Using Detailed Independent 3D Sub-models to Improve Facial Feature Localisation and Pose Estimation

Angela Caunce, Chris Taylor, and Tim Cootes

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

Abstract. We show that the results from searching 2D images or a video sequence with a 3D head model can be improved by using detailed sub-models. These parts are initialised with the full model result and are allowed to search independently of that model, and each other, using the same algorithm. The final results for the sub-models can be reported exactly, or optionally fed back into the full model to be constrained by its parameter space. In the case of a video sequence this can then be used in the initialisation of the next frame. We tested various data sets, constrained and unconstrained, including a variety of lighting conditions, poses, and expressions. Our investigation showed that using the sub-models improved on the original full model result on all but one of the data sets.

1 Introduction

Head and facial feature tracking can provide important information in various environments with respect to the activity and attitude of the subject. For example, in a driving scenario, head orientation alone can indicate attentiveness. Feature localisation and subsequent behaviour analysis can give a detailed picture of the driver's state and possible intent. This could identify critical or dangerous situations. Recently, 2D models have been used with great success to localise and analyse features of the face [1-3], however in some unconstrained scenarios with large pose variation this approach may be limited. Multiple 2D models and detectors may be required for different views [4]. There is also the additional problem of view-based occlusion. As a consequence, authors have been experimenting with augmented 2D [5, 6] and 3D [7]. Due to pose invariance, a 3D model requires less training data than its 2D counterpart, and is able to report critical pose information directly without additional calculation. In [8] the authors showed that their 3D method outperformed an established 2D approach on out of plane rotations and in [9] they extended this by integrating some limited facial actions for preliminary behaviour analysis. To do this they used two sparse, largely symmetrical, statistical point models of the whole face. However, the complex interplay of the various parts of the face may not have been fully realised due to global constraints. Also, the ability to deal with some individual quirks, like a crooked smile, is limited by the training set. This is a standard problem with deformable objects containing complex sub-parts. One solution is to define the sub-parts separately and model their inter-relationship. The search is performed by locating the

G. Bebis et al. (Eds.): ISVC 2012, Part II, LNCS 7432, pp. 398–408, 2012.

sub-parts and confirming an acceptable configuration. Methods include: pictorial structures [10]; star models [11]; Hierarchical Deformable Templates [12]; and probabilistic approaches [13, 14]. Martinez [15] used a probabilistic approach and weighted abstract sub-parts of the face based on their involvement in the test expression. However working in 3D has advantages in that many configurations of parts (those derived from the object's pose) are already constrained by the model structure and only the articulation problem remains. Generally, authors working in 3D have taken a bottom up approach, that is finding sub-parts and combining them in some meaningful way. In [16] Blanz and Vetter show that this produces visually pleasing results. Tena et al [17] illustrated that sub-parts improved performance when reconstructing motion capture data.

1.1 Contribution

Our method takes a top down approach, in that we localise the face with a full model using the method in [9], and then refine the result by allowing the parts to search independently afterwards. By extending the method in this way, we have increased its accuracy by improving its versatility. The system can deal with new feature configurations, without requiring additional training data. This may even reduce the amount of data required for training in future. We demonstrate the improvements by reporting results on 7 large datasets with various challenges. These show improvements either overall, or in some localised region of the face. The method has all the advantages of 3D outlined above, and can deal with greater pose variation than demonstrated in other works. We report median average point-to-point errors of less than 15 pixels (inter-ocular distance ~100 pixels) on headings up to 70 degrees and on all pitch angles tested (up to 60 degrees).

2 Search Method

The search method uses two sparse 3D statistical models of the form:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} \tag{1}$$

Where each example $\mathbf{x}$ is represented by a vector of n 3D co-ordinates $(x_1, y_1, z_1, \ldots\ldots x_n, y_n, z_n)$. Each is expressed in (1) as the mean vector, $\bar{\mathbf{x}}$, plus a linear combination of the principal components, $\mathbf{P}$, with coefficients, $\mathbf{b}$.

One of the two models is built from 'identity' training data, i.e. data from 923 individuals, with a close to neutral expression, eyes open, and mouth closed. The other is built from a small set of facial actions created from a neutral base.

Unlike other approaches which use a combined model strategy [18, 19], these two models are used in an alternating process to localise the features of the face and to provide a basic representation of some simple behaviours. This is done by substituting the results from the ID model into the actions model, and vice-versa, when

matching to the target points. Therefore, at each iteration of the algorithm, both models are fitted in sequence to the same target before moving on to the next iteration. This is represented in the following equations:

$$\mathbf{x}^{k(1)} = \bar{\mathbf{x}}^{k(1)} + \mathbf{P}_{\mathbf{ID}}\mathbf{b}_{\mathbf{ID}}^{k} \tag{2}$$

$$\bar{\mathbf{x}}^{k(1)} = \bar{\mathbf{x}}_{\mathbf{ID}} + \mathbf{P}_{\mathbf{A}}\mathbf{b}_{\mathbf{A}}^{k-1} \tag{3}$$

$$\mathbf{x}^{k(2)} = \bar{\mathbf{x}}^{k(2)} + \mathbf{P}_{\mathbf{A}}\mathbf{b}_{\mathbf{A}}^{k} \tag{4}$$

$$\bar{\mathbf{x}}^{k(2)} = \bar{\mathbf{x}}_{\mathbf{ID}} + \mathbf{P}_{\mathbf{ID}}\mathbf{b}_{\mathbf{ID}}^{k} \tag{5}$$

Where *k(1)* and *k(2)* refer to the 1st and 2nd fit at each iteration *k*, (5) is the current identity result and is used as the action model mean, (3) is devised from the current action result and is used as the identity model mean, and $\mathbf{b}_{\mathbf{A}}^{\mathbf{0}}$ is the zero vector. Notice that the action model mean, $\bar{\mathbf{x}}_{\mathbf{A}}$, is not used since this has no meaning in this context. See [9] for further details.

Most of the target points are located using an independent local template matching at each model point. There are 238 points in the model and each can search with a small (5x5) view based texture patch. This patch is extracted from a mean texture generated from 913 subjects. The population variation in texture is not modelled. The content of the patch is updated at every iteration to reflect the current pose, and is compared using normalised correlation to a neighbourhood around the point to find the best match. This has the advantage of providing some robustness to illumination variation over the face and between images.

Those points at approximately 90 degrees to the viewpoint do not use this technique but search along the surface normal for the strongest edge. Once the target points are established the whole model is fitted using the active shape model fitting method in [20] extended to 3D, assuming an orthogonal projection. This is a two-stage process. Firstly, the points are rigidly aligned to minimise the sum of squared distances between matched points, then the shape model parameters (**b** in (1)) are updated using a least squares approximation. The global fit has the advantage of minimising the effect of badly matched or occluded points.

The search is conducted at multiple resolutions, starting at the lowest, and is completed at each resolution before moving onto the next.

Initialisation is achieved using the Viola-Jones (V-J) face detector [21]. For the video sequences this occurs once at the start and again only when the search fails during the sequence. Failure is determined by comparing the average match value of the template patches to some threshold. If the match does not fail, the search on the next frame is initialised using the latest result. For still images the V-J detector is used on each example.

3 Sub-models

Fitting the model to the target using global constraints, as outlined above, has the advantage of keeping all of the areas of the model in their expected place, as well as having a neutralising effect on rogue matches, such as those found at occluded points. However, once this process is completed, it may be expected that the final result will also suffer, since un-correlated movements of the individual sub-parts may be lost as noise. Plus, individual quirks, such as a crooked smile, cannot be localised unless they are specifically included in the training set.

To provide the added flexibility necessary we propose extending the method by allowing sub-models to continue searching independently, after the full model search has completed.

The face models described in section 2 have a sparse representation. They are built from 238 vertices of a full 3D head mesh, warped [22] to fit each individual in the training set. However, for the sub-models we chose to use all the vertices but only from parts of the mesh around the eyes and mouth. We thus built two sub-models from the areas shown in Figure 1. As with the full face models, an identity and an actions model were constructed for each sub-part.

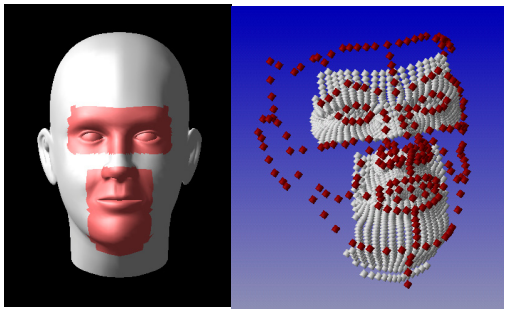

Fig. 1. The two sub-models are built from all the vertices in the eye and mouth areas *(left)*. The relationship between the sparse full face model and the sub-models is shown right.

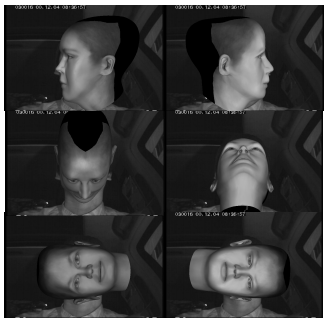

Fig. 2. The Artificial Driver data set contains extreme poses

4 Sub-model Search

The full model search is conducted over a series of three increasing image/texture resolutions. The search is completed at each resolution before continuing to the next. For the sub-models we use exactly the same search method except that searching always begins at the penultimate resolution. Experiments indicated that results were improved if two resolutions were used but there appeared to be no advantage to using all three.

When the full model search is complete the sub-models are initialised in their respective positions. Since the sub-models have many more points than the full model an initialisation target is constructed for each sub-part from just those points common to both. To do this the common points have a weighting of 1 whereas all other points have weight 0. The fitting method is the same method used in the image search, i.e. the 3D extension of that in [20], whereby an equation of the form in (6) is minimised to find the new model point positions.

$$\min_{\mathbf{b},\mathbf{t}} \left\{ \left| \mathbf{W}\left(\mathbf{T}_{\mathbf{t}}\left(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b}\right) - \mathbf{x}_{\mathbf{obs}}\right)\right|^2 \right\} \tag{6}$$

T represents the camera transform (in this case orthogonal) and pose with parameters **t**, **W** is the diagonal weighting matrix, and $\mathbf{x}_{\mathbf{obs}}$ is the observed, or target, set of points.

The identity and action sub-models are fitted to this target alternately as described above using equations (2)-(5). After this initialisation, each sub-part is allowed to search the image independently of the other and of the full model.

4.1 Feedback and Reporting Strategies

Once the results from the sub-models are obtained there are several ways that they can be integrated with the full model results for analysis. The most obvious is to use the parts to replace all the points in the full model that are common to the sub-models. We refer to this as 'Exact Parts'. The alternative is to feed back (FB) the common points into the full model by fitting it to the Exact Parts result. This fitting follows the same iterative alternating process described above for the sub-model initialisations. Normally, when analysing a video stream, the result of each frame is used to initialise the search on the next. This means that if the sub-model results are fed back into the full model these will influence subsequent frames. For a series of still images this will have no effect.

Therefore for still images we report these methods: 'NoParts' (results from the full model search); 'NFBExact' (exact parts substituted); and 'FB' (feedback – i.e. full model fitted to parts results). For video data, method 'FB' will affect future results and there is an additional combination: 'FBExact' (feedback affects future results but exact parts reported).

5 Data Sets

For our analysis we used 7 different data sets presenting various challenges:

- XM2VTS [23]: 2344 still images of 295 individuals at 720x576 pixels. The subjects are posed against a fairly uniform backdrop and in general have neutral expressions and near frontal poses.
- BioID [24]: 1520 still images of 23 individuals at 384x286 pixels. The data was acquired in an office setting so has cluttered backgrounds. It shows much more natural poses (although still mainly near frontal) and a variety of expressions.

- Expressions: 401 still images of 103 individuals at 1024x768 pixels. Each person is making some or all expressions selected from: eyes closed; smile; neutral; frown; and surprise. These images are near-frontal and posed against a uniform background so the main challenge is from the facial actions.
- 3 video sequences of 3 different drivers. These were taken in real driving conditions. Each sequence is 2000 frames but only a subset of each was used for evaluation: 150, 156, and 136 frames. There is a great deal of lighting variation, within and between frames, and there is a wider variation in poses than in the still images.
- Artificial driver short sequences [8]. This publicly available dataset was devised to assess the ability of search methods to deal with large poses. It comprises of a series of images of 20 synthetic subjects in known poses. The subjects were arranged against a real in-car background (Figure 2). Each sequence starts at zero rotations and runs in a single direction. The sequences are as follows: Heading +/- 90º; Pitch +/- 60º; and Roll +/- 90º.

6 Results

The available annotations for the different data sets do not mark the same features. Due to this, and the differences between the annotations and the model points, only a subset of 12 points was used for point to point error evaluation. The points chosen are located on the better defined features, common to all sets: the ends of the eyebrows; the corners of the eyes; the corners of the mouth; and the top and bottom of the mouth.

On the still image databases the search is initialised using the V-J detector on every image. Since the method relies on initialisation in the area of the face, we assessed the detector's performance by comparing the 12 points of interest to the box it returned. If any manually marked 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 therefore excluded from the analysis.

The generic model can handle a variety of challenges including: occlusion; pose variation; and variable lighting [9]. In many cases the sub-models will make only a small correction. However, Figure 3 illustrates the ability of the sub-parts to correct poorer results from the full model. If this is very inaccurate, sometimes only one part can recover (top row, column 2) and sometimes recovery is not possible (top row, column 3). The middle left image of Figure 3 illustrates the ability of the method to adapt to non-standard configurations, in this case the mouth is not symmetric with the face. Normally this would need to be included in a training set to be handled by a full face model of this kind.

Figure 4 shows the cumulative error proportion plotted against average point-to-point (P2P) distance for each dataset. In most cases this is a percentage of the inter-ocular distance (IOD). This normalises for the fact that the head size can vary across the data set, which is particularly true of the BioID images. For the artificial data the errors are presented as pixels because of the large pose variation. However, the head size is fairly constant on all images and IOD on the frontal faces is ~100 pixels.

Fig. 3. Examples showing how the sub-models can correct a poor full model result. Top row shows real-world driver video. The second image illustrates the independence of the models: only one has recovered. The third image shows that if the result is very poor the parts cannot always recover. The middle row shows BioID images, including glasses. The first image illustrates how the sub-models can allow adaptation to non-symmetric features (crooked mouth). The bottom row shows Expressions Surprise and Expressions Frown. Black points are the full model result and white points show the sub-models (common points).

Generally, including the parts improves on the original 'NoParts' result. However, examining the cumulative error curve for Video 1 up to the 15% threshold, we observed that performance got worse when the parts were introduced. Extending the curve to show errors beyond 15% revealed that, although the errors were increased the proportion of failures was reduced albeit at larger thresholds (shown in Figure 4).

Examining the results for different parts of the face revealed that the mouth area was much improved even at the lower thresholds (also shown in the figure). This indicates that the mouth model was able to correct many of the failures thus pushing the curve higher. For the other data sets the eyes showed the best performance. This is unsurprising since the corners of the eyes are the most easily located when marking

ground truth on the data. It is therefore likely to produce the lowest errors in a model search. The variant producing the best results for the eye corners is therefore shown on the graphs of Figure 4. For the artificial data the results of using a 2D CLM are included for comparison.

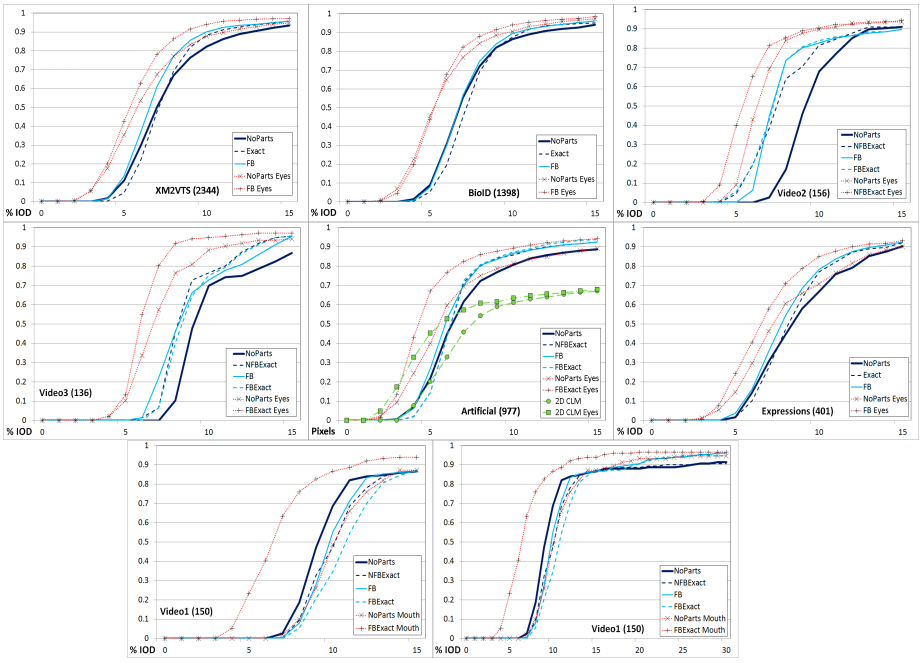

Fig. 4. Cumulative Error Distributions. The proportion of data is plotted against the point-to-point error as % inter-ocular distance (IOD), except for the artificial data which is shown in pixels. The number of images in the data set is given in brackets. The best local result is also shown, which is the eyes for all but Video 1. 2D CLM results are also given for the artificial data and an extended error plot is shown for Video 1.

Figure 5 shows the P2P errors broken down by rotation angle for the artificial images. Also shown are the errors on the estimated pose. Here the pose is reported as a quaternion, which represents a rotation about an axis vector. We examined the error on both the angle and the axis. The latter is calculated as the rotation angle between the known and estimated axes. Since the pose is not affected when using 'Exact' methods only the original and FB methods are shown. There are no axis results for 0 degrees rotation because of the high degree of uncertainty on the rotation axis at that point. As with figure 4, including the sub-parts has had a generally beneficial effect. For heading, the improved P2P errors are below 8 pixels for angles between +/- 50 degrees, and the estimated angle error is 5 degrees or less for the same range. For pitch the P2P error is below 13 pixels as far as +60 degrees and below 8 for -60 degrees. The angle estimation error is no more than 9 degrees for the entire range. Since

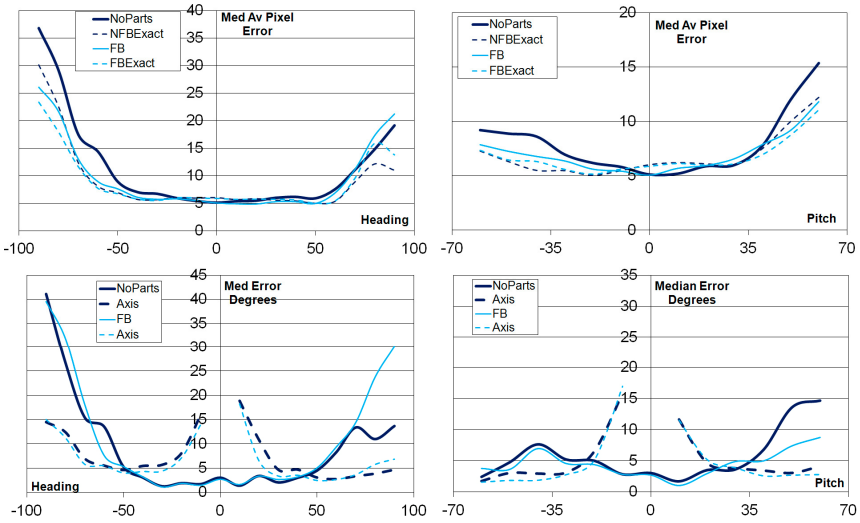

Fig. 5. The top row shows the point-to-point errors for the artificial data, broken down by out of plane rotation angle. The bottom row shows pose estimation errors over the same ranges.

the results for roll (i.e. in plane rotations) were consistent over all angles and methods they are not shown. The median average P2P error varied between 4.75 and 7.7 pixels. The angle estimation error was consistently less than 2.9 degrees.

6.1 Discussion

The easiest data sets in this group are XM2VTS and BioID and, as might be expected, the overall performance (all methods) is relatively better on these sets, closely followed by the expressions and the artificial data. The real-world videos obviously present the greatest challenge and the relative performance reflects this.

If we allow that Video 1 was improved in one area only, the mouth, then there is an overall improvement on every other data set from adding in the subparts. The other two videos and the artificial data show the most obvious improvements which implies that using sub-models has the most beneficial effect when used in the more difficult situations, as might be hoped, since it is a corrective technique. For the still images it seems that the best improvement is when the full model is re-fitted to the parts result (FB). This implies that the ability of the parts to correct the full model and deal with non-standard examples has aided the method but a final regularisation step is still needed. However, in the case of the sequence data this is less clear cut. For Videos 2&3 and the artificial data, 'NoParts' is the worst performer but whether exact or corrected parts should be reported needs further investigation.

When compared to state-of the-art 2D systems such as [1, 3], our method does not perform as well on the common BioID data set. However, in [3] the authors use near frontal faces only and in [1] state that their experimental data set, acquired from the

internet, does not have profile or near profile images and that all faces were detected by an off the shelf face detector. This will tend to exclude not only extreme poses but also unusual lighting and some occlusions. We have shown that our method deals with large rotations, a wide variety of data types and conditions and, in addition, can provide an estimate of pose, for those gaze-critical applications.

7 Summary

We have presented an extension to 3D model search which allows refinement of the results using independent 3D sub-parts. From the graphs presented in Figures 4 & 5 it can be seen that, for all data sets but Video 1, including the sub-parts has had a clearly positive effect. Even in the case of Video 1, when breaking down the errors between parts of the face, it can be seen that the mouth area is vastly improved. However, there is uncertainty as to whether reporting the exact parts points or refitting the model gives the best individual result. This requires further study and implies that an iterative approach, alternating between the full and parts models, may yield even more improvements.

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. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing Parts of Faces Using a Consensus of Exemplars. Computer Vision and Pattern Recognition, 545–552 (2011)
2. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition 41, 3054–3067 (2007)
3. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial Point Detection Using Boosted Regression and Graph Models. Computer Vision and Pattern Recognition, 2729–2736 (2010)
4. Pentland, A., Moghaddam, B., Starner, T.: View-Based and Modular Eigenspaces for Face Recognition. Computer Vision and Pattern Recognition, 1–7 (1994)
5. 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 Computer Vision, PP. 1–7 (2007)
6. 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)
7. 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)
8. Caunce, A., Cristinacce, D., Taylor, C., Cootes, T.: Locating Facial Features and Pose Estimation Using a 3D Shape Model. In: International Symposium on Visual Computing, Las Vegas, pp. 750–761 (2009)

9. Caunce, A., Taylor, C., Cootes, T.: Adding Facial Actions into 3D Model Search to Analyse Behaviour in an Unconstrained Environment. In: International Symposium on Visual Computing, Las Vegas, pp. 132–142 (2010)
10. Fischler, M.A., Elschlager, R.A.: The Representation and Matching of Pictorial Structures. IEEE Transactions on Computers C-22, 67–92 (1973)
11. Felzenswalb, P.F., Girshick, R.B., McAllester, D.: Cascade Object Detection with Deformable Part Models. Computer Vision and Pattern Recognition, 1–8 (2010)
12. Zhu, L., Chen, Y., Yuille, A.: Learning a Hierarchical Deformable Template for Rapid deformable Object Parsing. IEEE Trans. Pattern Analysis and Machine Intelligence 32, 1029–1043 (2010)
13. Burl, M.C., Weber, M., Perona, P.: A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 628–641. Springer, Heidelberg (1998)
14. Hua, G., Wu, Y.: Sequential Mean Field Variational Analysis of Structures De-formable Shapes. Computer Vision and Image Understanding 101, 87–99 (2006)
15. Martinez, A.M.: Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class. Pattern Analysis and Machine Intelligence 24, 748–763 (2002)
16. Blanz, V., Vetter, T.: A Morphable Model for the Synthesis of 3D Faces. SIGGRAPH, pp. 187–194 (1999)
17. Tena, J.R., Torre, F.D.l., Matthews, I.: Interactive Region-Based Linear 3D Face Models. SIGGRAPH (2011)
18. Amberg, B., Knothe, R., Vetter, T.: Expression Invariant 3D Face Recognition with a Morphable Model. In: International Conference on Automatic Face Gesture Recognition, Amsterdam, pp. 1–6 (2008)
19. Basso, C., Vetter, T.: Registration of Expressions Data Using a 3D Morphable Model. Journal of Multimedia 1, 37–45 (2006)
20. Cootes, T.F., Cooper, D.H., Taylor, C.J., Graham, J.: Active Shape Models - Their Training and Application. Computer Vision and Image Understanding 61, 38–59 (1995)
21. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57, 137–154 (2004)
22. Bookstein, F.L.: Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 567–585 (1989)
23. 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)
24. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust Face Detection Using the Hausdorff Distance. In: International Conference on Audio and Video-based Person Authentication, Halmstaad, Sweden, pp. 90–95 (2001)