3D Facial Feature Localization for Registration

Albert Ali Salah and Lale Akarun

Boğaziçi University
Perceptual Intelligence Laboratory
Computer Engineering Department, Turkey
{salah, akarun}@boun.edu.tr

Abstract. Accurate automatical localization of fiducial points in face images is an important step in registration. Although statistical methods of landmark localization reach high accuracies with 2D face images, their performances rapidly deteriorate under illumination changes. 3D information can assist this process by either removing the illumination effects from the 2D image, or by supplying robust features based on depth or curvature. We inspect both approaches for this problem. Our results indicate that using 3D features is more promising than illumination correction with the help of 3D. We complement our statistical feature detection scheme with a structural correction scheme and report our results on the FRGC face dataset.

1 Introduction

Automatic face recognition traditionally suffers from pose, illumination, occlusion and expression variations that effect facial images more than changes due to identity. With the emergence of 3D face recognition as a supporting modality for 2D face recognition, there is renewed interest in robust detection of facial features. Facial feature localization is an important component of applications like facial feature tracking, facial modeling and animation, expression analysis, face recognition and biometric applications that rely on 2D and 3D face data. Especially deformation-based registration algorithms require a few accurate landmarks (typically the nose tip, eye and mouth corners, centre of the iris, tip of the chin, the nostrils, the eyebrows) to guide the registration. The aim in landmark detection is locating selected facial points with the greatest possible accuracy.

The most frequently used approach in the detection of facial landmarks is to devise heuristics that are experimentally validated on a particular dataset [3,8,9]. For instance in 3D, the closest point to the camera can be selected as the tip of the nose [4,17]. This method will sometimes detect a streak of hair or tip of the chin as the nose, but depending on the dataset, it may produce better results than any statistical method we can devise. However, its value is limited as a method of pattern recognition, as it cannot be used in any other application or in 2D facial feature localization. In 2D, contrast differences in the eye region are used to detect the eyes [10,17]. The assumption that the eyes are open for detection

B. Gunsel et al. (Eds.): MRCS 2006, LNCS 4105, pp. 338–345, 2006.

can easily be violated, especially when there is simultaneous 3D acquisition with a laser scanner.

The second approach is the joint optimization of structural relationships between landmark locations and local feature constraints, which are frequently conceived as distances to feature templates [14,16]. The landmark locations are modeled with graphs, where the arcs characterize pairwise distances. In [16] local features are modeled with Gabor jets, and a template library (called *the bunch*) is exhaustively searched for the best match at each feature location. A large number of facial landmarks (typically 30-40) are used for graph based methods. Fewer and sparsely distributed landmarks do not produce a sufficient number of structural constraints.

A third and recent approach in 2D facial feature localization is the adaptation of the popular Viola-Jones face detector to this problem [2,6]. In this approach, patches around facial landmarks are detected in the face area with a boosted cascade of simple classifiers based on Haar wavelet features [15]. This approach is used for the coarse-scale detection, as a substitute for manual initialization.

There are very few techniques proposed in the literature to locate facial landmarks using 3D only. In [5], spin images are used with SVM classifiers to locate the nose and the eyes. This is a costly method, and the search area has to be greatly constrained by using prior face knowledge. In [8], the symmetry axes of the face and two planes orthogonal to it are used to locate the eye and mouth corners. In [9], the mean and Gaussian curvatures are combined with the first and second order derivatives of the range image to identify critical points of the face image. Several heuristics were listed for each type of landmark, with promising results. However, our preliminary studies with this approach indicate that robust curvature features require extensive pre-processing that comes with a high computational cost. Furthermore, the large number of false positives suggests that other features and methods should be used in assistance to obtain conclusive results.

Other approaches use 3D in conjunction with 2D [3,4,7]. In [4] 3D shape indices are used with 2D Harris corners to train statistical models for landmarks. In [3], 3D distances between facial points are used to constrain landmark search areas and to clear the background clutter. In [7], it was shown that 2D methods with 3D support can produce good results under relatively stable illumination conditions.

In this paper, we follow a statistical modeling approach for landmark localization that treats each landmark uniformly and independently. Mixture models are used instead of feature templates to make the system scalable (Section 2). Our 2D coarse localization method proposed in [7,13] is contrasted with a similar 3D method and a 3D assisted 2D method. We use a structural correction scheme that detects and corrects erroneous landmarks (Section 3). We evaluate our scheme on FRGC dataset, and report our results in Section 4, followed by our conclusions and indications of future work in Section 5.

2 Unsupervised Local Feature Modeling

The method we propose is based on unsupervised modeling of features sampled from around the landmark locations in the training set. We use mixtures of factor analyzers (MoFA) as an unsupervised model. MoFA is in essence a mixture of Gaussians where the data is assumed to be generated in a lower-dimensional manifold. For each component of the mixture, the $(d \times d)$ covariance matrix Σ is generated by a $(d \times p)$ dimensional factor matrix Λ and a $(d \times d)$ diagonal matrix Ψ:

$$\Sigma_j = \Lambda_j \Lambda_j^T + \Psi \tag{1}$$

Ψ is called the uniqueness, and it stands for the independent variance due to each dimension. With a small number of factors (represented with p) the model will have a much smaller number of parameters than a full Gaussian, even though the covariances are modeled. This is important, because a large number of parameters calls for an appropriately large training set.

In the mixture model, each distribution is potentially multimodal: We fit an arbitrary number of factor analysis components to each feature set. To determine the number of components in the mixture and the number of factors per component, we use an incremental model [12]. The IMoFA-L algorithm adds components and factors to the mixture one by one while monitoring a separate validation set for likelihood changes. With this approach, the number of parameters is automatically adapted to the complexity of the feature space.

2.1 3D Model

The preprocessing of the depth map (or range image) is accomplished by eroding and dilating it with a diamond-shaped structural element to patch the holes and to eliminate the spikes. After downsampling 480×640 range images to 60×80, a z-normalization is applied to depth values of valid pixels. 7×7 neighbourhoods are cropped from around each landmark. These 49-dimensional features are min-max normalized before modeling with MoFA.

In the test phase, the model produces a likelihood map of the image for each landmark. Working on the downsampled images, we determine the landmark locations on the coarse level. This may later be complemented with a fine-level search for greater accuracy.

2.2 2D Model

We use a 2D Gabor wavelet-based method that was shown to have a good accuracy for comparison [7]. In our 2D localization scheme, for each of cropped landmark patch, Gabor wavelets in eight orientations and a single scale are applied (See [13] for more details). Using more scales or neighbourhoods larger than 7×7 did not increase the success rate. 49-dimensional vectors obtained from each Gabor channel are min-max normalized and separately modeled with MoFA. The likelihood maps computed for each channel are summed to a master map to determine the most likely location for the landmark. In [13], we have

shown that this model is more powerful in determining the local similarity than the more traditional Gabor-jet based methods [16], producing large basins of attraction around the true landmark.

2.3 3D Assisted 2D Model

Under a Lambertian illumination assumption, we can use 3D surface normals to remove illumination effects from the 2D images. Basri and Jacobs have shown that projecting to a 9-dimensional subspace spanned by the first spherical harmonics adequately represents an arbitrary illumination on the object [1]. We use the algorithm presented in [18] to recover the texture image (*the albedo*) with spherical harmonics (See Figure. 1). From the albedo image, 7×7 patches are cropped and modeled with MoFA, as in the previous section.

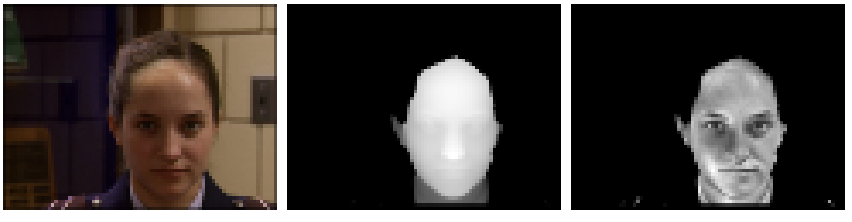

Fig. 1. (a) 2D intensity image. (b) 3D depth image. (c) Recovered albedo.

3 Structural Analysis Subsystem

To make the system robust to occlusions and irregularities, we have opted for independent detection of all landmarks. Therefore, we need to take into account that some of the landmarks may be missed. The purpose of the structural subsystem is to find and correct these landmarks. The structural correction scheme uses three landmarks (called the *support set*) for normalization. The normalization procedure translates the mean of the support set to the origin, scales its average distance from the origin to a fixed value, and rotates the landmarks so that the first landmark in the support set (*the pivot*) falls on the y-axis. After this transformation is applied to all the landmarks, the distribution of each landmark can be modeled with a Gaussian (See [13] for more details).

In the training phase, we find the distribution parameters for all possible support sets. In the test phase, a support set is selected and the corresponding normalization is applied to all the landmarks. Selecting the best support set is possible by looking at the number of inliers (i.e. landmarks that turn up within their expected locations) and the joint likelihood under the support set. We can also trade-off speed for accuracy, and stop at the first support set with at least one inlier. For seven landmarks, there are 35 support sets of size three. Once a support set is selected, we can re-estimate the location of a landmark that falls outside its expected location by an inverse transformation applied to the expected location.

Denoting the distribution parameters of a landmark l_j with $\mathcal{N}(\mu_j, \Sigma_j)$, it is labeled as an outlier if the likelihood value produced under this distribution falls below a threshold:

$$\mathcal{L}(l_j, \mu_j, \Sigma_j) < \tau \tag{2}$$

3.1 BILBO

In a recent paper, Beumer et al. proposed an iterative structural correction scheme with a similar purpose [2]. The proposed algorithm (BILBO) first registers landmark locations to an average shape. During training, the registered landmark locations are perturbed with small rotations, translations and scalings. Then a singular value decomposition is used to compute a lower dimensional subspace. During testing, the landmark locations are projected to this subspace and back. Deviations from the average shape are corrected when passing through the bottleneck created by the subspace projection. A threshold value is monitored to detect the change due to backprojection. This threshold is increased at each iteration, and the algorithm stops once the change is smaller than the threshold.

We have constrasted our structural correction method (termed GOLLUM for Gaussian Outlier Localization with Likelihood Margins) with BILBO. We have used the parameter settings indicated in [2]. The experimental results of the comparison is given in the next section.

4 Experimental Results

4.1 Experiment 1

For the first experiment, we have used the first part of the Notre Dame University 2D+3D face database (FRGC ver.1) [11]. There were 943 images, of which half were used for training, one quarter for validation, and the rest for the test sets. Samples with poor 2D-3D correspondence were left out to treat all methods fairly. The results are reported separately for each different landmark type. The same structural subsystem corrections are applied to landmarks located with 2D, 3D and 2D+3D methods. Table 1 shows the localization accuracies for each landmark type when the acceptable distance to ground truth is less than or equal to three pixels on the downsampled image.

It is observed that 2D performs better in localizing outer eye corners and mouth corners. When coupled with the structural correction subsystem, the performance of 2D and 3D systems are close. Since the 2D information is richer, we expect it to produce a more accurate system when the training and test conditions are similar. Our simulations show that the proposed GOLLUM scheme outperforms it competitor BILBO. The albedo corrected images lose their discriminative power, and perform sub-optimally.

4.2 Experiment 2

We have used the FRGC ver.2, Fall 2003 dataset for a more challenging experiment. This dataset contains 1893 2D+3D images from the same set of subjects,

Table 1. Localization results for the first experiment

Method	*Outer Eye Corners*	*Inner Eye Corners*	*Nose*	*Mouth Corners*
2D	96.9 %	98.0 %	98.7 %	94.6 %
2D+GOLLUM	**99.3** %	**99.6** %	**100.0** %	**99.3** %
2D+BILBO	98.0 %	97.1 %	98.7 %	94.0 %
3D	87.9 %	98.4 %	96.7 %	85.4 %
3D+GOLLUM	95.7 %	99.3 %	98.2 %	88.1 %
3D+BILBO	89.7 %	98.2 %	96.9 %	88.8 %
ALBEDO	37.0 %	84.8 %	59.2 %	58.8 %
ALBEDO+GOLLUM	72.7 %	87.7 %	78.9 %	72.4 %
ALBEDO+BILBO	43.1 %	84.3 %	60.1 %	59.6 %

acquired six months later under expression variations and different lighting conditions, some of them so challenging that even the manual landmarking is difficult. Without suitable illumination compensation, the 2D statistical model is not expected to generalize correctly. However, 3D information is expected to be robust to illumination changes. We have directly applied the IMoFA-L models previously learned on ver.1 to this new dataset. Table 2 gives the localization results at an acceptance threshold equal to three pixels.

The system based on 2D features fails in the absence of adequate illumination compensation, whereas 3D depth features produce good results. The left and right ends of the horizontal crevice between the lower lip and the chin produce false positives for the mouth corners in 3D, and since this pattern conforms to the general face configuration it is very difficult to detect. This is the source of most of the mouth corner errors. The decrease in the mouth corner detection accuracy is partly due open-mouthed expressions in ver.2. The albedo correction increases the recognition accuracy for some landmarks, but there is no overall improvement.

Table 2. Localization results for the second experiment

Method	*Outer Eye Corners*	*Inner Eye Corners*	*Nose*	*Mouth Corners*
2D	18.4 %	9.9 %	0.2 %	31.8 %
2D+GOLLUM	18.4 %	10.8 %	1.8 %	31.7 %
2D+BILBO	17.0 %	15.5 %	1.4 %	29.9 %
3D	78.3 %	**97.2** %	96.7 %	20.1 %
3D+GOLLUM	**83.4** %	97.1 %	**98.0** %	29.3 %
3D+BILBO	79.3 %	96.3 %	96.8 %	**37.8** %
ALBEDO	3.5 %	12.9 %	1.5 %	21.5 %
ALBEDO+GOLLUM	3.9 %	12.8 %	2.3 %	21.2 %
ALBEDO+BILBO	4.1 %	15.1 %	2.6 %	20.6 %

5 Conclusions

The 3D system based on range images has performed close to the 2D system in Experiment 1, which contains illumination controlled 2D images. In the more challenging Experiment 2, 3D has performed remarkably good at nose tip and eye corners; but has failed at mouth corners, while the 2D system and 3D-assisted 2D system have very low detection rate. Our simulations show that the simple albedo correction scheme improves 2D on some points, but the illumination effects still deteriorate recognition. More elaborate albedo correction schemes use synthetic images to find suitable bases and iteratively estimate the illumination coefficients. This is left as a future work.

The local features of the faces provide reliable cues to identify facial landmarks independently. This is particularly useful when some of the landmarks are not available for detection. There may be acquisition noise that we frequently see in the laser-scanned eye regions, the subject may have a scar or deformity that renders some of the landmarks unrecognizable, there may be partial occlusions by facial hair. In this case, an optimization approach that attempts to locate all landmarks simultaneously may not converge to the correct solution. We propose an alternative approach that treats each landmark individually, and uses the structural relations between landmarks separately. Our structural correction scheme is shown to be superior to a recent competing technique.

Employing mixtures of factor analyzers allows us to strike a balance between temporal and spatial model complexity and accuracy. Although a full-covariance Gaussian mixture model has more representational power, it requires much more training samples than the model presently employed. Our model is able to represent the data with a smaller number of parameters.

Once the landmarks are located in the coarse scale, a fine-resolution search can be employed to refine these locations. The methods employed for the coarse scale are available in fine scale as well. However, larger windows need to be sampled in order to do justice to the local statistical information. In [13] a discriminatory approach that uses 2D DCT coefficients was successfully used for large scale refinement.

Acknowledgements

This work is supported by DPT project grant 03K 120 250 and TUBITAK project grant 104E080. The authors thank Dr. Miroslav Hamouz and Berk Gökberk for sharing ideas and code.

References

1. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2003) **25(2)** 218-233
2. Beumer, G.M., Tao, Q., Bazen, A.M., Veldhuis, R.N.J.: A Landmark Paper in Face Recognition. In: 7^{th} International Conference on on Automatic Face and Gesture Recognition. (2006) 73-78

3. Boehnen, C., Russ, T.: A Fast Multi-Modal Approach to Facial Feature Detection. In: 7^{th} IEEE Workshop on Applications of Computer Vision. (2005) 135-142
4. Colbry, D., Stockman, G., Jain, A.K.: Detection of Anchor Points for 3D Face Verification. In: IEEE Workshop on Advanced 3D Imaging for Safety and Security. (2005)
5. Conde, C., Serrano, A., Rodríguez-Aragón, L.J., Cabello, E.: 3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification. In: IEEE Conference on Computer Vision and Pattern Recognition. (2005)
6. Cristinacce, D., Cootes, T.F.: Facial Feature Detection and Tracking with Automatic Template Selection. In: 7^{th} International Conference on on Automatic Face and Gesture Recognition. (2006) 429-434
7. Çınar Akakın, H., Salah, A.A., Akarun, L., Sankur, B.: 2D/3D Facial Feature Extraction. In: SPIE Conference on Electronic Imaging. (2006)
8. İrfanoğlu, M.O., Gökberk, B., Akarun, L.: 3D Shape-Based Face Recognition Using Automatically Registered Facial Surfaces. In: International Conference on Pattern Recognition. (2004) **4** 183-186
9. Li, P., Corner, B.D., Paquette, S.: Automatic landmark extraction from three-dimensional head scan data. In: Proceedings of SPIE. (2002) **4661** 169-176
10. Liao, C.-T., Wu, Y.-K., Lai, S.-H.: Locating facial feature points using support vector machines. In: 9^{th} International Workshop on Cellular Neural Networks and Their Applications. (2005) 296-299
11. Phillips, P.J., P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W.J. Worek: Overview of the Face Recognition Grand Challenge. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. (2005) **1** 947-954
12. Salah, A.A., Alpaydın, E.: Incremental Mixtures of Factor Analyzers. In: International Conference on Pattern Recognition. (2004) **1** 276-279
13. Salah, A.A., Çınar Akakın, H., Akarun, L., Sankur, B.: Robust Facial Landmarking for Registration. Accepted by Annals of Telecommunications for publication.
14. Senaratne, R., Halgamuge, S.: Optimised Landmark Model Matching for Face Recognition. In: 7^{th} International Conference on on Automatic Face and Gesture Recognition. (2006) 120-125
15. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: Computer Vision and Pattern Recognition Conference. (2001) **1** 511-518
16. Wiskott, L., Fellous, J.-M, Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence. (1997) **19**(7) 775-779
17. Yan, Y., Challapali, K.: A system for the automatic extraction of 3-D facial feature points for face model calibration. In: IEEE International Conference on Image Processing. (2000) **2** 223-226
18. Zhang, L., Samaras, D.: Face Recognition Under Variable Lighting Using Harmonic Image Exemplars. In: Computer Vision and Pattern Recognition Conference. (2003) **1** 19-25