Landmark Localisation in 3D Face Data

Marcelo Romero and Nick Pears
Department of Computer Science
The University of York
York, UK
{mromero, nep}@cs.york.ac.uk

Abstract— **A comparison of several approaches that use graph matching and cascade filtering for landmark localisation in 3D face data is presented. For the first method, we apply the structural graph 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 triplet is selected, by exhaustive search, as the minimum Mahalanobis distance over a six dimensional space, corresponding to three node variables and three arc variables. A second method uses state-of-the-art pose-invariant feature descriptors embedded into a cascade filter to localise the nose tip. After that, local graph matching is applied to localise the inner eye corners. We evaluate our systems by computing root mean square errors of estimated landmark locations against ground truth landmark localisations within the 3D Face Recognition Grand Challenge database. Our best system, which uses a novel pose-invariant shape descriptor, scores 99.77% successful localisation of the nose and 96.82% successful localisation of the eyes.**

Keywords – 3D feature descriptors; facial landmark localisation; cascade filter; relaxation by elimination; SSR histograms.

I. INTRODUCTION

Many face processing applications, such as face tracking, identification and verification require automatic landmark localisation. In the context of 3D face data, the nose is often quoted as the most distinctive feature [5], [6], [9] and, in addition, it is visible over a wide range of head poses. These facts make it an obvious landmark to include in a minimal set of three rigid features, which allow facial pose to be established and, if necessary, normalised [17]. One can noted that the inner eye corners are highly concave areas and so we use these additional two features to complete a minimal landmark triplet.

A further reason that we have selected this landmark triplet is that it has been shown that this particular facial area is more distinctive for recognition using 3D data [13], and it has been proved robust in presence of facial expressions [10], [19].

In this paper, we compare two approaches for facial landmark localisation. In the first approach, we localise the triplet of landmarks simultaneously, using simple descriptors embedded in a structural matching algorithm. We compare this approach, with an approach that first localises the nose tip (the easiest landmark to localise) using a cascaded filter of more sophisticated descriptors [17], and then localises the eyes relative to the nose tip. In total, we have generated a manual mark up of eleven facial landmarks across the entire Grand Challenge 3D data-set, for all images where there is an accurate registration between 2D and 3D data [18]. Our future aim is to extend the landmark localisation processes described here to the full set of eleven features.

A. Related work

There are relatively few techniques proposed in the literature to automatically locate facial landmarks using 3D data only. Conde et al. [2] use spin images and support vector machine SVM classifier to locate the nose and the eyes. Xu et al. [5] present a 3D nose tip hierarchical filtering approach constructed with an *effective energy sphere* and SVM classifier. Colbry et al. [8] use shape index for anchor point localisation. Segundo et al. [6] report a successful localisation result experimenting with the FRGC database, although it is a technique constrained to a facial frontal pose. Different pose-dependent and multimodal approaches to localise facial landmarks using the FRGC database have been reported [10]-[12], and still some problems are noted due to shirt collars and hair styles present in the dataset. Lu and Jain [7] propose a feature extractor based on the directional maximum to estimate the nose tip location and pose angle simultaneously.

We are presenting an approach robust to pose, clothing and facial expression variations which uses distinctive shape features. Our final objective is a landmark localisation approach robust to extreme pose and facial expression variations, which is relevant to unconstrained face recognition [1], [4] and the 3D face recognition challenge [14]. Results presented here are motivating and guiding our future work towards that final objective.

B. Preliminaries

Our complete experiment is outlined in figure 1. Firstly, we establish visually that 2D and 3D datasets are correctly registered. For those that are, we collect eleven ground-truth landmarks [18] by manually clicking 2D features on enlarged bitmaps, and mapping them to 3D (note that only three of these 3D landmarks are currently used). We down-sample the data by a factor of 4, mapping 3D landmarks to the nearest down-sample. We then process this data in order to establish a variety of feature descriptors, to be used for training purposes.

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.90

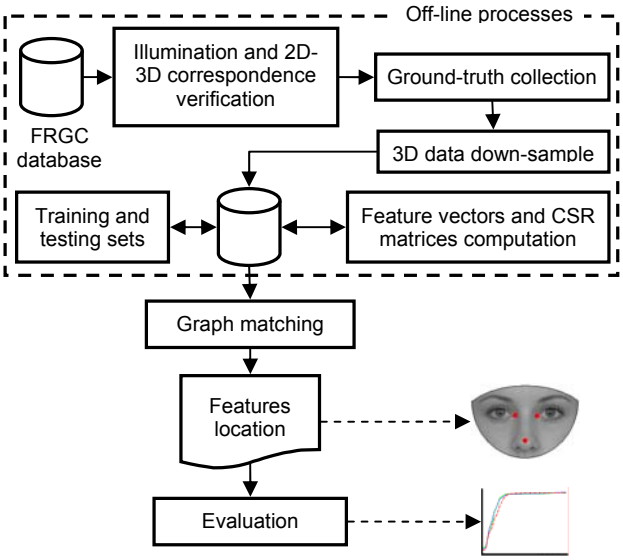

Figure 1. Block diagram of our complete experimentation.

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. Since the definition of successful landmark localisation is dependent on setting a threshold of acceptable error, we explore performance over the full range of possible thresholds. This allows us to identify both gross errors ('fails'), where the system completely fails to identify the correct landmark, and errors of poor localisation, which are due to the combined effect of any inaccuracies in the system.

The rest of this paper is structured as follows. Every feature descriptor employed in this paper is described in section 2. Our two main methods: the graph matching approach and the cascade filter are detailed in section 3. Our evaluation procedure, results and a comparison with other methods are presented in section 4. Finally, conclusions and future work are discussed in section 5.

II. FEATURE DESCRIPTORS

To localise landmarks in complex 3D data, we must extract descriptors that make such landmarks distinguishable from other points. A good landmark localisation algorithm should use descriptors invariant to rigid transformations, and robust to multi-resolution and poor quality raw data [5]. In these terms, we have selected a range of feature descriptors with various computational costs of extraction and powers of discrimination.

A. Distance to local plane (DLP)

To compute this feature descriptor, neighbouring points $X = \{x_1, x_2, \cdots, x_n\}$ in a radius r to a point p are used to interpolate a local plane π [18]. Thus, the signed distance d to local plane π (DLP) is calculated as the inner product of the vectors $p - \mu$ and $\vec{n}_\pi$: $d(\pi, p) = (p - \mu) \cdot \vec{n}_\pi$ This definition requires a normal which is estimated using the third Eigenvector of the covariance matrix $\Sigma = (X - \mu)(X - \mu)^T$, where μ is the mean vector of X. Using a simple sign check, the normal $\vec{n}_\pi$ always points toward the origin of the camera system, thus d indicates local convexity or concavity, see figure 2.

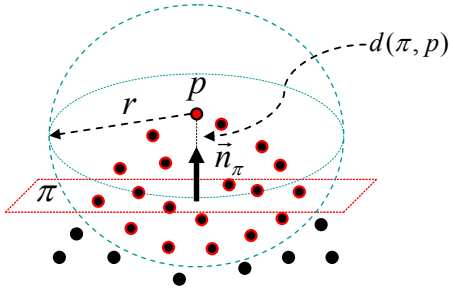

Figure 2. Distance to local plane, d, is a signed value which could indicate convexity or concavity from a common point of view.

B. Eigenshape descriptors for coarse local shape

A coarse local shape descriptor is a vector of the form: $(\lambda_1, \lambda_2, \lambda_3, DLP)$, containing the three Eigenvalues of the local point-cloud covariance matrix Σ and their signed DLP. We call this Coarse Eigen-Shape descriptor (or CES feature) in reference to the coarse shape information provided, see figure 3.

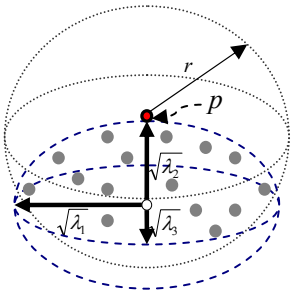

Figure 3. CES can be used to extract coarse shape information about a local surface represented by 3D points, e.g. if the shape encoded is an ellipsoid then $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}$ and $\sqrt{\lambda_3}$ are associated to its axial lengths.

C. Spherically sampled RBF (SSR) descriptor

An SSR shape histogram is a local surface representation derived from an implicit radial basis function (RBF) model, which is in fact a *signed distance to surface function* [17].

The concept of SSR shape histogram is to sample the RBF function in the locality of a candidate facial vertex in a pose-invariant way. Thus, a set of n sample points are evenly distributed across a unit sphere, centred on the origin. The sphere is then scaled by q radii, r_i, to give a set

of concentric spheres and their common centre is translated such that it is coincident with a facial surface point.

The RBF function, s, is then evaluated at the $N = nq$ sample points on the concentric spheres and these values are normalised by dividing by the appropriate sphere radius, r_i, giving a set of values in the range -1 to 1. A $(p \times q)$ SSR histogram is constructed by binning the normalised values $s_n = \frac{s}{r_i}$ over p bins.

D. Spin images

In this representation, each point belonging to a 3D surface is linked to an oriented point on the surface working as the origin [16]. As observed in equations (1), there is a dimension reduction: from 3D coordinates to a 2D system (α, β) which represents the relative distance between the oriented point p and the other points p_i in the surface. A spin image is produced by assigning the spin-map coordinates (α, β) into the appropriate spin-image bins.

$$S_0 : R^3 \rightarrow R^2$$
$$S_0(x) \rightarrow (\alpha, \beta) = \left(\sqrt{\|x - p\| - (\overline{n} \cdot (x - p))^2}, \overline{n} \cdot (x - p)\right) \quad (1)$$

III. Experimental framework

Each of our two main approaches has two variants, giving four systems in all (see table I). In system 1a, we localise the inner eye corners and the nose tip simultaneously, by applying graph matching to simple descriptors (DLP) and 'relaxing' until the best supported combination is obtained. System 1b is a variant of the first one, in which we integrate a local coarse shape feature descriptor (CES) and, like system 1a, we localise the three landmarks simultaneously. Systems 2a and 2b initially localise the nose tip, using feature descriptors in a cascade filter. The inner eye corners are then localised relative to the nose tip.

A. Structural graph matching algorithm

The graph model we fit in systems 1a and 1b is very simple and consists of three nodes and three arcs. 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 pair-wise relationships between a couple of nodes.

To do this we use a structural graph matching algorithm known as 'relaxation by elimination' [4], and in our implementation [18], we divide this into four steps: First, initial candidate lists for each of the three nodes are populated, using the appropriated mean and variance values from training data. Next, binary arrays are created which represent mutual support between two candidate nodes. Then, every least supported 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 computing the Mahalanobis distance in our 6-DOF training feature space. The candidate triplet with the minimum distance is considered to represent the estimated landmark locations and is stored.

TABLE I. Implementations for landmark localisation

	Feature descriptors	Method
System 1a	DLP and Euclidean distances.	Simultaneous localisation using graph matching
System 1b	DLP, Euclidean lengths and CES.	Simultaneous localisation using graph matching
System 2a	DLP, CES, SSR values, spin-images.	Cascade filtering w/spin-images to localise the nose tip. Local graph matching to localise the inner-eye-corners.
System 2b	DLP, CES, SSR features.	Cascade filtering w/SSR histograms to localise the nose tip. Local graph matching to localise the inner-eye-corners.

B. Cascade filtering

Our second localisation method, used in systems 2a and 2b, first localises the nose-tip, and then the inner-eye-corners, again using the trained Euclidean separation of features employed in systems 1a and 1b.

To localise the nose over all vertices is computationally expensive, thus we identify the raw nose tip vertex via a cascade filtering process, as illustrated in figure 4. Essentially this is a decision tree where progressively more expensive operations are employed to eliminate vertices. The constraints (thresholds) employed at each filtering stage are designed to be weak, by examining trained nose feature value distributions, so that the nose tip itself is never eliminated. Conceptually, this amounts to considering every vertex as a candidate nose position, where all but one vertex are 'false positives'. Then, at each stage, we apply a filter to reduce the number of false positives, until we have a small number of candidates at the final stage, at which point our most expensive and discriminating test (spin images and SSR histograms) is used to find the correct vertex.

The feature that we use in filter 1 is distance to local plane (DLP). The filter uses weak thresholding, so that candidates need to be within four standard deviations of the average DLP value for trained noses in order to survive. Local CES features are calculated in filter 2 using a 20 mm radius sphere and, all vertices not within four standard deviations are eliminated. In filter 3, we compute SSR convexity values [17] using a single sphere of radius 20 mm and, again, we set similar weak thresholds. At this stage, we have multiple local maxima in SSR convexity value [18]. We expect the nose to be situated at some local maximum in convexity value, so we find these local maxima and eliminate all other vertices. (This filter, filter 4, will not be useful when we expand to all eleven landmarks in our dataset, but we can adapt the filter stages and thresholds, as necessary for each landmark). Finally we use spin-images (system 2a) or SSR shape histograms (system 2b), by finding

the minimum Mahalanobis distance in the feature space, to select the correct nose vertex from the set of local maxima in SSR convexity value.

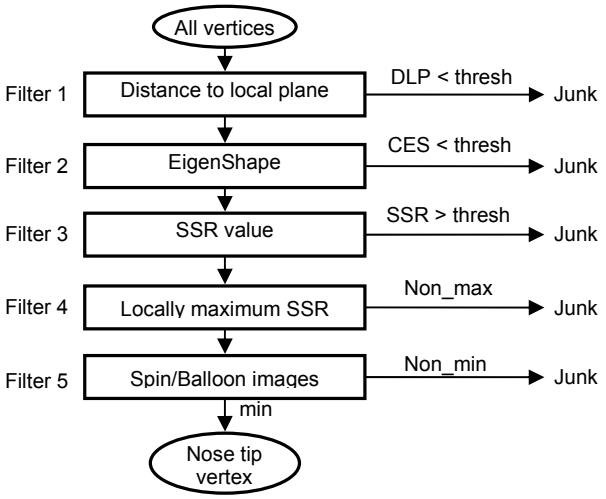

Figure 4. Cascade filter for nose tip detection.

IV. EVALUATION

We have evaluated our landmark localisation systems on the FRGC database [15]. 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 this has an associated intensity image. The files are divided into three subsets, named after their collection periods: Spring-2003, Fall-2003 and Spring-2004.

A. Methodology

We have created four localisation systems for the inner eye corners and the nose tip, each of which uses the same training and testing data. However, they use different feature descriptors (with particular training sets) and apply graph matching in different stages as follows: (1a) Graph matching applied directly. (1b) A graph matching variant which eliminates unlikely features using the CES descriptor before relaxation. (2a) & (2b) localise the nose-tip by cascade filtering followed by local graph matching to localise the eye-corners relative to the nose tip. Our experimental methodology was as follows:

1. For each record in the FRGC database, eleven landmarks (we only use three here) 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 [18].
2. We selected the first 200 subjects from the Spring-2003 subset, which have more than one image in this specific data subset. For each of these persons, we randomly selected a capture to give 200 training 3D images.
3. For each of these 200 training 3D images, SSR shape histograms at the ground-truth nose vertex were constructed, using 8 radii of 10 mm to 45 mm in steps of 5 mm and 23 bins for normalised RBF values. This gave SSR shape histograms of dimension (8×23).
4. For the same 200 training set above, spin-images at the ground-truth nose vertex were calculated: $\alpha_{\max} = 45$, $\beta_{\max} = 45$ and a mesh resolution of 3 mm where used. These α and β parameters, cover an equivalent surface area to the SSR histograms.
5. We evaluated our localisation systems in two scenarios, considering variations in depth and facial expressions. The FRGC database is already divided in this way and we adopted them as they are (see table II). Naturally, there are variations in illumination and small variations in pose.

TABLE II. TESTING SETS FOR EVALUATION PERFORMANCE ANALYSIS

Scenarios	Subset	Size
1. Depth variations, neutral expressions.	Spring2003	509
2. Facial expression variations and few depth variations.	Fall2003	1,507
	Spring2004	1,764

6. We applied principal component analysis (PCA) to reduce the spin-images and SSR histograms feature space dimensionalities.
7. For all nose candidates (filter 4 outputs in the cascade filter) on all test images, we calculated the Mahalanobis distance to the mean of the trained spin-images/SSR-histograms data. For each test image, the vertex with the minimum Mahalanobis distance was identified as the nose and stored.
8. 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 3-5 mm. This has implications on the achievable localisation accuracy. We set a distance threshold (specified in millimetres) and if the RMS error is below this threshold, then we label our result as a successful localisation. This allows us to present a performance curve indicating the percentage of successful feature localisations against the RMS distance metric threshold used to indicate a successful location. These results have the nice property that they are not dependent on a single threshold and, in general, these performance curves show two distinct phases: (i) a rising phase where an increased RMS distance threshold masks small localisation errors, 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. If the plateau is not at 100% success rate, this indicates the presence of some gross errors in landmark localisation. Of course, it is useful to choose some RMS threshold values and quote performance figures (e.g. categorisation in table III). A sensible place to choose the threshold is close to where the graph switches from the rising region to the plateau region.

TABLE III. THRESHOLDS TO EVALUATE ESTIMATED LOCATIONS

Success	$RMS \leq 12mm$
Poor	$12mm < RMS \leq 20mm$
Fail	$RMS > 20mm$

B. Results

The graphs on the top row of figure 5 show eye localisation performance (left and right averaged), and the bottom row shows nose localisation performance. These results were generated by averaging the results from the three data sets presented in table II. Table IV summarises localisation performance, where success is defined according to an RMS error threshold of 12 mm (see table III). System 2b clearly gives the most successful localisation performance.

TABLE IV. SUCCESSFUL LOCALISATION SUMMARY

		Scenario #1	Scenario #2	
		Spring-03	Fall-03	Spring-04
System 1a	Eyes	77.99 %	90.04 %	90.02 %
	Nose	62.47 %	74.19 %	74.26 %
System 1b	Eyes	87.03 %	96.08 %	95.91%
	Nose	73.47 %	86.39 %	86.22 %
System 2a	Eyes	92.33 %	75.77 %	72.67 %
	Nose	97.64 %	78.63 %	76.98 %
System 2b	Eyes	93.90 %	96.15 %	96.82 %
	Nose	99.41 %	99.60 %	99.77 %

We know that vertex-based spin-images are mesh resolution dependent, also, that the FRGC database was populated using different depths varying the number of vertices of the shape images. These facts could affect the performance if adequate training is not considered. Thus, in order to verify this assumption, we select 66 representative 3D shape files from different people (from the Spring-03 subset and different to the testing set) counting 33 scans for both middle and far focus. After that, training spin-images and SSR histograms were computed and they were used in filter 5 of the cascade filter. This modification produces new systems which we call: system 2a* and system 2b*. An improvement in spin-images' performance was obtained, but it is still lower than SSR histograms' performance (see table V).

C. Comparison with other methods

Although a quantitative comparison with other methods in the literature is out of the scope of this paper, it is clear (from table VI) that SSR descriptors (system 2b) give an excellent performance, considering that this landmark localisation approach uses only 3D pose invariant feature descriptors in extensive experimentation with the FRGC database.

TABLE V. SUCCESSFUL LOCALISATION USING A REPRESENTATIVE TRAINING SET

		Scenario #1	Scenario #2	
Descriptor	Landmark	Spring03	Fall03	Spring04
Spin images (System 2a*)	Eyes	91.94 %	91.04 %	90.02 %
	Nose	97.83 %	94.82 %	93.87 %
SSR histograms (System 2b*)	Eyes	91.94 %	94.35 %	95.80 %
	Nose	98.82 %	99.27 %	99.65 %

TABLE VI. PERFORMANCE COMPARISON

Approach	Database	Size	Eye corners	Nose tip
Xu et al. [5]	3DPEF	280	n/a	99.30%
Conde et al. [2]	FRAV3D	714	98.00%	99.50%
Mian et al. [12]	FRGC	4,950	n/a	98.30%
Segundo et al. [6]	FRGC	4,007	99.83%	99.95%
SSR features	FRGC	4,013	96.82%	99.77%

V. CONCLUSIONS

We have compared our graph matching and cascade filter approaches to localise the inner-eye corners and the nose-tip. Four systems have been implemented and evaluated for that purpose. System 1a presents a poor performance, because our graph matching approach is looking for the global maxima (the best supported triplet) and only simple descriptors (DLP and Euclidean lengths) are been used. An increase in performance is observed in system 1b when the coarse eigenshape (CES) descriptor is applied before relaxation. A significant improvement is achieved when the nose-tip is located robustly; and this is possible using a set of pose-invariant feature descriptors embedded in a cascaded filter. The inner-eye-corners' localisation performance is lower than that of the nose-tip, because they are located using only simple features and graph matching.

We found that SSR histograms outperformed spin-images even with a more representative training set designed to improve spin-image classification performance. SSR histograms appear to be more immune to mesh resolution.

Our future work includes extending not only the application of our feature descriptors, but also our localisation approaches to the eleven facial landmarks that exist in our ground truth of the 3D FRGC database.

ACKNOWLEDGMENT

M. R. is supported by the Mexican Council of Science and Technology (CONACYT, grant 207690) with assistance of the Mexican Department of Public Education (SEP) and Government.

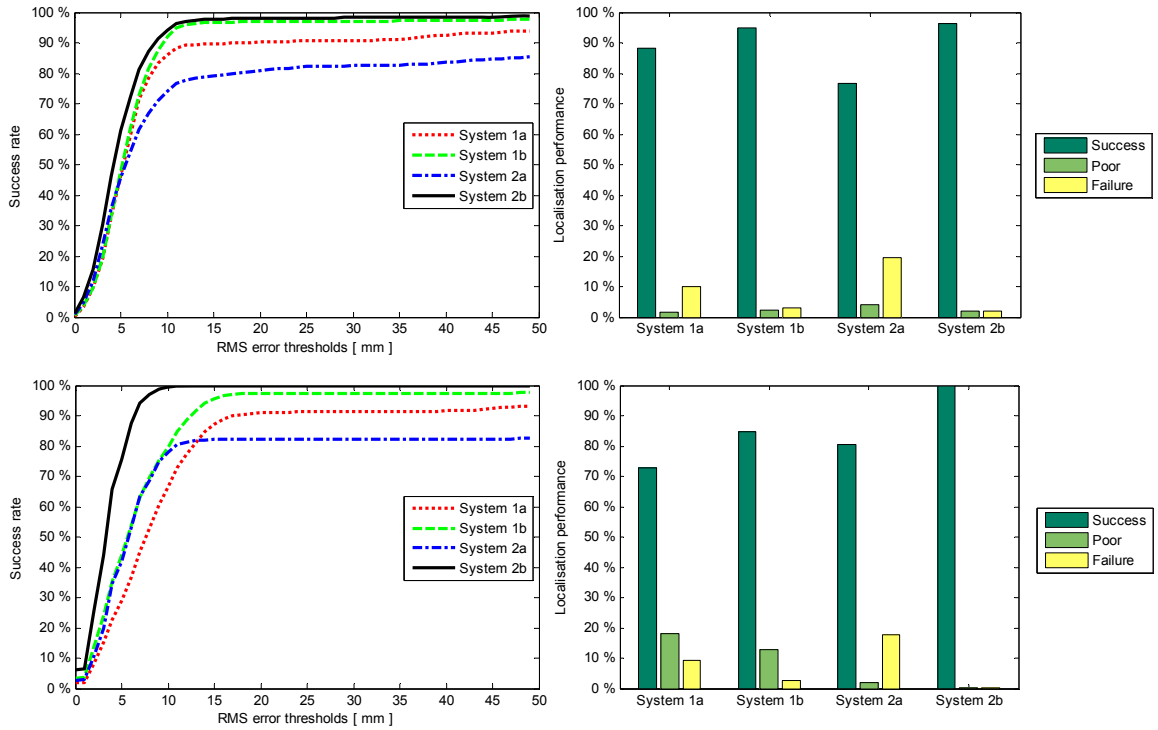

Figure 5. Averaged eye localisation performance (top row) and nose localisation performance (bottom row) using the four systems in table I. A randomly selected training set of 200 images of different people was used. The bar charts show categorisation of performance according to table III.

REFERENCES

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

[2] C. Conde, L. J. Rodriguez, E. Cabello, "Automatic 3D Face Feature Points Extraction with Spin Images", Proc. International Conference on Image Analysis and Recognition, LNCS 4142, 317 – 328, 2006.

[3] M. Turner and J. Austin, "Graph Matching by Neural Relaxation", Neural Computing & Applications, vol. 7 (3), Sep. 1998, 238 – 248.

[4] W. Zhao and R. Chellappa, Face Processing: Advanced Modeling and Methods, Elsevier, USA, 2006.

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

[6] M. Segundo, C. Queirolo, O. Bellon and L. Silva, "Automatic 3D Facial Segmentation and Landmark Detection", Proc. 14th Int. Conf. on Image Analysis and Processing, 2007, 431 – 436.

[7] X. Lu and A. Jain. "Automatic Feature Extraction for Mltiview 3D Face Recognition", Proc. 7th International Conference on Automatic Face and Gesture Recognition, 2006.

[8] D. Colbry, G. Stockman, and A. Jain, "Detection of Anchor Points for 3D Face Verification", Proc. Computer Vision and Pattern Recognition, 2005.

[9] S. Gupta, K. Aggarwal, M. Markey and A. Bovik, "3D Face Recognition Founded on the Structural Diversity of Human Faces", Proc. Computer Vision and Pattern Recognition, 2007.

[10] K. Chang, K. Bowyer and P. Flynn, "Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28 (10), Oct. 2006.

[11] A. Salah and L. Akarun, "3D Facial Feature Localization for Registration", International Workshop on Multimedia Content Representation, Classification and Security, 2006.

[12] A. Mian, M. Bennamoun and R. Owens, "An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29 (11), 2007, 1927 – 1943.

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

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

[15] P. J. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, "Overview of the Face Recognition Grand Challenge", Proc. Computer Vision and Pattern Recognition, vol. 1, 2005, 947- 954.

[16] A. Johnson and M. Herbert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes", Trans. Pattern Analysis and Machine Intelligence, vol. 21, number 5, 1997, 433 – 449.

[17] N. Pears, "RBF Shape Histograms and Their Application to 3D Face Processing", Proc. 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2008.

[18] M. Romero and N. Pears, "3D Facial Landmark Localisation by Matching Simple Descriptors", Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems, 2008.

[19] A. Bronstein, M. Bronstein and R. Kimmel, "Three-Dimensional Face Recognition", International Journal of Computer Vision, vol. 64 (1), Aug. 2005.