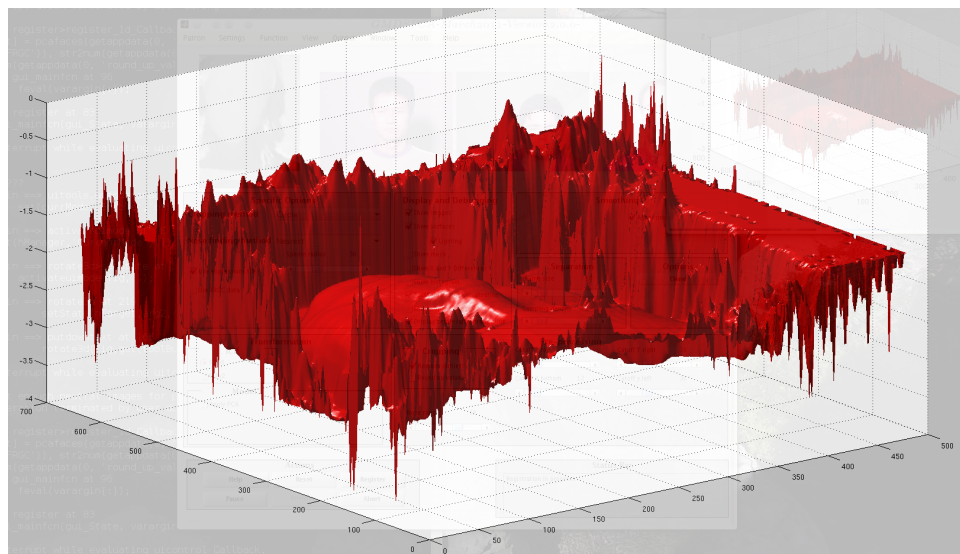


Figure 322: Difficulties identifying faces in GIP data



Figure 323: Difference images of the entire face surface before and after ICP-based registration

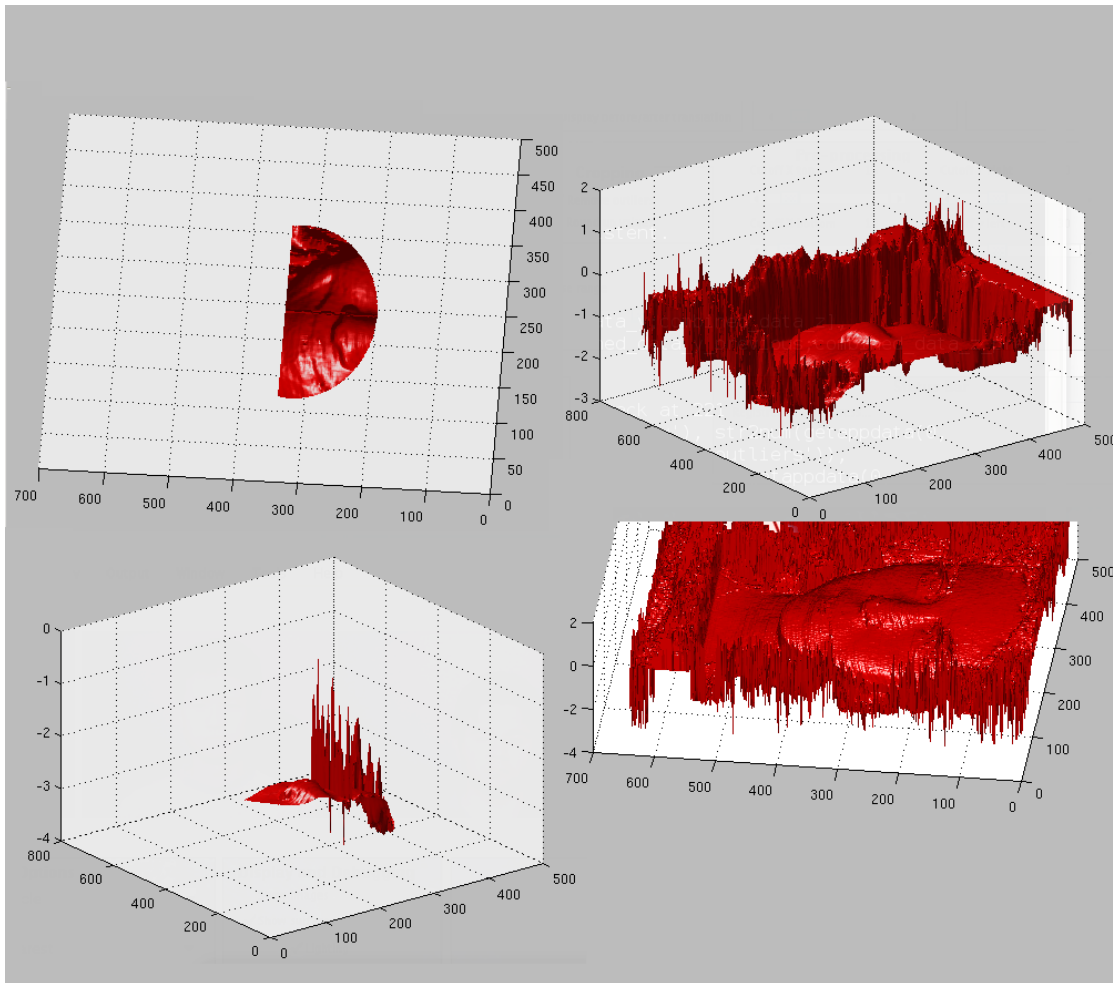


Figure 324: Assumed mark (left) extracted from accompanying GIP data (right), illustrating mis-detection

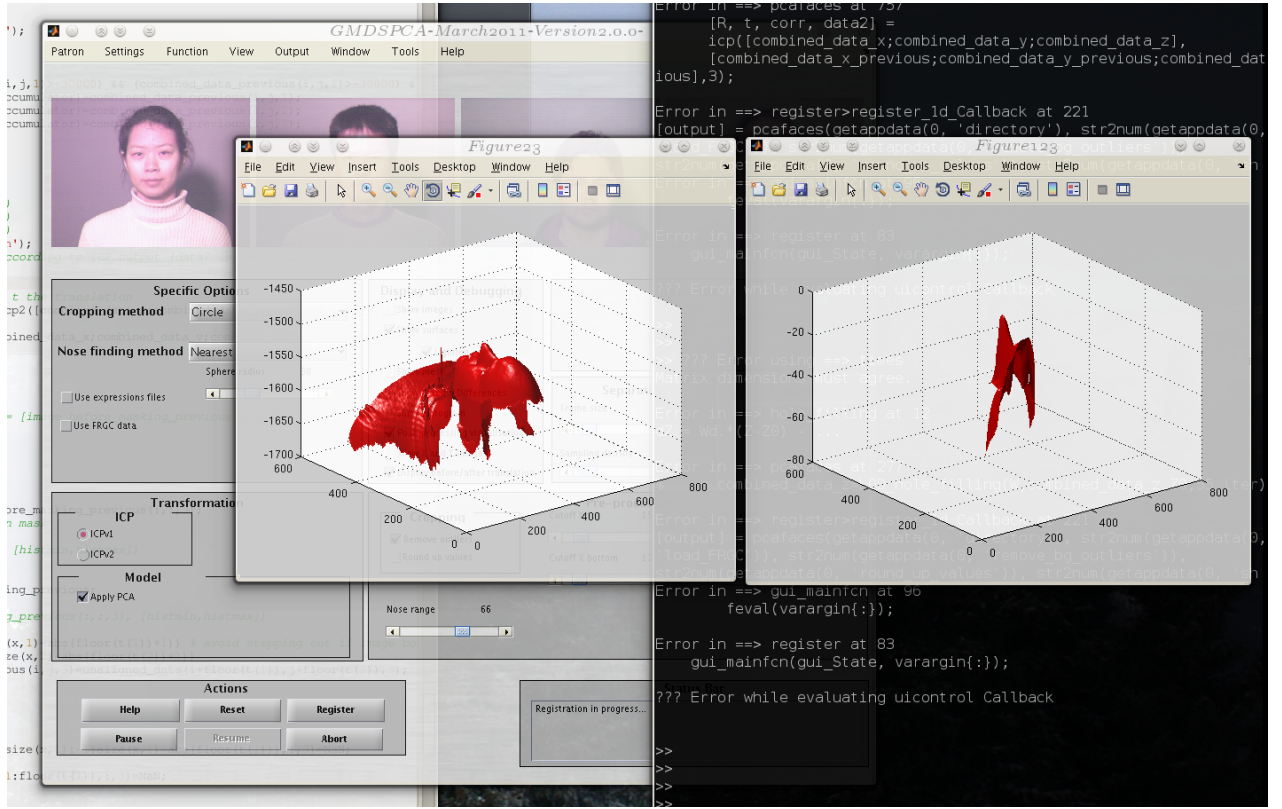


Figure 325: The offset problem visualised

I ended up writing some code to annul the effect of frames being divided upon themselves. It was reasonable to be surprised that code in GIP examples did not already have this, so suggested for it to go upstream, too.

Acknowledgements: the project was funded by the [European Research Council](#).

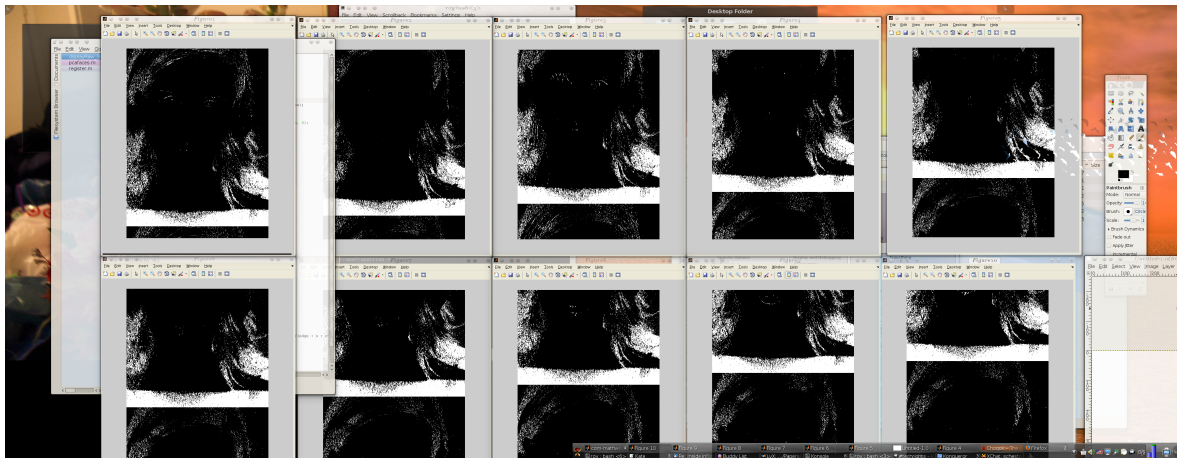


Figure 326: The face before (left) and after cropping (right)

References

- [1] J. Ahlberg and R. Forchheimer, “Face tracking for model-based coding and face animation,” *International Journal of Imaging Systems and Technology*, vol. 13, pp. 8–22, 2003.
- [2] B. Azouz, Z. Ben, A. Prosenjit, B. Chang, and S. S. Wuhrer, "Approximations of Geodesic Distances for Incomplete Triangular Manifolds," *CCCG 2007*. 465
- [3] F. Al-Osaimi, M. Bennamoun, and Ajmal Mian, “An Expression Deformation Approach to Non-rigid 3D Face Recognition,” *Int’l Journal of Computer Vision (IJCV)*, vol. 81(3), pp. 302–316, 2009. 50
- [4] P. R. Andresen, F. L. Bookstein, K. Conradsen, B. Ersboll, J. Marsh, and S. Kreiborg, “Surface-bounded growth modeling applied to human

- mandibles," IEEE Transactions on Medical Imaging, vol. 19, pp. 1053–1063, 2000.
- [5] D. Beymer and T. Poggio. "Image Representations for Visual Learning," in *Science*, vol. 272, issue 5270, pp. 1905–1909.
- [6] K. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Journal of Computer Vision and Image Understanding*, 101(1), pp. 1–15, 2006.
- [7] A. Bronstein, M. Bronstein, and R. Kimmel, "Three-dimensional face recognition. *International Journal on Computer Vision*," 64(1), pp. 5–30, 2005.
- [8] A. Bronstein, M. Bronstein, and R. Kimmel, "Generalized multi-dimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings National Academy of Sciences (PNAS)*," 103(5), pp. 1168–1172, 2006. 115
- [9] A. Bronstein, M. Bronstein, and R. Kimmel, "Robust expression-invariant face recognition from partially missing data. In *Proceedings of the European conference on computer vision*," 2006. 115
- [10] A. Bronstein, M. Bronstein, and R. Kimmel, "Efficient computation of isometry-invariant distances between surfaces," *SIAM J. Scientific Computing*, vol. 28/5, pp. 1812–1836, 2006.

- [11] A. Bronstein, M. Bronstein, and R. Kimmel, "Face2Face: an isometric model for facial animation," in Proceedings of Conference on Articulated Motion and Deformable Objects, pp. 38–47, 2006. 115
- [12] A. Bronstein, M. Bronstein, A. Bruckstein, and R. Kimmel, "Matching two-dimensional articulated shapes using generalized multidimensional scaling," in Proceedings of Conference on Articulated Motion and Deformable Objects (AMDO), pp. 48–57, 2006. 115
- [13] A. Bronstein, M. Bronstein, and R. Kimmel, "Calculus of non-rigid surfaces for geometry and texture manipulation," IEEE Trans. Visualization and Computer Graphics, vol. 13, pp. 902–913, 2007. 115
- [14] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant representations of faces," IEEE Trans. Image Processing, vol. 16(1), pp. 188–197, 2007.
- [15] A. Bronstein, M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching," Proc. National Academy of Sciences (PNAS), vol. 103(5), pp. 1168–1172, 2006.
- [16] W. R. Crum, O. Camara, D. Rueckert, K. Bhatia, M. Jenkinson, and D. L. G. Hill, "Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation," in Proceedings of Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 3749. pp. 99–106, 2005.

- [17] W. R. Crum, T Hartkens, and D. L. G. Hill, "Non-rigid image registration: theory and practice," *British Journal of Radiology*, vol. 77, pp. 140–153, 2004.
- [18] T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor, "A unified framework for atlas matching using Active Appearance Models," in *Information Processing in Medical Imaging, Proceedings*, vol. 1613, *Lecture Notes in Computer Science*, 1999, pp. 322–333.
- [19] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681–685, 2001.
- [20] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, *Lecture Notes in Computer Science*, vol. 1407. Springer, pp. 484–498, 1998. 45
- [21] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C.J. Taylor, "Groupwise diffeomorphic non-rigid registration for automatic model building," in *Proceedings of the European Conference of Computer Vision 2004*, pp. 316–32. 45
- [22] T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor, "Groupwise Construction of Appearance Models using Piecewise Affine Deformations." in *Proceedings of the British Machine Vision Conference (BMVC)*, Kingston UK, 2004. 45
- [23] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "View-based active appearance models," *Image and Vision Computing*, vol. 20, pp. 657–664, 2002.

- [24] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "Coupled-view active appearance models," British Machine Vision Conference, vol. 1, pp. 52–61, 2000.
- [25] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "3D statistical shape models using direct optimisation of description length," in Proceedings of the European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 2352, pp. 3–20, 2002.
- [26] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "Minimum description length approach to statistical shape modeling," IEEE Transactions on Medical Imaging, vol. 21, pp. 525–537, 2002.
- [27] R. H. Davies, C. J. Twining, P. D. Allen, T. F. Cootes, and C. J. Taylor, "Shape discrimination in the hippocampus using an MDL model," in Information Processing in Medical Imaging Proceedings, Lecture Notes in Computer Science, vol. 2732, pp. 38–50, 2003. 290
- [28] R. H. Davies, C. J. Twining, and C. J. Taylor, "Statistical Models of Shape: Optimisation and Evaluation," Springer, 2008.
- [29] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998. 57
- [30] O. Gervei, A. Ayatollahi, N. Gervei, "3D Face Recognition Using Modified PCA Methods," World Academy of Science, Engineering and Technology, 63 2010. 64

- [31] S. Gupta, M. Markey, and A. Bovik, "Anthropometric 3D Face Recognition," *International Journal of Computer Vision*, pp. 331–349, 2010. 85
- [32] C. A. Hack and C. J. Taylor, "Modelling 'talking head' behaviour," in *British Machine Vision Conference (BMVC03)*, 2003. 55
- [33] D. W. Hansen, M. Nielsen, J. P. Hansen, A. S. Johansen and M. B. Stegmann, "Tracking eyes using shape and appearance," *IAPR Workshop on Machine Vision Applications*, pp. 201–204, 2002.
- [34] T. Heimann and H. P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, issue 4, pp. 543–563, 2009.
- [35] A. Hill, C. J. Taylor, and A. D. Brett, "A framework for automatic landmark identification using a new method of nonrigid correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 241–251, 2000.
- [36] Y. Huang, Y. Wang, and T. Tan, "Discriminating 3D Faces by Statistics of Depth Differences," *8th Asian Conference on Computer Vision (ACCV)*, part 2, LNCS 4844, pp. 690–699, 2007. 144
- [37] S. M. S. Islam, R. Davies, M. Bennamoun, and A. S. Mian, "Efficient Detection and Recognition of 3D Ears," *Int. Journal of Computer Vision*, 2011. 53
- [38] I.T. Joliffe, "Principal Component Analysis," *Springer Series in Statistics*, Springer, New York, 1986.

104

- [39] M. Kaiser, G. Heym, N. Lehment, A. Arsic, and G. Rigoll, "Dense Point-to-Point Correspondences Between 3D Faces Using Parametric Remeshing for Constructing 3D Morphable Models," In Proc. of the Workshop on Applications of Computer Vision (WACV), pp. 39–44. 2011. 508
- [40] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models. International Journal of Computer Vision," 1(4), pp. 321–331, 1988. 56
- [41] R. Kimmel and J. A. Sethian, "Computing geodesic on manifolds," In Proceedings of US National Academy of Science, vol. 95, pp. 8431–8435, 1998.
- [42] A. C. W. Kotcheff and C. J. Taylor, "Automatic construction of eigen-shape models by genetic algorithm," in Information Processing in Medical Imaging, vol. 1230, pp. 1–14, Lecture Notes in Computer Science, 1997.
- [43] H. Kong, L. Wanga, E. K. Teoha, X. Lia, J. Wangb, and R. Venkateswarlub, "Generalized 2D principal component analysis for face image representation and recognition," Neural Networks 18, pp. 585–594, 2005. 77
- [44] A. C. W. Kotcheff and C. J. Taylor, "Automatic construction of eigen-shape models by direct optimization," Medical Image Analysis, vol. 2, pp. 303–314, 1998.

- [45] A. Lanitis, "PROSOPO - A face image synthesis system," *Advances in Informatics, Lecture Notes in Computer Science*, vol. 2563, pp. 297–315, 2003.
- [46] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic Face Identification System Using Flexible Appearance Models," *Image and Vision Computing*, vol. 13, pp. 393–401, 1995.
- [47] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 442–455, 2002.
- [48] Y. Li, S. Gong, and H. Liddel, "Constructing facial identity surfaces," in *Computer Vision and Pattern Recognition*, 2001.
- [49] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum, "Estimation of subspace arrangements with applications in modeling and segmenting mixed data," *Submitted to SIAM Review*, 2006. 128
- [50] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, pp. 1–36, 1998. 102
- [51] C. R. Maurer, G. B. Aboutanos, B. M. Dawant, R. J. Maciunas, and J. M. Fitzpatrick, "Registration of 3-D Images Using Weighted Geometrical Features," *IEEE Transaction on Medical Imaging*, vol. 15:6, pp. 836–849, 1996.
102
- [52] F. Mémoli and G. Sapiro, "A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*," 5(3), pp. 313–347, 2005. 49

- [53] J. Mena-chalco, I. Macedo, L. Velho, and R. Cesar-Jr, "PCA-Based 3D Face Photography," Brazilian Symposium on Computer Graphics and Image Processing, pp. 313—320, 2008. 64
- [54] A. 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, vol. 29, no. 11,, pp. 1927–1943, 2007. 145
- [55] C. Moenning and N. Dodgson, "Fast Marching farthest point sampling," 2003. 74
- [56] I. Mpiperis, S. Malassiotis, and M. G. Strintzis, "3-D Face Recognition With the Geodesic Polar Representation," IEEE Transactions on Information Forensics and Security, vol. 2, pp. 537 - 547, 2007. 490
- [57] K. Ouji, B. B. Amor, M. Ardabilian, F. Ghorbel, and L. Chen, "3D Face Recognition using ICP and Geodesic Computation Coupled Approach," SITIS 2006. 490
- [58] Z. Pan, G. Healey, and B. Tromberg, "Comparison of Spectral-Only and Spectral/Spatial Face Recognition for Personal Identity Verification," EURASIP Journal on Advances in Signal Processing, 2009. 491
- [59] V. S. Petrovic, T. F. Cootes, C. J. Twining, and C. J. Taylor, "Automatic Framework for Medical Image Registration, Segmentation and Modeling," Proceedings of Medical Image Understanding and Analysis. Vol. 2, pp.141–145, 2006.

- [60] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, "Overview of the Face Recognition Grand Challenge," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp.947-954, 2005. 147
- [61] S. Romdhani, S. Gong, and A. Psarrou, "A multi-view nonlinear active shape model using kernel PCA." in Proceedings of the British Machine Vision Conference, pp. 483-492, 1999.
- [62] D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic construction of 3-D statistical deformation," IEEE Transactions on Medical Imaging, vol. 22, issue 8, pp. 1014–1025, 2003.
- [63] D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic construction of 3D statistical deformation models using non-rigid registration," presented at Medical Image Computing and Computer-Assisted Intervention 2001.
- [64] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images," IEEE Transactions on Medical Imaging, vol. 18, issue 8, pp. 712–729, 1999.
102
- [65] T. Russ, C. Boehnen, and T. Peters, "3D Face Recognition Using 3D Alignment for PCA," Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1391–1398, 2006. 63
- [66] R. S. Schestowitz, C. J. Twining, T. F. Cootes, V. S. Petrovic, C. J. Taylor, and B. Crum, "Assessing the Accuracy of Non-Rigid Registration

- With and Without Ground Truth,” IEEE International Symposium on Biomedical Imaging (ISBI), 2006. 106
- [67] R. S. Schestowitz, C. J. Twining, T. F. Cootes, and C. J. Taylor, “Image Registration by Model Criteria,” presented in Proceedings of MIAS-IRC Plenary Meeting, pp. 16–17, 2004. 106
- [68] R. S. Schestowitz, C. J. Twining, T. F. Cootes, V. S. Petrovic, and C. J. Taylor, “A Generic Method for Evaluating Appearance Models,” presented in Proceedings of MIAS-IRC Plenary Meeting, 2006. 106
- [69] I. M. Scott, T. F. Cootes, and C. J. Taylor, “Improving appearance model matching using local image structure,” in Proceedings of Information Processing in Medical Imaging, Lecture Notes in Computer Science, vol. 2732, pp. 258–269, 2003.
- [70] M. Sonka, V. Hlavac, and R. Boyle, “Image processing, analysis and machine vision,” Pacific Grove, Calif. ; London: PWS Publishing, 1999.
- [71] L. Spreeuwiers, “Fast and Accurate 3D Face Recognition Using Registration to an Intrinsic Coordinate System and Fusion of Multiple Region Classifiers,” Int. Journal of Computer Vision, pp. 389–414, 2011. 54
- [72] M. B. Stegmann, B. K. Ersboll, and R. Larsen, “FAME - A flexible appearance modeling environment,” IEEE Transactions on Medical Imaging, vol. 22, pp. 1319–1331, 2003. 57
- [73] V. Surazhsky, T. Surazhsky, D. Kirsanov, S. J. Gortler, and H. Hoppe, “Fast exact and approximate geodesics on meshes,” ACM SIGGRAPH 2005, vol. 24, pp. 553–560, 2005. 464

- [74] J. R. Tena 1 F. De la Torre, and I. Matthews, “Interactive Region-Based Linear 3D Face Models,” *ACM Transactions on Graphics (Proc. ACM SIGGRAPH)*, Aug. 2011 (to appear). 67
- [75] C. J. Twining and S. Marsland, “Constructing diffeomorphic representations of non-rigid registrations of medical images,” presented at *Information Processing in Medical Imaging 2003*.
- [76] C. J. Twining, S. Marsland, and C. J. Taylor, “Groupwise non-rigid registration: the minimum description length approach,” *British Machine Vision Conference 2004*.
- [77] C. J. Twining and C. J. Taylor, “The use of kernel principal component analysis to model data distributions,” *Pattern Recognition*, vol. 36, pp. 217–227, 2003.
- [78] C. J. Twining and C. J. Taylor, “Specificity as a Graph-Based Estimator of Cross-Entropy,” in *Proceedings of the British Machine Vision Conference*, vol. 2, pp. 459–468, 2006.
- [79] C. J. Twining, S. Marsland, and C. J. Taylor, “Measuring geodesic distances on the space of bounded diffeomorphism,” in *Proceedings of the British Machine Vision Conference (BMVC’02)*, 2002.
- [80] C. J. Twining, T. F. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. J. Taylor, “A unified information-theoretic approach to groupwise non-rigid registration and model building,” in *Proceedings of Information Processing in Medical Imaging (IPMI)*, *Lecture Notes in Computer Science*, vol. 3565. Springer, 2005, pp. 1–14.

- [81] R. Vidal, Y. Ma, and Jacopo Piazzzi, “A New GPCA Algorithm for Clustering Subspaces by Fitting, Differentiating and Dividing Polynomials” CVPR 2004, pp. 510–517, 2004.
- [82] R. Vidal and S. Sastry, “Generalized principal component analysis (GPCA),” IEEE TPAMI, vol. 27, pp. 621–628, 2003. 74
- [83] L. Wang, Y. Zhang, and J. Feng, “On the euclidean distance of images,” IEEE Trans. Pattern Anal. Machine Intell., vol. 27, pp. 1334–1339, 2005.
- [84] F. Wang, J. Wang, C. Zhang, J. Kwok, "Face recognition using spectral features," Pattern Recognition, 2007 491
- [85] X. Wu, Z. Wang, H. Wang, and Y. Leng, “Generalized PCA Face Recognition by Image Correction and Bit Feature Fusion,” Neural Information Processing Published, vol. 4233, pp. 227–235, 2006. 78
- [86] J. Xie, D. Alcantara¹, N. Amenta¹, E. Fletcher, O. Martinez, M. Persianinova, C. DeCarli, and O. Carmichael, "Spatially-Localized Hippocampal Shape Analysis in Late-Life Cognitive Decline," MIC-CAI 2008 Workshop on Computational Anatomy and Physiology of the Hippocampus, pp. 2–12, 2008.A. Y. Yang, S. R. Rao, and Yi Ma, “Robust statistical estimation and segmentation of multiple subspaces,” CVPR workshop on 25 years of RANSAC, 2006. (<http://picsl.upenn.edu/caph08/papers/paper16.pdf>) 290

- [87] A. Y. Yang, S. R. Rao, and Yi Ma, "Robust statistical estimation and segmentation of multiple subspaces," CVPR workshop on 25 years of RANSAC, 2006. 128
- [88] J. Ye, "GPCA: An Efficient Dimension Reduction Scheme for Image Compression and Retrieval," 2004.
- [89] L. Younes, "Deformations, Warping and Object Comparison: A Tutorial," 2000.
- [90] A. Yuille, D. Cohen, and P. Hallinan, "Feature extraction from faces using deformable templates," In Proceedings of CVPR, pp. 104--109, 1989.
- [91] B. Zitova and J. Flusser, "Image registration methods: A survey," Image and Vision Computing, vol. 21, pp. 977--1000, 2003.
- [92] W. Zhao,, R. Chellappa, P. Phillips, and A. Rosenfeld. "Face recognition: a literature survey," ACM Computing Surveys, 35(4), pp. 399--458, 2003.
- [93] H. Zhou and A.H. Sadka, "Combining Perceptual Features With Diffusion Distance for Face Recognition," IEEE Transactions on Cybernetics, vol. 41, pp. 577 - 588, 2011. 492
- [94] Y. Zhu and E. Sung, "The Spectral-Face Analysis for Face Recognition," The 5th Asian Conference on Computer Vision, 2002.

