Resolution-Aware 3D Morphable Model

Guosheng Hu
g.hu@surrey.ac.uk
Chi Ho Chan
chiho.chan@surrey.ac.uk
Josef Kittler
j.kittler@surrey.ac.uk
William Christmas
w.christmas@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

Abstract

The 3D Morphable Model (3DMM) is currently receiving considerable attention for human face analysis. Most existing work focuses on fitting a 3DMM to high resolution images. However, in many applications, fitting a 3DMM to low-resolution images is also important. In this paper, we propose a Resolution-Aware 3DMM (RA-3DMM), which consists of 3 different resolution 3DMMs: High-Resolution 3DMM (HR- 3DMM), Medium-Resolution 3DMM (MR-3DMM) and Low-Resolution 3DMM (LR-3DMM). RA-3DMM can automatically select the best model to fit the input images of different resolutions. The multi-resolution model was evaluated in experiments conducted on PIE and XM2VTS databases. The experimental results verified that HR-3DMM achieves the best performance for input image of high resolution, and MR-3DMM and LR-3DMM worked best for medium and low resolution input images, respectively. A model selection strategy incorporated in the RA-3DMM is proposed based on these results. The RA-3DMM model has been applied to pose correction of face images ranging from high to low resolution. The face verification results obtained with the pose-corrected images show considerable performance improvement over the result without pose correction in all resolutions.

1 Introduction

Human face analysis has become a popular research topic in computer vision. One of the tasks required by many applications is to perform face reconstruction from the input image. Many models have been proposed for this purpose. These models can be classified into two groups: 2D face models and 3D face models.

The 2D-based group includes Active Appearance Models (AAMs) [4]. AAMs model the shape and texture variations statistically. Principal Component Analysis (PCA) is applied to compress the texture and shape information during the training process. The reconstruction process involves fitting the statistical model to an input image so as to minimise the difference between the input image and the fitted image. The authors in [7] investigated the relationship between the resolution of the model and the resolution of the input image, and concluded that the best fitting performance is obtained when the model resolution is slightly

http://dx.doi.org/10.5244/C.26.109

higher than the input image resolution. [5] proposed a Resolution Aware Fitting (RAF) algorithm to fit a low-resolution image with a high-resolution model. RAF changes a high-resolution AAM to a low resolution AAM by averaging neighbouring pixel intensities of high resolution AAM. Although AAMs exhibit promising face analysis performance, there is still a problem that the reconstruction with AAMs fails if the in-depth rotation of the face becomes large. Almost all the 2D-based models suffer from the same problem.

To overcome the limitations of 2D-based face models, 3D-based methods have been proposed. The main advantage of 3D-based models is that 3D shape does not change under different viewpoints [6]. 3D morphable model (3DMM) [1] is also a statistical model for face reconstruction, and is represented by its shape and texture: the shape is captured as vertices in three dimensions and the texture of the face is conveyed by the colour information of the polygonal patches created by the triplets of neighbouring vertices. Most of the current research just focuses on fitting high resolution 3DMM to high resolution images. However, in many applications images to be processed are of low resolution. It is therefore very important to investigate how the fitting performance is affected by the resolution of 3DMM and of the input image. The aim of this study is to identify the best fitting performance for a spectrum of input image resolutions.

In this paper, the relationship between 3DMM resoluton and input image resolution is studied and a Resolution-Aware 3DMM (RA-3DMM) model is proposed. RA-3DMM consists of a series of 3DMMs of different resolutions. RA-3DMM can select the best 3DMM according to the resolution of the input image. 3DMM has achieved promising performance in face recognition [1, 2]. The reason is that 3DMM has the capacity to capture general information about the natural variation of 3D shape and texture of faces in a vector space spanned by a database of 3D faces. This information helps to solve the ill-posed problem of reconstructing 3D shape from a single 2D image. The proposed model extends the applications of the approach to low resolution face recognition. Experimental results show the benefits of this model in pose-invariant face recognition.

The outline of the paper is as follows: Section 2 briefly explains the theory of 3D morphable model. We propose the resolution-aware 3D morphable model (RA-3DMM) in Section 3. Section 4 presents the experiments. In the first part we demonstrate how to construct the RA-3DMM. The second part presents the results of applying RA-3DMM to pose-invariant face recognition.

2 3D morphable model

The 3D Morphable Model (3DMM) was proposed by Blanz and Vetter [1]. The main application of 3DMM is to synthesize a 2D face image, which resembles the input face image. It is a parametric model based on a vector space representation of faces. This space is constructed so that any convex combination of shape and texture vectors belonging to the space describes a human face. Given a single face image, the fitting algorithm [1] automatically estimates 3D shape, texture, and all relevant 3D scene parameters like pose, illumination, etc. For constructing the model, the input 3D scans have to be preprocessed to establish the dense point-to-point correspondence. This preprocessing step is named registration, after which 3D meshes can be treated as vectors of the same dimensionality. The number of vertices of the 3D mesh determines the resolution of 3DMM. The face scans used by Blanz and Vetter are represented in cylindrical coordinates, and an optical flow algorithm is used to create dense correspondence [1]. Tena et al. proposed an alternative registration strategy called the Iterative Multi-Resolution Dense 3D Registration (IMDR) [9]. In this work, IMDR is used for registration. After registration, the information of shape S_0 and texture T_0 is conveyed by:

$$S_0 = (x_1, y_1, z_1, ..., x_n, y_n, z_n)$$

$$T_0 = (R_1, G_1, B_1, ..., R_n, G_n, B_n)$$

where n denotes the number of vertices. (x_n, y_n, z_n) is the coordinate of the nth vertex. (R_n, G_n, B_n) is the RGB value of the nth vertex. Once the dense correspondence of all 3D scans is created, PCA is performed separately on shape and texture vectors to decorrelate the data:

$$S = \bar{s} + \sum_{i=1}^{m-1} \alpha_i \cdot s_i, \quad T = \bar{t} + \sum_{i=1}^{m-1} \beta_i \cdot t_i \tag{1}$$

where $\bar{s}$ and $\bar{t}$ are the mean of shape and texture, respectively. s_i and t_i are the eigenvectors of shape and texture covariance matrices [1]. Different $\vec{\alpha} = [\alpha_1, ...\alpha_m - 1]$ and $\vec{\beta} = [\beta_1, ...\beta_m - 1]$ determine different S and T. $\vec{\alpha}$ and $\vec{\beta}$ are the main parameters optimised during the fitting process.

Once the model is constructed, the model can be used for fitting to a given 2D image. The primary goal of fitting is to minimise the sum of square differences over all colour channels and all pixels between the synthesized image and the input image,

$$E_I = \sum_{x,y} \| I_{input}(x,y) - I_{model}(x,y) \|^2 \tag{2}$$

where E_I is called the cost or energy function. $I_{input}(x,y)$ is the intensity of the input image at pixel(x,y). $I_{model}(x,y)$ is the model's texture value projected from 3D space to 2D image coordinates (x,y) [9]. The fitting problem is an ill-posed problem. In general, such an energy function would be highly non-convex with numerous local minima, which makes the fitting difficult. In [10] Romdhani and Vetter proposed an alternative fitting strategy called multiple features fitting (MFF) strategy. In this strategy, besides pixel intensity, other image features, such as the edges, the location of highlights and texture smoothness, are used for constructing a cost function. The resulting cost function is smoother and exhibits fewer local minima. Consequently, this multi-feature strategy has a wider radius of convergence and achieves a higher level of precision. In this work, this MFF strategy is adopted.

3 Resolution-aware 3D morphable model

To construct the mesh for a 3DMM, an upsample technique is applied [9]. It means any newly constructed 3DMM is upsampled from a low resolution 3D mesh called generic model. The generic model is upsampled with the 4-8 mesh subdivision algorithm [12]. The more subdivisions are performed, the more vertices and polygons the 3DMM will have. Thus the resolution of 3DMM can be controlled by the number of subdivision iterations.

The resolution-aware 3D morphable model (RA-3DMM) is proposed in this work. The construction of RA-3DMM is motivated by the assumption that **a high resolution 3DMM fits high resolution input images better and a low resolution 3DMM fits low resolution input images better**. The validity of the assumption will be verified in Section 4.2. Based on this assumption, a set of 3DMMs of different resolution should work better than a single

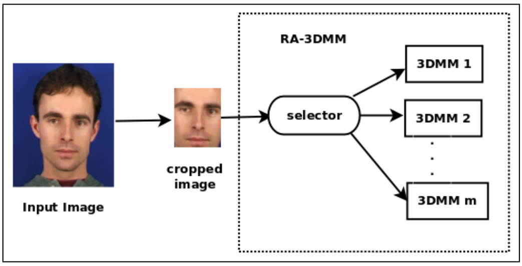


Figure 1: RA-3DMM

3DMM if the input images are of different resolutions. The structure of RA-3DMM is shown in Fig. 1.

From Fig. 1, RA-3DMM consists of *m* 3DMMs of different resolution. The model selector automatically selects the best 3DMM to fit the input face image according to its resolution. Clearly, the selector is the key part of RA-3DMM. In Section 4.2, the selection strategies will be determined experimentally on two face databases.

In this work, RA-3DMM consists of these 3DMMs: High Resolution 3DMM(HR-3DMM), Medium Resolution 3DMM(MR-3DMM) and Low Resolution 3DMM(LR-3DMM). Table 1 shows the details of these three models.

Table 1: The details of RA-3DMM

Model	vertices	polygons
HR-3DMM	29587	59763
MR-3DMM	16759	33211
LR-3DMM	1724	3368

4 Performance evaluation

4.1 Experimental Setup

We evaluated the proposed RA-3DMM model on two face databases: XM2VTS [8] and PIE [11]. XM2VTS contains face images of 295 subjects with 5 different poses under uniform illumination conditions. PIE consists of face images of 68 subjects. Each person was imaged across 13 different poses, under 43 different illumination conditions and with 4 different expressions. For building the RA-3DMM (see Section 4.2), only frontal images in the two databases are used. XM2VTS database does not include any illumination variations. The images are recorded under diffused light. Although images in PIE are recorded under different in-door controlled lights, in our experiments, we only use images of PIE taken under frontal illumination. In order to evaluate the performance of RA-3DMM, input images should be at different resolutions. So the images in XM2VTS and PIE are down-sampled to different resolutions. In this work, the down-sample rate (DSR) is 1, 1/2, 1/4, 1/6, 1/8, 1/10. Table 2 shows the average facial region pixel numbers in the two databases. All these down-sampled images were fitted by HR-3DMM, MR-3DMM and LR-3DMM. In order to

Table 2: face resolution with different DSR

DSR	1	1/2	1/4	1/6	1/8	1/10
XM2VTS	58061	14515	3629	1613	907	584
PIE	43546	10886	2721	1210	680	435

measure the fitting performance, we use the L1-Norm to estimate the fitting error

$$err = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{3}(|R_{i1} - R_{i2}| + |G_{i1} - G_{i2}| + |B_{i1} - B_{i2}|) \tag{3}$$

where R_{i1}, G_{i1}, B_{i1} are the RGB values of the input image, R_{i2}, G_{i2}, B_{i2} are the RGB values of the fitted image, and N is the number of facial pixels. The greater the err is, the worse the fitting is. In order to improve the fitting efficiency, not all of the 3D polygons of 3DMM are used for optimisation. Only 1000 polygons were selected in this work according to Romdhani's suggestion [10]. Gaussian filtering was applied to the input image for denoising and smoothing [9]. The higher the resolution of the input image is, the greater the kernel size of the Gaussian filter should be. We set the Gaussian kernel size in pixels empirically as follows:

Table 3: Kernel size settings

DSR	1	1/2	1/4	1/6	1/8	1/10
KS	7	5	3	1	1	1

4.2 Building RA-3DMM

The aim of this experiment is to verify the assumption in Section 3 and then train RA-3DMM to develop the model selection strategies mentioned in Section 3.

Before presenting quantitative results, some fitting examples are first shown in Fig. 2. Once the model fitting has been accomplished, the recovered parameters ($\vec{\alpha}$ and $\vec{\beta}$ in Equ.1) are used to render the face in any desired resolution, pose or illumination. Fig. 2 shows the fitting results with different models.

The first row shows the low resolution fitting results. The low resolution input image (top left) with DSR=1/10 is from XM2VTS (ID=003). It is not hard to see that the top middle reconstruction does not look like the input image, but the top right one does. It suggests that the LR-3DMM works better than HR-3DMM if the input image is of low resolution. The reason for this is that the use of HR-3DMM leads to over-fitting the low resolution input image. The second row shows the high resolution fitting results. The bottom left one is the high resolution input image with DSR=1. Both fitted images look like the input image, but there are more texture details in the bottom middle one than the bottom right one. Thus LR-3DMM yields an inferior result to that of HR-3DMM for fitting high resolution input images because it is under-fitting them.

Quantitative experimental fitting results on XM2VTS are shown in Fig. 3(a). It is clear that different models perform differently with input images of different resolutions: Obviously, HR-3DMM works best with DSR=1, 1/2; MR-3DMM with DSR=1/4, 1/6; and LR-3DMM works best with DSR=1/8, 1/10. We also note that the L1 error of these 3 models diminishes with decreasing DSR. It means that lower resolution input images with less information and less noise are smoother and easier to fit. However, the L1 error for DSR=1/10

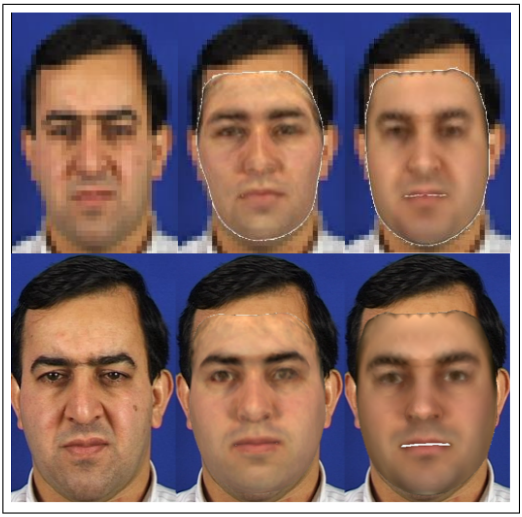
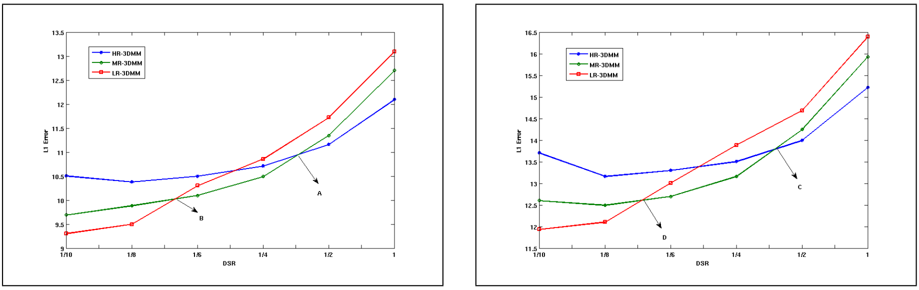


Figure 2: Comparisons of the fitting results with low resolution input image (top row) and high resolution input image (bottom row). Top left: low resolution input image, top middle: fitted result with HR-3DMM, top right: fitted result with LR-3DMM, bottom left: high resolution input image, bottom middle: fitted result with HR-3DMM, bottom right: fitted result with LR-3DMM

obtained using HR-3DMM is not smaller than that for DSR=1/8 because the model is overfitting if its resolution is much higher than that of the input image. As discussed in Section 3, the selection strategies are important for RA-3DMM. Specifically, it is important to know point A and point B in Fig. 3(a) to define the selection strategies. Actually, the resolutions that point A and point B correspond to are more interesting and useful than DSR for RA-3DMM. Combining the information of Table 2, it is not hard to conclude that the resolutions corresponding to A and B are 9397 and 1344 respectively. The fitting results on PIE are shown in Fig. 3(b). Similar to Fig. 3(a), HR-3DMM works best with DSR=1, 1/2; MR-3DMM works best with DSR=1/4, 1/6; and LR-3DMM works best with DSR=1/8, 1/10. Also, HR-3DMM at point DSR=1/10, and MR-3DMM at point DSR=1/10 are over fitted. Using the same argument, the resolutions corresponding to C and D are 7422 and 985 respectively. Note that the error for every DSR in Fig. 3(b) is greater than that of Fig. 3(a). The reason is that the images in XM2VTS are under diffused light, whereas the images in PIE are obtained under point source light, and hence are more difficult to fit.

In summary, the assumption in Section 3 has been verified in this experiment. The selection strategies of RA-3DMM are determined as follows: under the diffused light (such as XM2VTS), HR-3DMM is selected for fitting if the input image is of high resolution (greater than 9397 pixels). MR-3DMM is selected for fitting if the input image is of medium resolution (between 1344 pixels and 9397 pixels). LR- 3DMM is selected for fitting if the input image is of low resolution (smaller than 1344). Under the point source light (such as frontal illumination in PIE), HR-3DMM is selected for fitting if the input image is of high resolution (greater than 7422 pixels). MR-3DMM is selected for fitting if the input image is of medium resolution (between 985 pixels and 7422 pixels). LR-3DMM is selected for fitting if the input image is of low resolution (smaller than 985).



(a) (b)

Figure 3: Fitting Results on (a) XM2VTS and (b) PIE

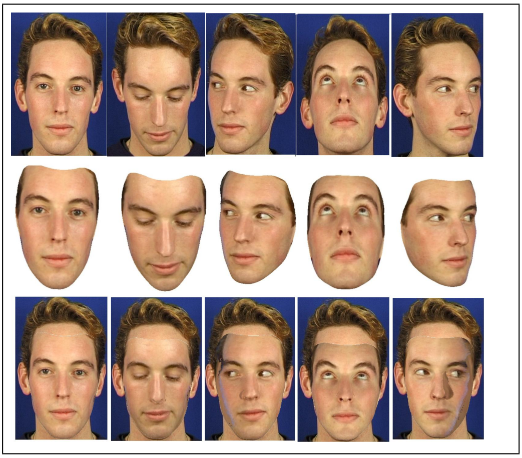


Figure 4: From the original images (top row), we recover shape and texture parameters for reconstructing the new images (second row), which extract textures from the input images. The bottom row shows the pose-corrected images. Note that we use the same background from the frontal image.

4.3 Experiments in face verification with RA-3DMM

3DMM has been used successfully in face recognition [1, 2]. Thus it is natural to apply RA-3DMM to face recognition. Traditional 2D methods suffer from ill-posed problems because it is impossible to register images of arbitrary poses. However, 3DMM can generate frontal face images from non-frontal images through a viewpoint-transform [2]. The normalised frontal images are then used for face recognition. However the frontal images reconstructed using the estimated parameters $\vec{\alpha}$ and $\vec{\beta}$ will miss some local information[2], such as moles or scars, because the texture parameter (T_i) cannot reproduce all local characteristics. But these local characteristics will affect face recognition, so we extract the person's true texture from the image wherever it is visible. Any occluded regions are reconstructed with the estimated texture parameters. One example of pose nomalisation is presented in Fig. 4. Note that the examples in Fig. 2 are reconstructed by 3DMM shape and texture parameters, but the examples in Fig. 4 extract true textures from the input images. So there are more local textures with the examples in Fig. 4 than in Fig. 2.

We apply RA-3DMM to face verification with input images ranging from high resolution to low resolution. The texture of the visible parts is extracted from the input image, and RA-3DMM automatically selects the best model to reconstruct the occluded part as discussed above. In this experiment, XM2VTS MPEG7 and the standard frontal datasets are used in conjugation with Configuration 1 of the Lausanne protocol[3]. This combined dataset contains 4100 images of 295 subjects. According to the protocol, the training and evaluation sets are taken from the standard frontal datasets. In the test set, each subject contains 10 images of 5 different poses (see Fig. 4). Table 4 provides a summary of the data used for each step of the evaluation protocol.

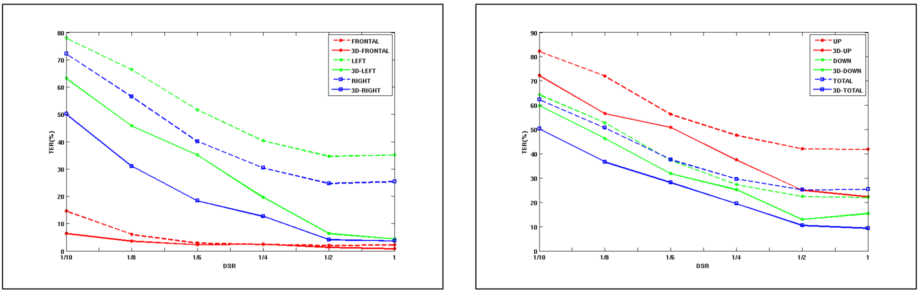
Table 4: Number of image access of each dataset in Configuration 1

Number of training samples	600 (3 × 200 client subjects)
Evaluation Client accesses	600 (3 × 200 client subjects)
Evaluation Imposter accesses	40,000 (25 imposter subjects × 8 shots × 200 client subjects)
Test Client access in MPEG7	2000 (200 client subjects×5 poses × 2 samples)
Test Imposter access in MPEG7	140000 (70 imposter subjects × 5 poses × 2 samples × 200 client subjects)

To evaluate the performance in different resolutions, we follow our previous experiment to down-sample the images to 6 different image sizes. In order to be robust to the blur caused by downsampling, the Multiscale Local Phase Quantisation histogram (MLPQH) [3] descriptor is used for image representation.

First, the image is geometrically normalised to 142 rows×120 columns based on the manually annotated eye positions. To perform the LPQ extraction, quadrature filters are convolved with the image and then zero crossing is detected in the filtered images to create a 3rd-order binary image tensor. The pixel value of the LPQ image is the decimal number obtained by converting the pixel value of the image tensor. The MLPQ images are extracted from a set of quadrature filters in a mulitscale manner. In each scale, the image is divided into 25 non-overlapping regions and the regional histogram is extracted as a regional descriptor. In each region, the multiresolution regional (MLPQH) descriptor is formed by concatenating the regional descriptors in different scales to a single vector. In order to reduce the redundancy of information and improve the discriminative power, Linear Discriminant Analysis (LDA) is applied. The regional LDA (MLPQHLDA) is trained on the training set without downsampling. In the matching process, the nomalised correlation is used to measure the similarity between a pair of images projected to the MLPQHLDA space. Following the Lausanne protocol, the total error rate (TER) is reported. TER is the sum of false acceptance rate (FAR) and false rejection rate (FRR) at a threshold. The threshold is chosen on the evaluation set at FAR=FRR using full-size images.

Fig. 5 shows the face verification results. It compares the performance with RA-3DMM pose correction and without pose correction. The TER of all poses with RA-3DMM pose correction (3D-TOTAL in Fig. 5(b)) is much smaller than that without pose correction(TOTAL) for all resolutions. Even for low resolution face verification, which is a hard task in face recognition, RA-3DMM shows considerable improvement over the method without pose correction. In particular, the verification performance with RA-3DMM pose correction is much better than that without pose correction for left-rotated and right-rotated images, indicating that RA-3DMM could reconstruct the occluded regions successfully for these 2 poses. The RA-3DMM based up and down pose correction also works better than that without pose correction, but worse than left and right pose correction. The reason is that there are some



(a) (b)

Figure 5: Face verification results of (a) frontal, left and right poses (b) up, down and total poses. 3D-LEFT means the left-rotated input images are pose-nomalised by RA-3DMM, and then used for face verification, while LEFT means the original left-rotated input images are directly used for face verification, and so on.

regions, such as eye balls of the down pose in Fig. 4, where there is no information. So it is impossible to reconstruct these parts successfully. For frontal images, there is not much difference with and without pose correction because these images have no occluded regions.

5 Conclusions

In this work, a RA-3DMM model is proposed. RA-3DMM, which consists of three 3DMMs, can select automatically the best 3DMM to fit input images of different resolutions. We validated the assumption that a high resolution 3DMM fits high resolution input images better and a low resolution 3DMM fits low resolution input images better. The face recognition performance with RA-3DMM pose correction is much better than that without pose correction for all resolutions.

Acknowledgement This work was carried out as part of EPSRC project ACASVA (Adaptive cognition for automated sports video annotation) under contract EP/F069421/1. The EPSRC financial support is gratefully acknowledged.

References

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, sept. 2003.

[2] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 454–461. IEEE, 2005.

[3] C.H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen. (multiscale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 633–640. IEEE, 2009.

[4] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, jun 2001.

[5] G. Dedeoğlu, S. Baker, and T. Kanade. Resolution-aware fitting of active appearance models to low resolution images. *Computer Vision–ECCV 2006*, pages 83–97, 2006.

[6] J. Kittler, A. Hilton, M. Hamouz, and J. Illingworth. 3d assisted face recognition: A survey of 3d imaging, modelling and recognition approaches. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 114–114. IEEE, 2005.

[7] X. Liu, P. Tu, and F. Wheeler. Face model fitting on low resolution images. In *Proc. 17th British Machine Vision Conference, Edinburgh, UK*, volume 3, pages 1079–1088, 2006.

[8] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, et al. Face verification competition on the xm2vts database. In *Audio-and Video-Based Biometric Person Authentication*, pages 1056–1056. Springer, 2003.

[9] J. T. Rodriguez. *3D Face Modelling for 2D+3D Face Recognition*. PhD thesis, Surrey University, Guildford, UK, 2007.

[10] S. Romdhani. *Face image analysis using a multiple features fitting strategy*. PhD thesis, University of Basel, Switzerland, 2005.

[11] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. 5th IEEE International Conference on*, pages 46–51, 2002.

[12] L. Velho and D. Zorin. 4-8 subdivision. *Computer Aided Geometric Design*, 18(5): 397–427, 2001.