Comparison of geometric and texture-based features for facial landmark localization in 3D

Luke Gahan[1], **Federico M. Sukno**[2,1] **and Paul F. Whelan**[1]
[1]Centre for Image Processing & Analysis
Dublin City University, Dublin, Ireland
[2]Department of Information and Communication Technologies
Pompeu Fabra University, Barcelona, Spain

Abstract

The localisation of facial landmarks is an important problem in computer vision, with applications to biometric identification and medicine. The increasing availability of three-dimensional data allows for a complete representation of the facial geometry, overcoming traditional limitations inherent to 2D, such as viewpoint and lighting conditions. However, these benefits can only be fully exploited when the processing concentrates purely on the geometric information, disregarding texture. This fact is particularly interesting when addressing the localisation of anatomical landmarks, as it is not clear to date whether geometric information can be used to fully replace texture (e.g. the localisation of the eye corners and the lips is believed to be strongly linked to texture clues).

In this paper we present a quantitative study of 3D landmark localization based on geometry, texture or a combination of both, integrated in a common framework based on Gabor filters that has reported state of the art results. We target 10 facial landmarks and find that, while the algorithm performs poorly for the nose tip with a mean 3D error of 6.15mm, the remaining landmarks are all localised with an error under 3.35mm, with the outer eye corners and mouth corners performing particularly well. Interestingly, geometry and texture achieved comparable results for the inner eye corners and mouth corners, while texture clearly outperformed geometry for the outer eye corners.

1 Introduction

Facial landmark localisation is the primary step in a number of computer vision systems including facial recognition, facial pose estimation, medical diagnostics and multimedia applications. Historically most landmark localisation algorithms have used standard 2D images. Such systems, no matter how accurate, are always going to be limited by the fact that they are operating on dimensionally reduced representations of 3D objects. A significant amount of extra information about the human face is contained in the 3D spatial dimension.

A number of different approaches have been taken with regard to localising facial landmarks in 3D images. Geometry based techniques have received a good deal of attention. Segundo et al. present an effective system which uses surface classification techniques in order to localise landmarks [Segundo et al., 2010]. The authors record a 3D localisation error of under 10mm for 90% of images in their test set. Creusot et al. combine machine learning and a large number of geometric techniques in their system [Creusot et al., 2013]. The authors note that while this system does not outperform others in terms of accuracy, it does perform quite well in terms of robustness. Since the algorithm used is not sequential in nature, a failure to detect certain landmarks does not influence the localisation of subsequent landmarks. This system provides a framework for landmark localisation and leaves potential for future improvement.

Zhao et al. present a statistical model based approach in [Zhao et al., 2011]. This system works well in challenging situations where there is facial occlusion and/or very expressive faces. This system learns the spatial relationships between different landmarks and uses this in conjunction with local texture and range information. The authors use Principal Component

Analysis (PCA) to create a statistical facial feature map. This is essentially a combination of individual geometry (landmark coordinates), shape (range images) and texture (texture images) models. The authors report a mean 3D error rate of below 5.07mm for all 15 facial landmarks.

Perakis et al. use local shape descriptors to localise facial landmarks [Perakis et al., 2013, Passalis et al., 2011]. These local shape descriptors characterise the shape profile at a given landmark. By evaluating the shape index at a landmark in a number of training images a model can be constructed. These descriptors are generated by examining the principal curvature and spin image at a landmark. A facial landmark model is then created. This is used to constrain the relative locations of detected landmarks. Models are also created for the left and right hand side of the face. These are used to deal with profile or semi-profile faces. The systems achieves relatively good results with a mean 3D error of below 5.58mm for all 8 targeted landmarks.

One particular approach which has received increased attention in recent years is the use of Gabor filters for facial landmark localisation [Movellan, 2002]. Jahanbin et al. use Gabor filter banks for landmark localisation in [Jahanbin et al., 2008]. This technique implements the same landmark localisation procedure as Wiscott et al. used in their Elastic Bunch Graph Match system (without the elastic constraint) [Wiskott et al., 1997]. While the authors do not present in depth results in this particular paper, it does serve as a basis for later work carried out by the same research group [Gupta et al., 2010b]. This particular system combines curvature detection, Gabor filters and expert knowledge of the human face to localise landmarks using anthropometric information based on the work carried out by Farkas et al. in the medical field [Farkas and Munro, 1987]. This information plays a vital role in establishing a sensible search region which is then examined to further improve the accuracy of localisation.

An interesting element of the work by Gupta et al. [Gupta et al., 2010b] is that Gabor filters are applied to both range and texture and their framework allows for a direct integration of both sources of information. However, the authors did not provide a detailed analysis of this aspect and results were limited to 2D standard deviation errors, which hampers a thorough comparison to other approaches. In this work we present a quantitative analysis of landmark localization errors when using texture, range or both sources of information at the same time. We use the framework developed by Gupta et al. and reproduce the results reported originally, which allows to also calculate the mean 3D error to make results comparable to related work. We find that the inclusion of both texture and range information always yields the best results, although the benefit of range was negligible in some cases. Interestingly, for the inner eye corners and mouth corners the error results were similar for all three tested alternatives.

2 Automatic Landmark Localisation Using Anthropometric Information

The landmark localisation procedure carried out remains as faithful as possible to the method developed by Gupta et al. [Gupta et al., 2010b]. Generally speaking the algorithm first uses curvature information to detect an approximate location for a particular landmark. Using anthropometric information a search region is defined around this approximation and the position is then refined using as described below. The 10 landmarks localised are the nose tip, with points and root center, inner and outer eye corners and mouth corners.

Nose Tip (prn): The Iterative Closest Point (ICP) algorithm is used to register each face in the database to a frontal template face. These aligned images are used in all subsequent steps. Once all images have been aligned the manually localised tip of the template face is taken as an approximate location for tip of the nose in all images. A window of 96 mm x 96mm is then defined around this approximated nose tip. Since all faces have been frontally aligned, the actual nose tip is present in this large window for all cases. This means that the method is not fully automated since it relies on the manually localised tip of the template face.

It has been observed that the Gaussian surface curvature of the tip of the nose is distinctly elliptical (K >0,) [Moreno et al., 2003, Segundo et al., 2010, Creusot et al., 2013]. For this rea-

son the Gaussian surface curvature (σ = 15 pixels) is evaluated within the search region about the nose tip approximation. The maximum Gaussian curvature within the region is taken as final location of the nose tip (prn).

Nose Width Points (al-al): These points are localised by first defining a search region around the detected nose tip. The size of this window (42 mm x 50 mm) is defined based on the mean and standard deviation values published by Farkas [Farkas and Munro, 1987]. A Laplacian of Gaussian edge detector (σ = 7 pixels) is then used within this region. Moving in a horizontal direction from the nose tip, the first edge encountered is considered to be the nose contour and is retained. Then, points of negative curvature are detected by generating an unwrapped chain code for the nose contour and using a derivative of Gaussian filter on this one dimensional signal to detect points of critical curvature [Rodriguez and Aggarwal, 1990]. Nose width points are finally selected from the critical points immediately above and below the vertical coordinate of the nose tip. The widest of these are selected as nose width points.

Inner Eye Corner (en-en) & Center of Nose Root (m'): A search region for the left and right inner eye corners is defined using the location of the detected nose tip and nose width points. The vertical limit defined based on the fact that for the average adult, the distance between inner eye corners and the tip of the nose in the vertical direction is 0.3803 times the distance between the tip of the nose and the top point of the head [Farkas and Munro, 1987, Gupta et al., 2010b]. Gupta et al. allow for variations in the measure by setting the upper vertical limit at $(prn_y + 0.3803 \times 1.5|prn_y - V_y|)$, where V_y is the Y coordinate of the highest vertical point in the 3D model. The horizontal limit is obtained by using the locations of the nose width points and the nose tip. Specifically, horizontal limits are defined from the nose tip to $al_{x,left/right} \pm 0.5|al_{x,left} - al_{x,right}|$ for the left and right inner eye corners.

The Gaussian curvature within this region is evaluated and the location of maximum curvature is used as an approximation for the location of the inner eye corner ($\sigma = 15$ pixels). Finally a region of 20mm x 20mm is defined around this peak of Gaussian curvature.

The location of inner eye corners are then refined with a modified version of the EBGM technique [Jahanbin et al., 2008, Wiskott et al., 1997]. In brief, this technique involves comparing the Gabor coefficients generated for each pixel in the search region with the coefficients for the landmarks of 89 training images. These 89 images consist of neutral and expressive faces. The images are selected in an attempt to cover as much feature variance as possible (i.e. closed/open mouth and eyes). 80 Gabor coefficients (known as a Gabor jet) are generated at each landmark for each of the example images. A filter bank of 40 Gabor filters is used (5 scales x 8 orientations). 40 coefficients are generated for both range (3D) and texture (2D) images. While the specific parameters of these filters are not provided in [Gupta et al., 2010b], we used the filter bank outlined in by Wiscott et al. [Wiskott et al., 1997]. Note that, for the database used, all images should be scaled by $\frac{1}{3}$ when Gabor filtering is applied.The final location of the inner eye corner is obtained by finding the pixel which has a Gabor jet most similar to that of any training landmark. The similarity score is given in equation (1):

$$S(\overrightarrow{J}, \overrightarrow{J'}) = \frac{\sum_{i=1}^{40/80} a_i a'_i \cos(\Phi_i - \Phi'_i)}{\sqrt{\sum_{i=1}^{40/80} a_i^2 \sum_{i=1}^{40/80} a_i'^2}} \tag{1}$$

where J and J' are the jets to be compared, defined as $J_j = a_j\, e^{i\phi_j}$. Where a is the magnitude and ϕ is the phase of the Gabor coefficient at a given pixel.The jets contain either 40 or 80 coefficients depending on which form of EBGM is to be used. Gupta et al. chose to use 2D and 3D Gabor coefficients. In this work 2D, 3D and 2D+3D results are compared. The center of the nose root is determined by finding the mid-point between the two inner eye corners.

Outer Eye Corners (ex-ex): A search region for the outer eye corners is defined based on the location of the detected inner eye corners as per [Gupta et al., 2010b]. This 20 x 34 mm region is evaluated using the same search procedure as used for the inner eye corners. Gupta et al.

chose to use 2D EBGM search as the outer eye corner region does not have distinct enough curvature characteristics. In this work all three EBGM techniques are evaluated.

Mouth Corners (ch-ch): The lip curvature is examined in order to determine a search region for the mouth corners. The Gaussian curvature of both the upper and lower lips is elliptical in nature. The regions immediately above the upper lip and below the lower lip are hyperbolic ($K < 0$). These properties can be used to define upper and lower search limits for the mouth corners. The horizontal limits are defined by $[(al_{x,left} - 0.7|al_{x,left} - al_{x,right}|), (al_{x,left})]$ for ch_{left} and analogously for ch_{right}. In order to remove noise a certain amount of smoothing must be carried out when calculating Gaussian curvature. In some cases the Gaussian curvature of the upper or lower lip is too weak and cannot be localised. In such cases the troughs in Gaussian curvature immediately above and below the lip region are used as limits. While these are usually stronger features than the lips, errors can arise when searching for peak mean curvature in the next stage of the algorithm as there is a high mean curvature along the jaw line.

The mean curvature ($\sigma = 2$ pixels) is then calculated for the defined search region. Since the mouth corners are regions of high mean curvature the peak curvature value in this region is taken as an estimate for of the mouth corner. A 30mm x 11mm search region is defined around these mouth corner estimates. The same EBGM procedure used to localise the eye corners is also used to precisely localise the mouth corners. Gupta et al. chose to use 2D+3D EBGM. In this work 2D, 3D and 2D+3D EBGM results are compared.

3 Experimental Results & Discussion

3.1 Test Data

The performance of the landmark localisation algorithm is evaluated using the Texas 3DFR database [Gupta et al., 2010a]. It contains high resolution (751 x 501 pixels, 0.32 mm per pixel) pairs of portrait and range images from 118 healthy adult subjects. 25 facial landmarks have been manually located. Both range and portrait images were acquired simultaneously using a regularly calibrated stereo vision system and the data was filtered, interpolated and smoothed to remove impulse noise and large holes [Gupta et al., 2010a]. From the 1149 portrait-range pairs of the database, 89 were used in the EBGM search and the remaining 1060 were used as test data.

3.2 Landmark Localisation Results

The landmark localisation results obtained for the Texas 3DFR database are given in Table 2. All results are given in millimetres. As mentioned previously Gupta et al. do not provide 3D error results [Gupta et al., 2010b]. Thus, we compared our results to the ones originally provided, in terms of 2D standard deviation and confirmed that our implementation a faithfully reproduced the original method (Table 1).

The mean error result of the nose tip is noticeably larger than the localisation of the other landmarks. On closer examination it appears that in all cases the detected nose tip is above the manually localised nose tip (in the Y direction). This can clearly be seen in the boxplot in Figure 1. This figure shows clearly that the median value for the X error is 0mm as expected in a normal error distribution. The Y distribution is extremely skewed to one side of the manually localised nose tip (a negative Y error is above the manual location for an upright face). Since the standard deviation of the Y error is relatively small it seems that the issue is that the peak of Gaussian curvature does not correspond to the same location the manual annotators have identified as the nose tip.

The mean error results obtained for the nose width points are reasonable while the standard deviations are impressive, especially when using the modified EBGM technique. A 3D mean error of under 2mm is recorded for both inner eye corners. The outer eye corners which are slightly more difficult to localise are detected with a mean error of under 2.6mm. A mean

	X std. dev (mm)		Y std. dev (mm)		2D std. dev (mm)	
Landmark	**Gupta**	**This Method**	**Gupta**	**This Method**	**Gupta**	**This Method**
PRN	1.045	0.766	1.680	1.714	1.978	1.705
AL Left	0.721	0.647	1.655	0.710	1.805	0.739
Al Right	0.798	0.546	1.646	0.814	1.829	0.818
EN Left	1.488	1.249	1.245	0.908	1.940	1.363
EN Right	1.354	1.378	1.344	0.792	1.908	1.417
M’	1.355	1.415	1.811	1.010	2.261	1.417
EX Left	1.795	1.727	1.285	1.047	2.208	1.850
EX Right	2.126	1.940	1.384	1.248	2.537	2.149
CH Left	1.948	1.749	0.933	1.692	2.160	2.321
CH Right	1.976	1.429	1.045	0.844	2.235	1.460

Table 1: Error standard deviation results comparison with Gupta et al. [Gupta et al., 2010b]

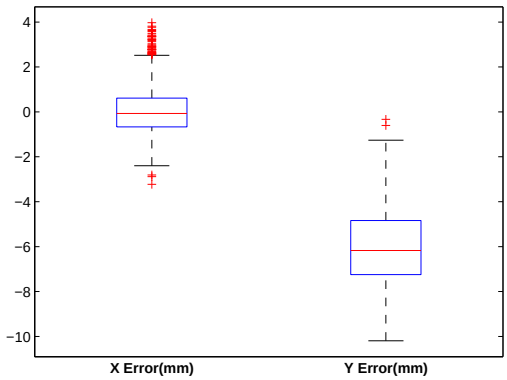

Figure 1: Vertical Prn Error Bias

error of below 2.16mm is achieved for both mouth corners. The algorithm does have particular difficultly with faces where facial hair is present. This is as expected when using Gabor filters as there is a significantly different response to a Gabor filter when facial hair is present.

One interesting point to note is that the three worst results obtained are for the three landmarks localised using techniques which do not involve training. The training stage of EBGM uses manual landmark locations. This means that when EBGM is used, the algorithm searches for a location on an unknown image which is most similar to the training data, which is based on manual locations. For the nose tip and width points the algorithm searches for a particular image feature (e.g. maximum Gaussian curvature) which is said to be present at that landmark. Perhaps using EBGM for all landmarks might yield better performance. Another possible issue could be marker bias. No details are provided about how many annotators are used but using separate annotators for test and training data could be a possible solution.

Landmark	Prn	Al_L	Al_R	En_L	En_R	M’	Ex_L	Ex_R	Ch_L	Ch_R
3D mean	6.15	3.35	3.31	1.82	1.75	2.76	2.48	2.59	2.16	2.02
3D stdev	1.75	1.65	1.88	1.50	1.52	1.59	2.58	2.99	3.04	2.15

Table 2: Landmark localisation error

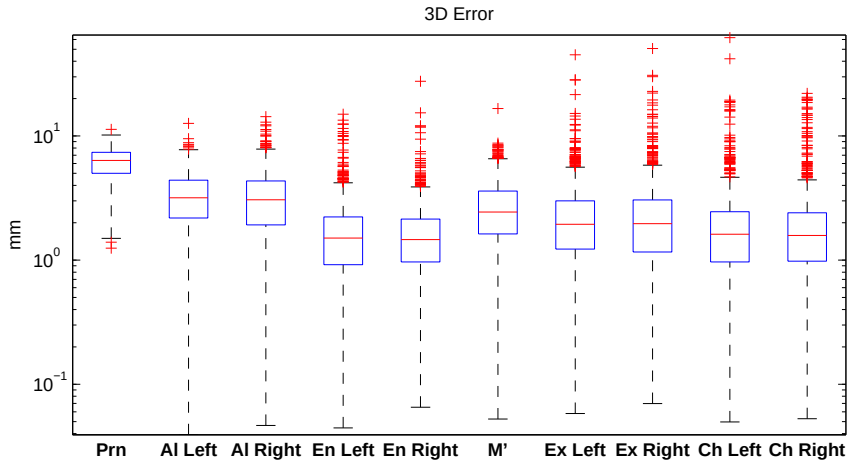

Figure 2: 3D error boxplot

3.3 Texture & Range Comparison

The inner eyes and outer mouth corners are detected using 2D + 3D EBGM while 2D EBGM is used for the outer eye corners. The same similarity metric is used in each case (1) with the only difference being the coefficients examined.

Landmark	2D EBGM	3D EBGM	2D+3D EBGM
En Left	1.83 ± 1.53	2.18 ± 1.70	1.82 ± 1.50
En Right	1.75 ± 1.55	1.99 ± 1.58	1.75 ± 1.52
Ex Left	2.48 ± 2.58	5.10 ± 5.28	2.39 ± 2.12
Ex Right	2.59 ± 2.99	8.91 ± 7.22	2.49 ± 2.27
Ch Left	2.20 ± 2.83	2.54 ± 2.89	2.16 ± 3.04
Ch Right	2.15 ± 2.44	2.20 ± 1.61	2.02 ± 2.15

Table 3: 2D, 3D & 2D+3D EBGM comparison, in terms of 3D error (mean $\pm$ std. dev.)

Interestingly, Table 3 shows that for the inner and outer eye corners the inclusion of range coefficients improves localisation results. Gupta et al. use 2D + 3D for the inner eye corner while they choose to use just 2D for the outer eye corners. The results obtained here suggest that a similar improvement in localisation could be achieved with the inclusion of range information. While it is clear that just using 3D information results in poor localisation performance it should be noted that the 3D information only influences the result of localisation when a 3D coefficient is more similar to one of the training image coefficients than any of the 2D coefficients. This means that in some individual cases the inclusion of 3D information may adversely affect localisation but for the entire database the average error is reduced.

With regard to the mouth corners the use of texture and range information results in the best mean error performance. This is the same as the behaviour for the other landmarks. Once again the worst mean error is recorded when just range information is used.

It is clear that in all cases examined the inclusion of more information (texture & range) in the EBGM stage results in better overall localisation. This suggests that the similarity score and the procedure Gupta et al. use for choosing the landmark location works quite well. It suggests that in the majority of cases the inclusion of extra information leads to enhanced localisation performance. Obviously there is a computational overhead to be considered when including this extra information but in cases where speed isn't an issue it seems that the inclusion of 2D and 3D information leads to the best localisation performance.

Since the 2D and 3D EBGM techniques are directly comparable, Table 3 shows that for all landmarks examined texture information yields better results. Though for the inner eye corners and mouth corners this difference is quite small.

4 Conclusion

We have shown that the method developed by Gupta et al. achieves state of the art landmark localisation results. The one weak point is the localisation of the nose tip which is quite poor. Even though the localisation of the tip is poor it does not appear to adversely affect the localisation of subsequent landmarks where the location of the nose tip is used to define a search region. Another better performing method, such as that used by Segundo et al., could perhaps be used for the localisation of the nose tip [Segundo et al., 2010].

It was determined that for the EBGM stage, the inclusion of both texture and range information yields the best results. Interestingly, for the inner eye corners and mouth corners the error results recorded are similar for each of the EBGM methods. For the outer eye corner 3D EBGM performed quite poorly, with 2D and 2D+3D obtaining similar results. This suggests that for outer eye corner detection, 2D EBGM could be used without a significant ($\sim$ 0.3mm) decrease in mean error.

Acknowledgments

The authors would like to thank their colleagues in the Face3D Consortium (www.face3d.ac.uk), and financial support from the Wellcome Trust (WT-086901 MA) and the Marie Curie IEF programme (grant 299605, SP-MORPH)..

References

[Creusot et al., 2013] Creusot, C., Pears, N., and Austin, J. (2013). A machine-learning approach to keypoint detection and landmarking on 3D meshes. *Int. J. Comput. Vis*, 102(1-3):146–179.

[Farkas and Munro, 1987] Farkas, L. G. and Munro, I. R. (1987). *Anthropometric facial proportions in medicine*. Charles C. Thomas Publisher.

[Gupta et al., 2010a] Gupta, S., et al. (2010a). Texas 3D face recognition database. In *Proc. SSIAI 2010*, pages 97–100.

[Gupta et al., 2010b] Gupta, S., Markey, M. K., and Boxxxvik, A. C. (2010b). Anthropometric 3D face recognition. *Int. J. Comput. Vis*, 90(3):331–349.

[Jahanbin et al., 2008] Jahanbin, S., Bovik, A. C., and Choi, H. (2008). Automated facial feature detection from portrait and range images. In *Proc. SSIAI 2008*, pages 25–28.

[Moreno et al., 2003] Moreno, A. B., et al. (2003). Face recognition using 3D surface-extracted descriptors. In *Proc. IMVIP 2003*.

[Movellan, 2002] Movellan, J. R. (2002). Tutorial on gabor filters. *Open Source Document*.

[Passalis et al., 2011] Passalis, G., et al. (2011). Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1938–1951.

[Perakis et al., 2013] Perakis, P., et al. (2013). 3d facial landmark detection under large yaw and expression variations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1552–1564.

[Rodriguez and Aggarwal, 1990] Rodriguez, J. J. and Aggarwal, J. (1990). Matching aerial images to 3D terrain maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(12):1138–1149.

[Segundo et al., 2010] Segundo, M. P., et al. (2010). Automatic face segmentation and facial landmark detection in range images. *IEEE Trans. Syst., Man, Cybern. B*, 40(5):1319–1330.

[Wiskott et al., 1997] Wiskott, L., et al. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779.

[Zhao et al., 2011] Zhao, X., et al. (2011). Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model. *IEEE Trans. Syst., Man, Cybern. B*, 41(5):1417–1428.