Locating Facial Features and Pose Estimation Using a 3D Shape Model

Angela Caunce, David Cristinacce, Chris Taylor, and Tim Cootes

Imaging Science and Biomedical Engineering, The University of Manchester, UK

Abstract. We present an automatic method for locating facial features and estimating head pose in 2D images and video using a 3D shape model and local view-based texture patches. After automatic initialization, the 3D pose and shape are refined iteratively to optimize the match between the appearance predicted by the model, and the image. The local texture patches are generated using the current 3D pose and shape, and the locations of model points are refined by neighbourhood search, using normalized cross-correlation to provide some robustness to illumination. A key aspect is the presentation of a large-scale quantitative evaluation, comparing the method to a well-established 2D approach. We show that the accuracy of feature location for the 3D system is comparable to that of the 2D system for near-frontal faces, but significantly better for sequences which involve large rotations, obtaining estimates of pose to within 10º at headings of up to 70º.

1 Introduction

There are many potential applications which require the location of facial features in unseen images - from in-car safety to crime-prevention. One of the major challenges stems from the fact that the pose of the head, relative to the camera, is often unknown. Although 2D statistical model-based approaches have proved quite successful, they do not deal well with large variations in pose, because the models lose specificity when significant pose variation is included in the training set [1]. Some authors [2, 3] have attempted to augment a 2D approach with a 3D shape model, and, in recent years, other authors have begun to experiment with fully 3D matching algorithms (see [4] for a review).

We present our 3D matching approach and with it attempt to progress two areas. The first is to provide a comprehensive quantitative evaluation of performance in both feature detection and pose estimation. The second is to show that the 3D approach performs as well as a well-developed 2D system on large datasets of near frontal images [5, 6], and surpasses it on large rotations.

1.1 Comparison to Other Methods

The first stage in any 3D modelling approach is building the model. In some work the 3D model is generated from multiple 2D search results [3]. Some authors use an artificial head model [7] or prior knowledge of face deformation [2]. Others find

G. Bebis et al. (Eds.): ISVC 2009, Part I, LNCS 5875, pp. 750–761, 2009.

correspondences between 3D head scans [4, 8] or generate artificial examples [9]. We use manual markups as the basis for our model thus overcoming the correspondence problem.

In early work, authors have used manual methods to show that a good initialisation leads to good pose estimation accuracy [10, 11]. Some authors therefore use automated means to not only locate the face beforehand, but also to make some estimate of the pose before searching begins [3, 9, 12, 13]. Also, integrating the search into a tracking strategy [3, 7, 13] enables systems to deal with the larger rotations without the need for complex initialisations on every image. In our experiments, we use a face detector [14] on images where an independent initialisation is required, and a tracking strategy on sequences.

To compensate for illumination variation, some approaches use illumination models [11, 12]. We use normalised view-based local texture patches similar to Gu and Kanade [9], but continuously updated to reflect the current model pose.

In summary, our approach uses a sparse 3D shape model [15] for pose invariance, and continuously updated view-based local patches for illumination compensation. On images with small rotations the system can locate the features well with a face detector and no specialized pose initialisation. For larger rotations the system works best when integrated into a tracking strategy and we successfully tackle images at headings of up to 70° to within 10° accuracy.

2 Shape Model

We built a 3D statistical shape model [15] from 923 head meshes. Each mesh was created from a manual markup of photographs of an individual. The front and profile shots of each person were marked in detail and the two point sets were combined to produce a 3D representation for that subject (Figure 1 top). A generic mesh, with known correspondence to the 3D points, was warped [16] to fit the markup giving a mesh for each individual (Figure 1 bottom). Since the same mesh was used in each case the vertices are automatically corresponded across the set.

Any subset of vertices from this mesh can be used to build a sparse 3D shape model. We used 238 points (Figure 1 right) which are close to features of interest such as eyes, nose, mouth, etc.

Each example is represented as a single vector in which the 3D co-ordinates have been concatenated:

$$(x_1,\dots x_n, y_1,\dots y_n, z_1,\dots z_n)^T \tag{1}$$

Principle Component Analysis is applied to the point sets to generate a statistical shape model representation of the data. A shape example $\mathbf{x_i}$ can be represented by the mean shape $\overline{\mathbf{x}}$ plus a linear combination of the principle modes of the data concatenated into a matrix $\mathbf{P}$:

$$\mathbf{x_i} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b_i} \tag{2}$$

where the coefficients $\mathbf{b_i}$ are the model parameters for shape $\mathbf{x_i}$. We established that the model performance improved when the number of modes was restricted. The results here are quoted for a model with 33 columns in $\mathbf{P}$ which accounts for approximately 93% of the variation in the training data. None of the subjects used in training was present in any of the images or videos used in the experiments.

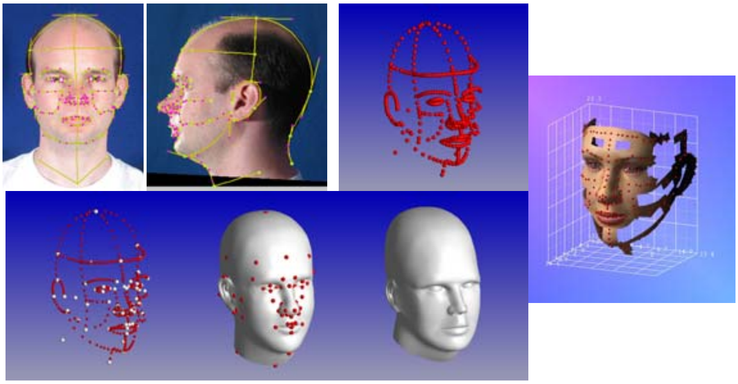

Fig. 1. The front and profile markups are combined to create a 3D point set (*top*). Using known correspondences between the markup and a generic head mesh, an individual mesh can be created for each subject (*bottom*). Only a subset of the mesh vertices are used to build the statistical shape model (*right*).

3 View-Based Local Texture Patches

The local patches are sampled from an average texture generated from 913 subjects. The individual examples are in the form of faces 'unfolded' from the meshes described in Section 2. Because all the vertices of the meshes have the same (UV) coordinates into the texture, all the unwrapped examples correspond directly pixel for pixel and it is easy to obtain the mean (Figure 2). Variation in the texture was not modelled for these experiments.

In order to successfully match the 3D shape model to the face in a 2D image, a texture patch is required at each point for comparison to the image. This patch is always the same size and shape throughout the matching process (5x5 pixels) but changes content at every iteration. It is updated based on the surface normal of the point and the current orientation of the model, and represents the view of the texture at that point (Figure 3). It is assumed for this purpose that the head is a globe and the texture lies tangential to the surface at each point with its major (UV) axes aligned to the lines of latitude and longitude. Black pixels are substituted outside the texture. To reduce speed, only a subset of 155 points are considered in the search. Many of the points excluded are from around the outside of the face where there is less information. Further, of the 155, only points which have surface normals currently facing forwards (less than 90 degrees to the view axis) are actually used to search at each iteration.

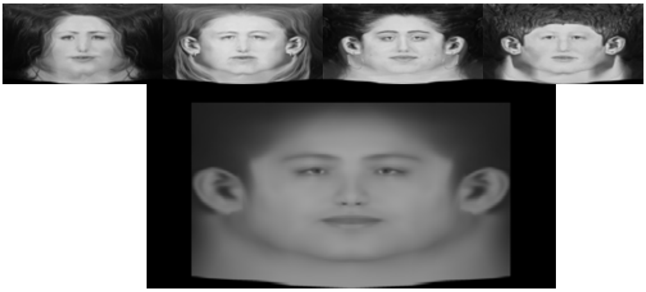

Fig. 2. The texture patches are sampled from a mean texture (*bottom*) averaged over a set of faces ‘unwrapped’ from the head meshes. Some example faces are shown (*top*).

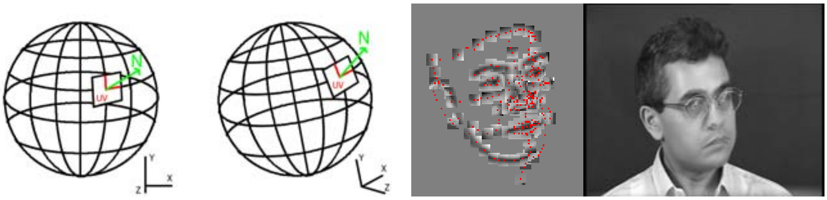

Fig. 3. The local texture patches are generated based on the current pose of the model. The UV axes of the texture are assumed to follow the lines of latitude and longitude of the head at 0º rotation (*left*). During matching the patches are a ‘window’ onto the texture oriented by the estimated pose of the head. Only forward facing points, determined from the surface normal are used to search.

4 Locating the Features

The model is initialised using the Viola-Jones (V-J) face detector [14]. The detector returns the location of a box, bounding the most likely location of a face in the image. The 3D shape model is placed within the box adopting its default (mean) shape and facing forwards (0° rotation).

The view-based patches are normalized and compared to the image using an exhaustive neighbourhood search. This is done for several iterations at each of a series of resolutions of both the model and the target image. Beginning with the lowest, the search is completed at each resolution before moving on to the next, and the shape and pose parameters are inherited at each resolution from the previous one. As the resolution increases (x2 at each step) the neighbourhood is increased by 2 pixels in each direction, which gradually concentrates the search.

The method begins at the lowest resolution and, at each point, a match value is calculated for all surrounding pixels in a 9x9 neighbourhood using normalised correlation. The best value gives the new target for each point. The targets are weighted in importance by the improvement in match value from the current position. Greater improvements are weighted more strongly. Once each point has a new 2D target location the z-component is estimated as the current z co-ordinate of the point. This assumes an orthogonal projection. Finally, the shape model is fitted in 3D to give a new estimate of the shape and pose parameters. This is a 2 stage process extended from the 2D case [15]. Firstly the points are rigidly aligned (rotation, scale, and translation) to minimise the sum of squared distances between matched points, then the shape model parameters (**b** in (2)) are updated using a least squares approximation.

4.1 Search Summary

- Initialise the model to the mean shape with 0° rotation using the V-J face detector.
- At the lowest resolution of the model find the best matching image resolution.
- At each model resolution and matching image resolution.
    - For a number of iterations.
        - For each forward point construct a patch based on the current model pose.
        - Search the neighbourhood around each point for the best match using normalised correlation to get a target point position.
        - Estimate pose and shape parameters to fit to the target points.

5 3D to 2D Comparison

In order to test the efficacy of the 3D in 2D search, it was compared to an implementation of a well-developed 2D shape matching approach: the Constrained Local Model (CLM) [17]. The two search methods were applied to images from two large publicly available datasets, neither of which contain large variations in pose:

- XM2VTS [6]: We used 2344 images of 295 individuals at 720x576 pixels.
- BioID [5]: We used 1520 images of 23 individuals at 384x286 pixels.

Both of these sets have manual markups but not the same features are located in each. Because of this, and the difference in model points from the 2D and the 3D model, only a small subset of 12 points was used for evaluation. The points chosen are located on the better defined features, common to all sets: the ends of the eyebrows (least well localised); the corners of the eyes (well localised); the corners of the mouth (well localised); and the top and bottom of the mouth (moderately well localised).

Both the 2D and 3D systems are initialised using the Viola-Jones face detector. We assessed the detector's performance by comparing the box returned by the algorithm to the 12 points of interest in the manual markup. If any points fell outside the box the detection was considered a failure. It was found that the detector failed on 8% of the BioID data set. These examples were excluded from the analysis, since both methods require initialisation in the location of the face.

6 Comparison Results

Figure 5 shows the cumulative distribution of the average point-to-point accuracy for the two methods and Table 1 provides a summary of these results. Due to the wide variation in size of the faces in the images, particularly in the BioID data, the errors are presented as a percentage of the inter-ocular (between pupils) distance and those in the table are the median of the average errors in each example. Also shown are the numbers of poorly located results. This is defined by a median average error of over 15%. The table distinguishes between the average results over all 12 points and the results just for the eyes, which are an easy feature to localise when marking manually. From the Figure and Table it can be seen that the 3D model results are generally better than those of the 2D CLM. Figure 8 shows some sample results for the 3D system.

7 Pose Handling and Estimation

The images in the data sets used in the experiments of section 5 are mainly near front facing. Although still challenging, it would not be expected that either system would dramatically outperform the other. However the advantages of a 3D model search are more apparent when dealing with larger rotations, both in terms of performance and pose estimation. To test this, a series of artificial images were generated with known poses.

Using the full mesh statistical model of the head described in Section 2, and a texture model built from the unwrapped textures described in Section 3, 20 synthetic subjects were generated. These were posed against a real in-car background to generate the artificial images. Figure 4 shows some examples. Feature marking was done automatically by extracting the 2D positions of selected mesh vertices.

The heads were posed as follows (Figure 4):

- Heading +/- (*r/l*) 90° in 10 degree intervals (right and left as viewed)
- Pitch +/- (*d/u*) 60° in 10 degree intervals
- Roll +/- (*r/l*) 90° in 10 degree intervals

For each rotation direction, the images were presented to the 2D and 3D systems as a sequence starting at zero each time. To perform tracking, the 3D system uses its latest result, if successful, to initialise the next search. Otherwise, the search is initialised as in section 4, using the V-J detector. The success of a search, for this purpose, is measured in 2 ways: by the final scaling of the model with respect to the image size; and by the average matching value over all the texture patches. The upper limits for both tests were fixed for all sequences.

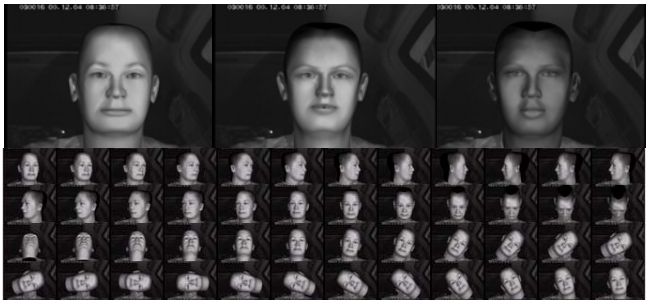

Fig. 4. Some of the synthetic subjects (*top*). And the 48 poses (in addition to zero).

8 Pose Results

Figure 5 shows the cumulative distribution of the point-to-point distances for the two methods on the artificial images. Table 1 summarises the results. The point-to-point errors are presented as pixels (inter-ocular distance is approximately 100 pixels). The automatic markup process described in the previous section failed in 3 cases therefore the results are reported on 977 images.

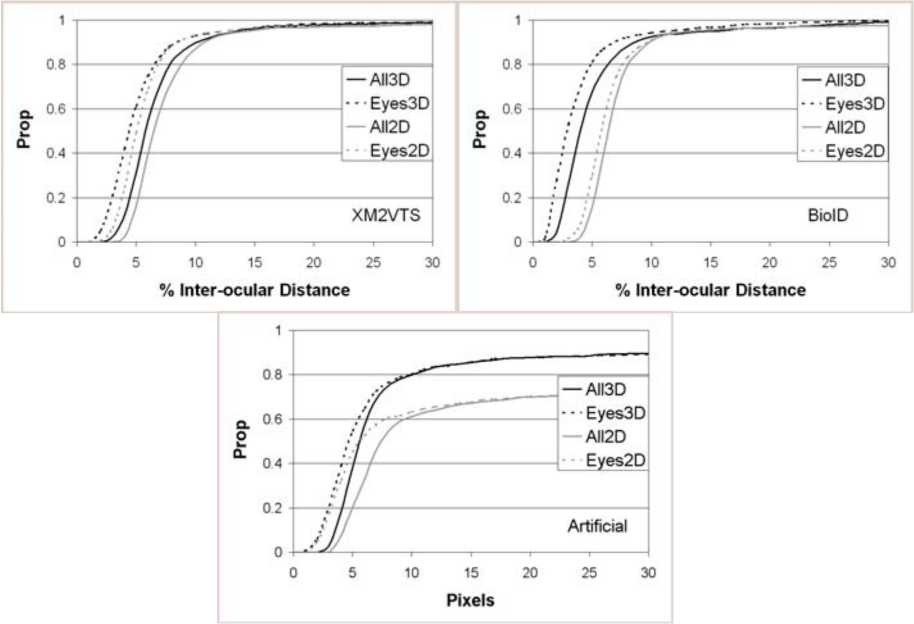

Fig. 5. The cumulative distribution of average point-to-point accuracy as % of inter-ocular distance for the real datasets (*top*), and as pixels for the artificial driver (*bottom*)

It can be seen that the 3D system out performs the 2D system and has a much lower failure rate. The graphs of Figure 6 indicate, as might be expected, that this is related to the larger rotations.

In addition to feature location the 3D model also provides an estimate of pose. Figure 7 shows the accuracy of the pose estimation at each rotation and Table 2 shows the ranges for which the median estimate lies within 10 degrees.

The pose estimation is returned as a quaternion which represents an angle of rotation about an axis and takes the form:

$$Q=(w,x,y,z); \qquad w=\cos\left(\frac{\theta}{2}\right); \qquad (x,y,z)=\sin\left(\frac{\theta}{2}\right)(x',y',z') \quad (3)$$

Where $\boldsymbol{\theta}$ is the angle of rotation and ***(x', y', z')*** is the axis. The graphs of Figure 7 show two error values for the angle and the axis (angle from actual rotation axis). It can be seen that at smaller rotations the axis error is larger than that at higher rotations. This is due to the ambiguity in head pose close to zero. It is compensated for by the shape model in feature detection.

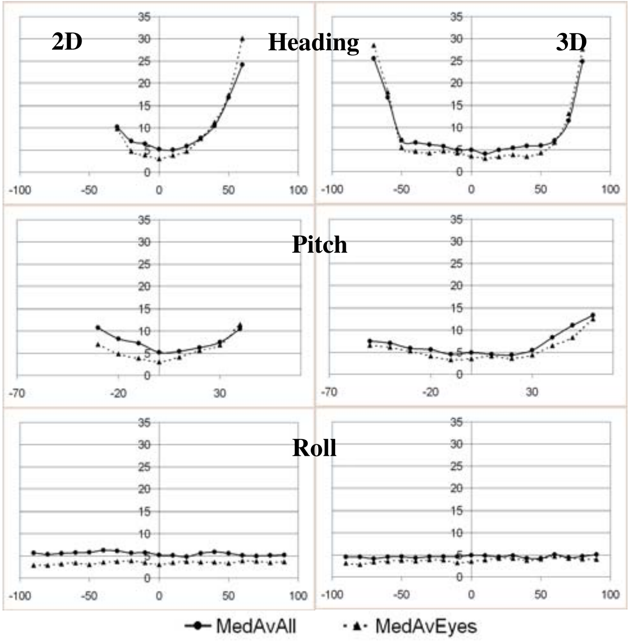

Fig. 6. The median average pixel errors at each rotation for the 2D system (*left*) and the 3D system (*right*). Each graph shows the angle across the horizontal and the pixel error on the vertical.

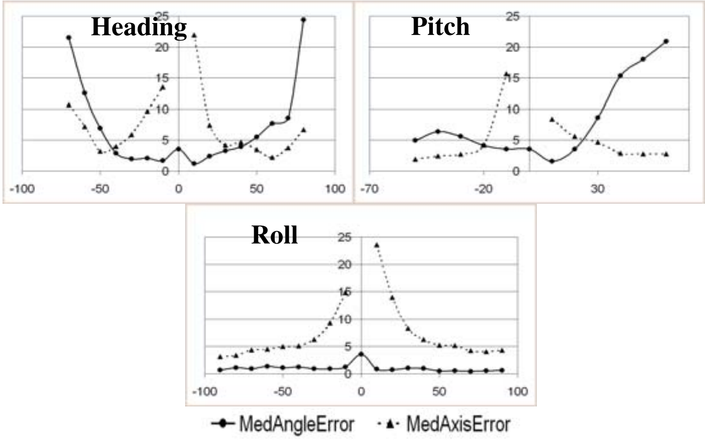

Fig. 7. The median angle errors for the pose estimation at each rotation. Each graph shows the rotation angle across the horizontal and the error, in degrees, on the vertical.

Table 2 indicates the ranges handled by each system where median average errors are within 15 pixels and median angle estimation is within 10º. Both systems handled all roll angles within these limits so these are not included. The 3D system has difficulty estimating pitch at positive rotations above 30º. This is probably due to the disappearance of the nostrils and mouth as the head rotates downwards. In contrast these features can be seen in larger upward rotations.

Figure 8 shows some sample search results from the artificial images for the 3D model.

9 Tracking Video Sequences

The 3D model was used to track features in 3 real-world in-car video sequences of 2000 frames each. The camera is located behind the steering wheel and below the head, therefore the model was initialized with a 40 degree upward pitch. Tracking was performed as described in section 7. Every 10th frame was manually marked but, due to occlusion from the steering wheel, not all 200 frames were suitable for inclusion. Table 1 shows the number of frames used for each assessment and the median average % errors and failures (>15% error). These sequences present difficult challenges because, as well as changing pose and expression, the illumination changes quite dramatically, and there are harsh shadows partially over the face at times. Figure 8 shows some search results. The errors are comparable to those on the datasets of section 5 although the failure rates are somewhat higher.

Table 1. Median average point to point errors for all data sets, presented as % inter-ocular distance or pixels, as indicated. The values are shown for all 12 pts and for just the eyes.

			All	Eyes	All	Eyes
Data Set	Images	Model	Med. Av. Error %		Fails (>15% Error) As % of set [No.]	
XM2VTS	2344	2D	6.44	5.03	4.56 [107]	3.58 [84]
		3D	5.80	4.44	3.75 [88]	3.41 [80]
BioID	1398	2D	6.33	5.75	4.51 [63]	3.86 [54]
		3D	3.98	2.83	5.01 [70]	3.22 [45]
Video 1	150	3D	8.11	6.37	16.00 [24]	14.67 [22]
Video 2	156		8.85	5.68	24.36 [38]	24.36 [38]
Video 3	136		5.57	4.45	14.71 [20]	13.97 [19]
			Med. Av. Error (Pixels)		Fails (>15 pixels Error) As % of set [No.]	
Artificial Driver	977	2D	7.47	5.54	32.86 [321]	32.24 [315]
		3D	5.51	4.61	14.53 [142]	14.33 [140]

Table 2. The ranges handled by each system to within the tolerance shown

	Point to Point Median Average Error <15 pixels		Pose Estimation Median Error <10 degrees	
	Heading	Pitch	Heading	Pitch
2D	-30 to 40	-30 to 40		
3D	-50 to 70	-50 to 60	-50 to 70	-50 to 30

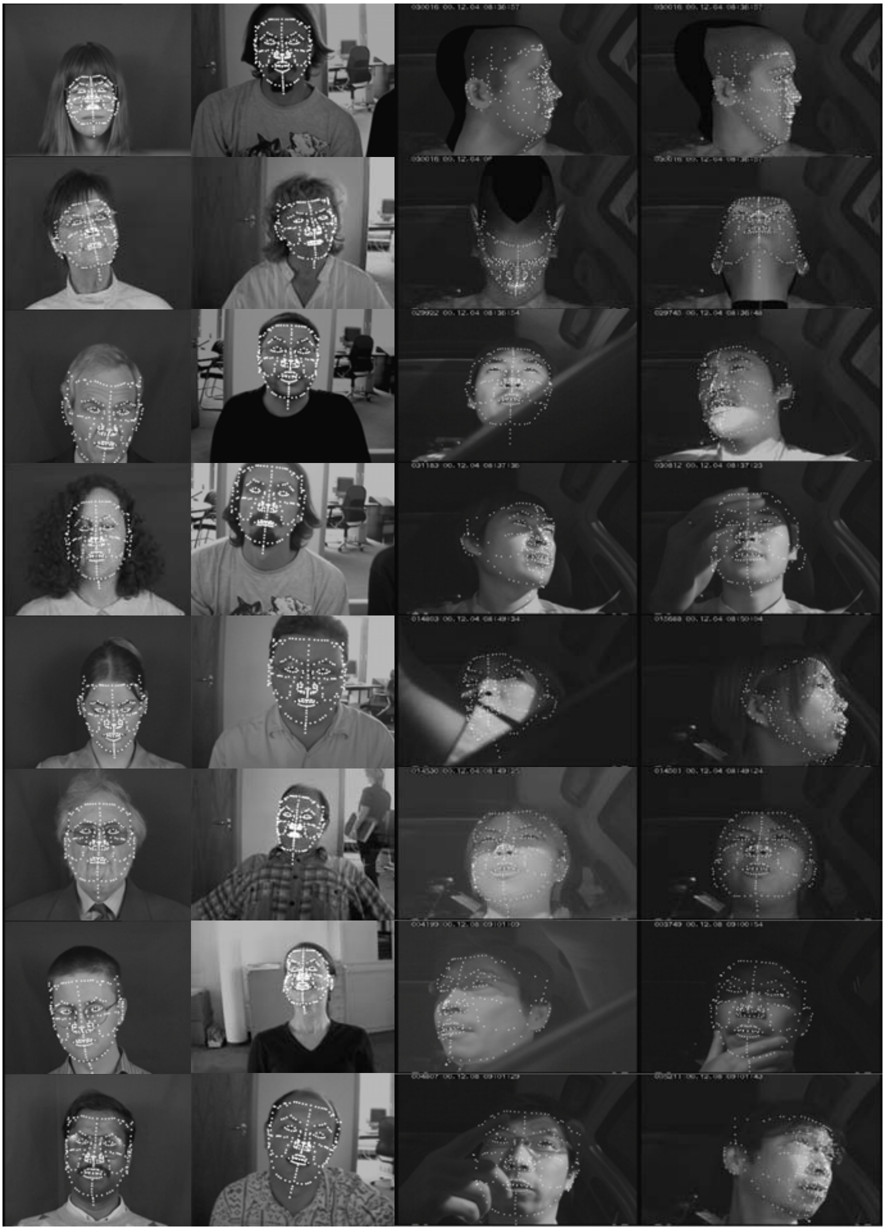

Fig. 8. Some sample 3D search results from: XM2VTS (*1st column*); BioID (*2nd column*); Artificial images (*4 top right*); Video sequences (*rest*). The two images at top left illustrate the failure modes of the search which generally result from the mouth confused with a moustache or the nose. In some cases the ears are not well shaped. This is because only forward facing points are used in the search and therefore the backs of the ears are generally not used. The model deals well with occlusion, glasses, variable illumination, low contrast, and, in many cases, facial hair.

10 Discussion and Future Work

On large datasets of near frontal images the 3D model has been shown to be comparable to a well developed 2D shape matching method. In addition, it has proved superior when handling large rotations and can provide an estimate of pose, critical for gaze dependant applications such as in-car safety.

On the XM2VTS dataset we achieved a median feature detection error of less than 6% inter-ocular distance with only 3.75% of examples falling outside a distance limit of 15%. On artificial images, with known poses, the 3D search exhibited similarly low errors at up to 50° headings and handled rotations of up to 70° with <15 pixels median average error. The system was able to estimate the pose in these images to within a median of 10° for rotations up to 70° right (as viewed) and 50° up.

Currently, the 3D system is initialised using a detector tuned to frontal faces and is instantiated in a frontal pose. One of the key ways that this system may be improved is by developing a more versatile initialisation for unseen sequences, which may not conform to these assumptions.

Acknowledgements

This project is funded by Toyota Motor Europe who provided the driver videos. We would like to thank Genemation Ltd. for the 3D data markups and head textures.

References

1. Matthews, I., Xiao, J., Baker, S.: 2D vs. 3D Deformable Models: Representational Power, Construction, and Real-Time Fitting. International Journal of Computer Vision 75, 93–113 (2007)
2. Vogler, C., Li, Z., Kanaujia, A., Goldenstein, S., Metaxas, D.: The Best of Both Worlds: Combining 3D Deformable Models with Active Shape Models. In: International Conference on Compute Vision, pp. 1–7 (2007)
3. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-Time Combined 2D+3D Active Appearance Models. In: Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 535–542 (2004)
4. Romdhani, S., Ho, J., Vetter, T., Kriegman, D.J.: Face Recognition Using 3-D Models: Pose and Illumination. Proceedings of the IEEE 94, 1977–1999 (2006)
5. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust Face Detection Using the Hausdorff Distance. In: International Conference on Audio- and Video-based Biometric Authentication, Halmstaad, Sweden, pp. 90–95 (2001)
6. Messer, K., Matas, J., Kittler, J., Jonsson, K.: XM2VTSDB: The Extended M2VTS Database. In: International Conference on Audio- and Video-based Biometric Person Authentication, Washington DC, USA (1999)
7. Dornaika, F., Ahlberg, J.: Fitting 3D face models for tracking and active appearance model training. Image and Vision Computing 24, 1010–1024 (2006)
8. Blanz, V., Vetter, T.: A Morphable Model for the Synthesis of 3D Faces. In: SIGGRAPHH, pp. 187–194 (1999)
9. Gu, L., Kanade, T.: 3D Alignment of Face in a Single Image. In: International Conference on Computer Vision and Pattern Recognition, New York, vol. 1, pp. 1305–1312 (2006)

10. Blanz, V., Vetter, T.: Face Recognition Based on Fitting a 3D Morphable Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1063–1074 (2003)
11. Ishiyama, R., Sakamoto, S.: Fast and Accurate Facial Pose Estimation by Aligning a 3D Appearance Model. In: International Conference on Pattern Recognition, vol. 4, pp. 388–391 (2004)
12. Romdhani, S., Vetter, T.: 3D Probabilistic Feature Point Model for Object Detection and Recognition. In: Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, pp. 1–8 (2007)
13. Zhang, W., Wang, Q., Tang, X.: Real Time Feature Based 3-D Deformable Face Tracking. In: European Conference on Computer Vision, vol. 2, pp. 720–732 (2008)
14. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57, 137–154 (2004)
15. Cootes, T.F., Taylor, C.J.: Active Shape Models - 'Smart Snakes'. In: British Machine Vision Conference, pp. 266–275 (1992)
16. Bookstein, F.L.: Principal Warps: Thin-Plate Splnes and the Decomposition of Deformations. IEE Transactions on Pattern Analysis and Machine Intelligence 11, 567–585 (1989)
17. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition 41, 3054–3067 (2007)