3D Facial Landmark Localisation by Matching Simple Descriptors

Marcelo Romero-Huertas and Nick Pears

Abstract— **We present our graph matching approach for 3D facial feature localisation. The work here uses a basic graph model (three vertices and three arcs) to locate the inner eye corners and the nose tip simultaneously. We intend to extend this to a larger set of the eleven features that exist in our ground truth of the Face Recognition Grand Challenge (FRGC) database. We apply the structural matching algorithm "relaxation by elimination" using a simple "distance to local plane" node property and a "Euclidean distance" arc property. After the graph matching process has eliminated unlikely candidates, the most likely feature combination (left eye, right eye and nose tip) is selected, by exhaustive search, as the minimum Mahalanobis distance over a six dimensional space, corresponding to three node variables and three arc variables. Our results on the 3D FRGC database are presented and discussed.**

I. Introduction

Automatic facial feature (landmark) localisation is an important component in many face processing applications, such as face tracking, face modelling, animation, expression analysis, face identification and face verification. Eye centres are suitable facial features for location when 2D intensity images are used, because of their dark, uniform texture and rounded shape. Similarly, the nose-tip is often quoted as the most distinctive feature in 3D images [10]. However, eye centres tend to be locally flat in 3D, whereas the inner eye corner is often highly concave, especially in Caucasian racial types, making them more distinctive on the 3D facial surface. Furthermore, it has been shown [14] that the area between the eyes and the nose of the human face is more distinctive for recognition using 3D data, and it has been proved robust in presence of facial expressions [11].

In this paper, we provide a solution to the facial feature (landmark) localisation problem using 3D data, initially focussing on the largely rigid triplet of landmarks that consists of the inner eye corners and the nose tip.

There are relatively few techniques proposed in the literature to automatically locate facial landmarks using 3D only. Conde and Serrano [3] use spin images and support vector machine (SVM) classifiers to locate the nose and the eyes. In Xu et al [6], a 3D nose tip approach is presented. Here the SVM is used in a hierarchical filtering scheme and 99.3% successful nose tip localisations were reported, although testing was not done on widely used benchmark datasets. Different pose-dependent approaches for 3D feature location have been reported using the FRGC database [11]-[13], and still some problems are noted due to shirt collars and hair styles present in the dataset. In other papers [8], [9], alternative pose-dependent approaches to localise facial features or the face are presented (using a variety of databases). A similar approach to our work is reported by Colombo et al [2], where the same three features are located simultaneously using curvature of the face. However, our technique is different in essence and it is tested using the FRGC benchmark database.

In the work of Segundo et al [7], 99.9% successful landmark detection is reported using the FRGC database, but it is a technique constrained to a facial frontal pose. We are presenting an approach robust to pose and facial expression variations which uses distinctive shape features.

Our final objective is a graph matching approach robust to extreme pose and facial expression variations, which is relevant to unconstrained face recognition [1], [5]. The preliminary results presented here are motivating and guiding our future work toward that final objective.

The rest of this paper is structured as follows. Our experimental design is detailed in section 2. Results are presented in section 3. Finally, conclusions and future work are discussed in section 4.

II. Experimental design

Our complete experiment is outlined in figure 1. In summary, we analyzed the FRGC database and collected pre-processed data, namely distance to local plane (DLP), from down-sampled 3D data, using a down-sample factor of 4. We also manually verify 2D-3D correspondence in the FRGC database in order to be able to collect ground-truth landmarks localisations using shape and intensity images simultaneously. Separate training and testing sets were defined within the FRGC data. After our feature localisation process has finished, performance results are collected by comparing the localised feature landmarks against our ground-truth data.

A. Benchmark database (FRGC)

The FRGC database contains the largest 3D face dataset that is widely available to the research community. In it, there are 4,950 shape images and each of these has an associated intensity image. The files are divided into three subsets, named after their collection period: Spring-2003, Fall-2003 and Spring-2004.

Manuscript received May 31, 2008.

Authors are with the Computer Science Department, The University of York, Heslington, York, YO105DD, UK, {mromero, nep}@cs.york.ac.uk.

M. R. author is supported by the Mexican National Council of Science and Technology (CONACYT), grant 207690.

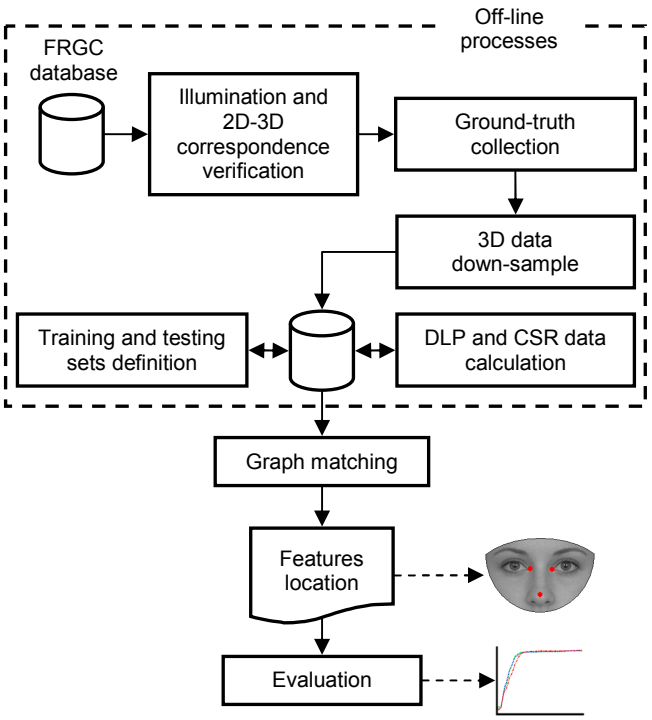


Fig. 1. Block diagram of the complete experimentation reported in this paper.

The Spring-2003 subset was collected under controlled illumination. Participants in this term were positioned at various depths from the camera. As a consequence, several images include not only the face but also the upper part of the body, i.e. shoulders and chest. Generally, all images present an unoccluded fronto-parallel pose in a neutral expression.

Fall-2003 and Spring-2004 subsets were collected under uncontrolled illumination and varying facial expression. In contrast to the Spring-2003 subset, in most of the images, only the participant's face was captured. Again, there are no extreme pose variations and a fronto-parallel pose is used. Table I shows how the FRGC database is populated.

TABLE I
ORIGINAL FRGC DATABASE POPULATION

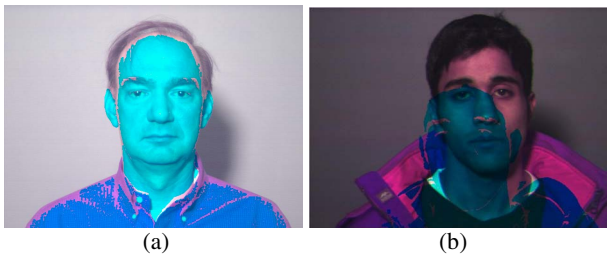
Imgs/person	Spring-2003	Fall-2003	Spring-2004
1	77	47	43
2	32	31	24
3	47	42	22
4	33	47	30
5	28	45	28
6	30	38	33
7	15	29	35
8	13	30	32
9	77	36	29
10	32	25	32
11	-	-	27
12	-	-	10
Files	943	1893	2114
People	275	370	345

B. *Filter data with poor 2D-3D registration*

The 3D sensor used to collect the FRGC data acquires the texture image just after the shape image acquisition. Thus subject motion can cause poor registration between the intensity and its shape counterpart [16]. For an objective performance evaluation, we manually eliminated from the FRGC database those files with a visually poor 2D-3D correspondence.

We visually verified correspondence using a composed image which effectively is an orthographic projection of the 3D data into 2D (the z dimension is discarded). Figure 2 shows an example, where the 3D projection is visually observed as a blue translucent film layer over the intensity image. Poor registration is visually identified if there is a mismatch between this projection and the intensity image.

Table II shows a summary of files with correspondence between its shape and intensity images. Note that records with extreme lighting variations are difficult to verify using this technique and so those files are not considered in our experimentation.



(a) (b)

Fig. 2. Two examples of 2D-3D correspondence verification: (a) good registration and (b) poor registration.

TABLE II
FRGC FILES WITH 2D-3D CORRESPONDENCE

Imgs/person	Spring-2003	Fall-2003	Spring-2004
1	85	63	49
2	46	40	27
3	39	52	31
4	35	51	43
5	21	44	32
6	21	34	33
7	4	24	30
8	2	23	37
9	-	22	25
10	-	3	20
11	-	-	12
12	-	-	2
Files	709	1507	1813
People	253	356	341

C. *Data pre-processing*

The FRGC database was collected using a resolution of 640 by 480; which is standard for intensity images, but rather high resolution for 3D processing. We firstly down-sampled data by a factor of four, so that a typical batch processing job on a FRGC 3D dataset using MATLAB was generally achievable in an overnight processing session. We chose a down-sample factor of four, as our preferred trade-off between 3D shape resolution and processing-time.

Even under controlled illumination for a given sensor, it is common for 3D errors to occur in and around the facial regions, for example due to the poor reflectivity of hair [15]. These errors consist of spikes, pits (negative spikes) and holes (data absence). To overcome these problems, a basic data

filtering step was used as a pre-process on our training data. This consisted of first spike/pit elimination (thus creating extra holes), followed by interpolation over all holes.

D. Ground-truth data collection

For an objective performance evaluation, it is necessary to have a good ground-truth to estimate the error in feature localisation. However, the FRGC database is only provided with limited ground-truth data (4 landmarks). We felt that we needed more facial landmarks in our ground truth dataset (we marked up 11 landmarks) and that this data needed to be more meticulously populated.

As mentioned before, the most distinctive facial features of the face are the eyes and the nose, for this reason we focus our attention on them. The anatomy of the face, specifically its bone structure, divides the face in two parts: rigid (largely) and non-rigid. A complete approach needs to consider both areas and their features, for this reason we selected 11 landmarks: eye corners, nose-bridge, nose-tip, mouth corners and the chin. Figure 3 and table III illustrate these ground-truth feature points.

To obtain our ground-truth data we take advantage of both intensity and shape images. Eleven facial feature points were collected by very carefully manually clicking on enlarged intensity images and then computing the corresponding 3D point using the registered 3D shape information. We use a dual (2D and 3D) view to verify 2D-3D landmark correspondences.

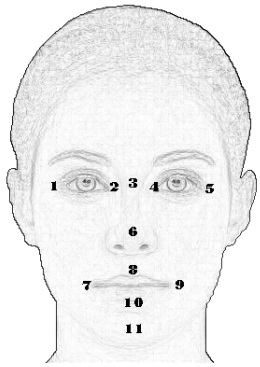


TABLE III
LANDMARKS MANUALLY COLLECTED

1	Outer right eye corner
2	Inner right eye corner
3	Nose bridge
4	Inner left eye corner
5	Outer left eye corner
6	Nose tip
7	Right mouth corner
8	Centre upper lip
9	Left mouth corner
10	Centre lower lip
11	Centre chin

Fig. 3. Landmarks manually collected. In this paper only 3 landmarks were used, but we will extend our graph model to use all eleven landmarks in future work.

E. Data representation

Our graph matching approach is flexible and different features can be used to represent the nodes and the arcs. We begin by exploiting "distance to local plane" (DLP) as our node representation, because it is stable, computationally inexpensive and can be implemented with any linear algebra package.

As illustrated in figure 4, for each point in one point cloud we can find its neighbouring points $X=\{x_1, x_2, \cdots, x_n\}$ lying within a sphere of radius r and centred at this point. Let π be the plane that better fit the neighbouring set X with normalized normal $\vec{n}_\pi$. Thus, we can calculate d as the inner product of the vectors $p-\mu$ and $\vec{n}_\pi$:

$$d(\pi, p) = (p-\mu)\cdot\vec{n}_\pi \quad (1)$$

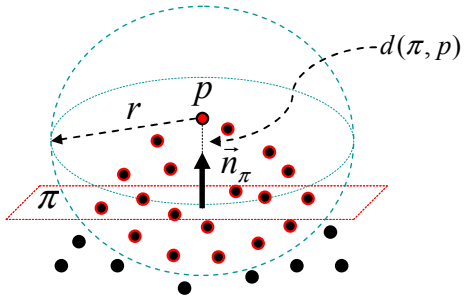


Fig. 4. A plane π is fit using neighbouring points around p in a radius r, so d is calculated using dot product of the normal $\vec{n}_\pi$ and p.

Definition (1) requires the normal of the plane which can be estimated using singular value decomposition (SVD) of the covariance matrix $\Sigma=(X-\mu)(X-\mu)^T$, where X is the set of neighbouring points of p and μ the mean vector of X.

Hence, Σ is a 3 by 3 positive semi-definite matrix, we determine that λ_1, λ_2 and λ_3 $(\lambda_1 \geq \lambda_2 \geq \lambda_3)$ denote the largest eigenvalues of Σ corresponding to the eigenvectors v_1, v_2 and v_3 respectively. Commonly, the perpendicular direction of the local area has the fewest points. So, v_3 similarly corresponds to the normal direction. Then, we use v_3 to approximate the normal $\vec{n}_\pi$ of the local plane π.

By using this representation with an appropriate radius r, we can identify points over convex (DLP>0), flat (DLP close to zero) and concave (DLP<0) areas of the facial surface. By bounding the allowable values of DLP with Mahalanobis distance, referenced to the mean and variance of our training data, we can identify both nose (suitably convex) and eye corner (suitably concave) candidate vertices. Training data consists of 200 images from 200 different subjects from Spring-2003 FRGC subset. For each training image, we compute DLP for nodes 2, 4 and 6 (table III) and the Euclidean distance between each node pair.

F. Graph matching

The graph model we fit is very simple and consists of three nodes and three arcs, as shown in figure 5. Obviously, exhaustively testing every possible vertex triplet against training data is too computationally expensive and we seek to significantly reduce the number of vertex triplets that we have to test, first by checking for appropriate nodal attributes, and then by checking pairwise relationships between node pairs.

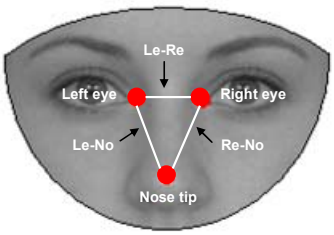


Fig. 5. Our graph model consists of three vertices (inner eye-corners and nose-tip) and three arcs: leftEye-rightEye (Le-Re), leftEye-noseTip (Le-No) and rightEye-noseTip (Re-No).

To do this we use a structural graph matching algorithm known as relaxation by elimination (RBE) [4], and in our implementation, we divide this into four steps: 1) Initialization, 2) Generation, 3) Iteration and 4) Selection, as shown in figure 6.

Initialization populates an initial candidate list for each of the three nodes, based on the Mahalanobis distance of the DLP value, using the appropriate mean and variance from the training data. For a data vertex to become a candidate its Mahalanobis distance must be less than three.

After that, binary arrays are created (generation) to represent pairwise 'Euclidean distance' relationships between nodes and arcs in our model. We refer to these binary arrangements as contextual support relationship (CSR) matrices and we have three in our model:

$$CSR_{leftEye-rightEye} \rightarrow [candidateLeftEyes,\ \ candidateRightEyes]$$
$$CSR_{leftEye-noseTip} \rightarrow [candidateLeftEyes,\ \ candidateNoseTips]$$
$$CSR_{rightEye-noseTip} \rightarrow [candidateRightEyes,\ \ candidateNoseTips]$$

Essentially, a '1' exists in a CSR matrix, if the two node candidates are mutually supportive. This means that the two DLP values (one for each node candidate) and the Euclidean distance between them must fall sufficiently close to the multivariate (3-DOF) distribution of these values in the training data. Again, a Mahalanobis distance of less than three is required for the candidates to be deemed mutually supportive.

We have noted that vertices close to each other have very similar DLP values and hence we often get clusters of candidate vertices around the ground truth landmark. This means, for example, that a particular left eye candidate can have many right eye candidates that are mutually supportive. This is where the 'elimination' in the RBE iteration comes in. Every least supported feature candidate is iteratively eliminated, until a stop condition is obtained, i.e. either a minimum number of candidates remain or a maximum number of iterations is reached.

Finally, the best combination is selected by exhaustive search of the remaining possible candidate triplets. This is done by computing the Mahalanobis distance in the multivariate (6-DOF) feature space [DLP_leftEye, DLP_rightEye, DLP_noseTip, E_left-right, E_left-nose, E_right-nose]. Again, the mean and covariance matrices are determined from the training data. If the node triplet with the minimum Mahalanobis distance has a distance value of less than 3, this triplet is retained as a successful graph matching output. Otherwise, remaining candidates after relaxation are considered false positives. These are deleted and the process is restarted from the generation stage, as shown by the dotted line in figure 6. Note that this happens in less than 1% of our 3D test images.

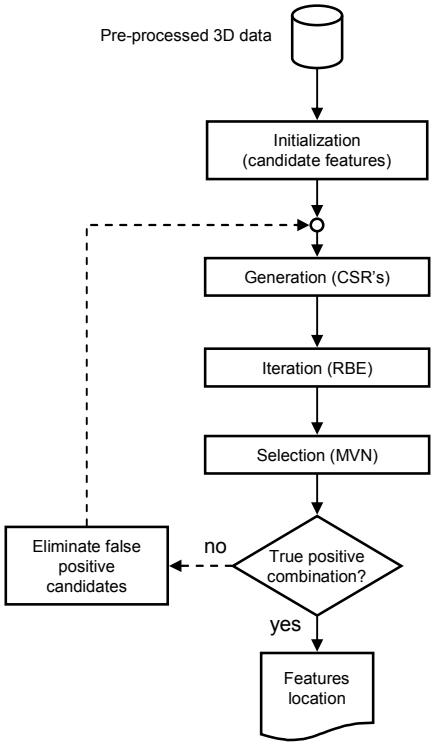


Fig. 6. Flowchart of our graph matching approach.

III. Results

Our graph matching approach was tested in two scenarios, considering both variations in depth and facial expression. The FRGC database is already divided in this way and we adopted them as they are. Naturally, there are variations in illumination and small variations in pose.

A. Scenario #1: Depth variations, neutral expressions

Although the Spring-2003 subset was created under controlled illumination and generally neutral expressions, large variations in depth are presented. This subset originally consists of 943 files, 200 were used to train our system and 509 were used for testing. The rest were not considered because they showed poor 2D-3D correspondence.

We gather results by computing the root mean square (RMS) error of the automatically localised landmarks with respect to the landmarks manually labelled in our ground truth. Remember that localisation is done at the 3D vertex level and we are using a down-sample factor of four on the FRGC dataset, which gives a typical distance between vertices of around 3-5mm. This has implications on the achievable localisation accuracy. We set a distance threshold (mm) and if the RMS error is below this threshold, then we label our result as a successful localisation. Then, by varying this distance threshold, we can observe how the percentage of

successful localisations changes for each feature, as shown in Figure 7. This allows us to present results which are not dependent on a single threshold and indicates two distinct phases in the success rate: (i) a rising phase where an increased RMS distance threshold masks small localisation errors at the "down-sample 4" resolution, and (ii) a plateau in the success rate, where an increased RMS threshold does not give a significant increase in the success rate of localisation. This indicates the presence of gross errors in localisation. Of course, it is useful to choose some RMS threshold values and quote performance figures. A sensible place to choose the threshold is close to where the graph switches from the rising region to the plateau region. Using this idea, we note that 80% of the eye-corners and nose-tip are localised around 12 and 15 millimetres respectively. Figure 8 summarises the performance of this scenario using the criteria in table IV.

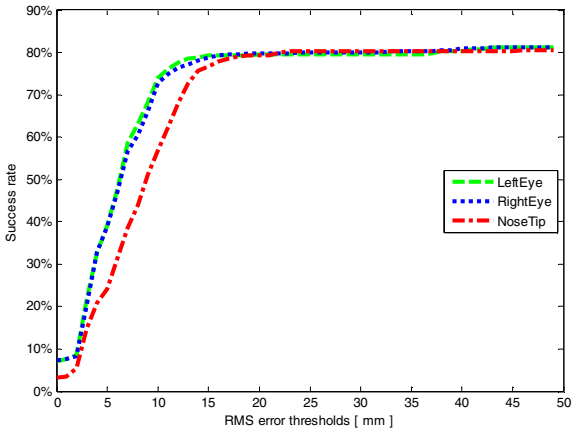


Fig. 7. Fractional success rates against RMS error thresholds from Spring-2003 subset (training = 200 and testing = 509 shape images).

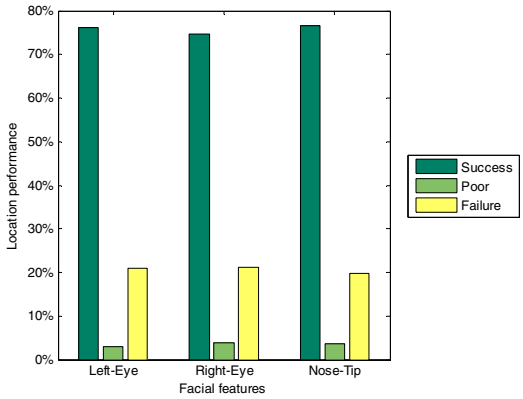


Fig. 8. Overall location performance from Spring-2003 subset (training = 200 and testing = 509 shape images).

TABLE IV
THRESHOLDS TO EVALUATE LOCATIONS ESTIMATED

	Inner eye corners	Nose tip
Success	$RMS \leq 12mm$	$RMS \leq 16mm$
Poor	$12mm < RMS \leq 16mm$	$16mm < RMS \leq 24mm$
Failure	$RMS > 16mm$	$RMS > 24mm$

Examples of successful, poor and failed landmark localisations are shown in figures 9 and 10. Those figures show the triplet with the minimum Mahalanobis distance in the 6-DOF feature space.

(a) (b)

Fig. 9. Successful feature localization, shape images: (a) 04336d211 and (b) 02463d662, with neutral and facial expression respectively.

(a) (b)

Fig. 10. Poor (a) and failed (b) feature localisation, shape images: 04297d210 and 04385d239 respectively. This failed location is a typical example where collars with better support than valid candidates are selected.

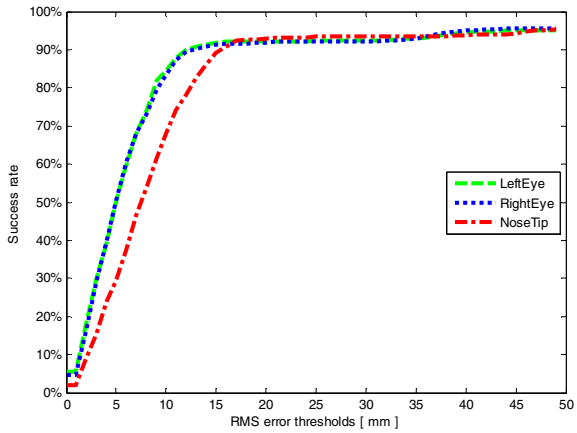


Fig. 11. Fractional success rates against RMS error thresholds by testing 1,507 people from Fall-2003 subset using the training set of the first scenario (200 different people).

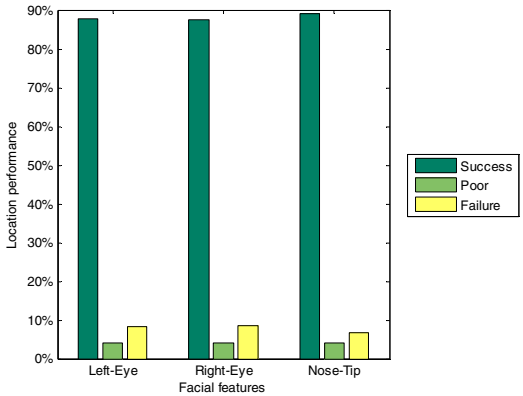


Fig. 12. Overall location performance by testing Fall-2003 subset (1,507 shape files) using a training set of 200 (from the first scenario).

B. *Scenario #2: Facial expression variations and few depth variations*

This scenario was tested using the Fall-2003 and Spring-2004 FRGC subsets which present facial expression variations, but relatively few depth variations. The same training set from scenario #1 was used; two testing sets were integrated with 1,507 (Fall-2003) and 1,764 (Spring-2004) shape images, all of which were deemed to have acceptable 2D-3D correspondence and illumination.

Figure 11 shows the fractional success rate curve, using the Fall-2003 testing set, and it can be noted that 90% of the eye-corners and the nose-tip are located around 12 and 15 millimetres respectively. Whereas, figure 12 resumes this location performance by categorising as 'success', 'poor' and 'failure' according to table IV.

We obtained a similar performance with the Spring-2004 testing set (1,764 shape images) as can be observed in figures 13 and 14.

Results in this scenario demonstrate our approach's robustness to facial expression variations, an example from this dataset is shown in figure 9-b.

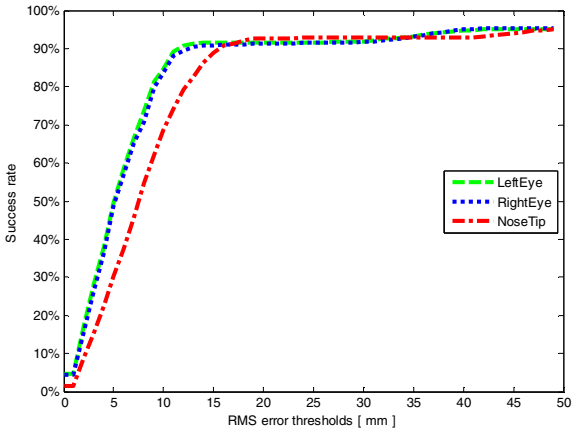


Fig. 13. Fractional success rates against RMS error thresholds by testing 1,764 shape images from Spring-2004 subset using a training set of 200 different people (from the first scenario).

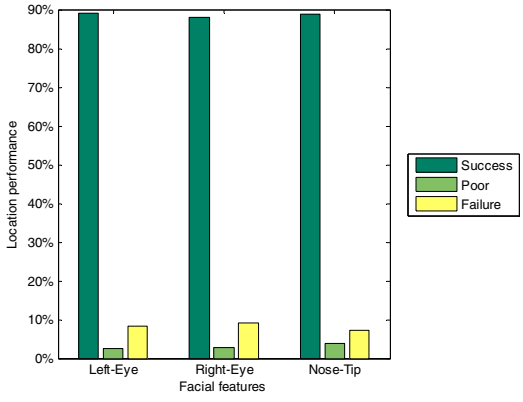


Fig. 14. Overall location performance by testing Spring-2004 subset (1,764 shape files) using a training set of 200 (from the first scenario).

IV. CONCLUSION

We have presented our graph matching approach, which is robust to facial expression variations, as shown by our results on the Fall-2003 and Spring-2004 subsets. Results from the Spring-2003 subset show a lower performance. In this subset, there are many features in the upper torso area, such as shirt collars, which have similar descriptor values to the facial landmarks that we seek. We aim to address this problem by implementing a more sophisticated approach, using a richer set of descriptors within the graph matching process.

We have presented results with the most commonly used benchmark database. In future, we aim to extend our model to consider more facial features (e.g. mouth-corners, chin). Finally, a more advanced stage in our research will include dealing with self occlusion and the associated absence of data, such as occurs in profiled poses.

REFERENCES

[1] Zhou S. K., Chellappa R. and Zhao W, Unconstrained Face Recognition, Springer, USA. 2007.

[2] Colombo, Cusano, and Schettini, "3D Face Detection Curvature Analysis", Pattern Recognition, Vol. 39, Issue 3, March 2006, pages 444-455.

[3] Conde and Serrano, "3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification", Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference on, Vol. 3, Issue 20-26.

[4] Turner and Austin, "Graph Matching by Neural Relaxation", Neural Computing & Applications (1998)7:238-248.

[5] Zhao and Chellappa, "Face Processing: Advanced Modeling and Methods", Elsevier, USA, 2006.

[6] Xu, Tan, Wang and Quan, "Combining Local Features for Robust Nose Location in 3D Facial Data", Pattern Recognition Letters 27, 2006, 1487-1494.

[7] Segundo, Queirolo, Bellon and Silva, "Automatic 3D Facial Segmentation and Landmark Detection", 14th International Conference on Image Analysis and Processing (ICIAP 2007) on proceedings, 431-436.

[8] Lu, Colbry and Jain, "Three-Dimensional Model Based Face Recognition", in proceedings of ICPR, (8) 2004.

[9] Colbry and Stockman, "Canonical Face Depth Map: A Robust 3D Representation for Face Verification", CVPR 2007.

[10] Gupta, Aggarwal, Markey and Bovik, "3D Face Recognition Founded on the Structural Diversity of Human Faces", CVPR 2007.

[11] Chang, Bowyer and Flynn, "Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 10, October 2006.

[12] Salah and Akarun, "3D Facial Feature Localization for Registration", International Workshop on Multimedia Content Representation, Classification and Security (MRCS'06), 2006.

[13] Mian, Bennamoun and Owens, "Automatic 3D Face Detection, Normalization and Recognition", 3D Data Processing, Visualization, and Transmission, Third International Symposium on, (6) 2006, 735-742.

[14] Nagamine, Uemura and Masuda, "3D Facial Image Analysis for Human Identification", Pattern Recognition, Conference A: Computer Vision and Applications, Proceedings, 11th IAPR International Conference on Volume I (9) 1992, 324 – 327

[15] Bowyer, Chang and Flynn, "A Survey of Approaches and Challenges in 3D and Multi-modal 3D+2D Face Recognition", Computer Vision and Image Understanding 101 (2006) 1-15

[16] Phillips, Flynn, Scruggs, Bowyer, Chang, Hoffman, Marques, Min and Worek, "Overview of the Face Recognition Grand Challenge", on Proceedings Computer Vision and Pattern Recognition 2005, Vol. 1, 947- 954.