3D-assisted Facial Texture Super-Resolution

Pouria Mortazavian
p.mortazavian@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
http://www.ee.surrey.ac.uk/cvssp

Abstract

In this paper we propose a new framework for super-resolving facial images under arbitrary pose. While example-based super-resolution methods have demonstrated impressive results for face super-resolution under given pose and imaging conditions, they have limitations dealing with different poses and illuminations. Due to these limitations their application to face super-resolution in generalized situations is either impractical or sub-optimal.

The proposed framework utilizes a 3D morphable face model in order to address the problem of face super-resolution under arbitrary pose. This framework does not assume any pre-defined pose for the subject thus it can be readily applied to any pose.

The main contribution of this work is defining a framework in which a 3D morphable model can be used in conjunction with any of the example-based super-resolution methods. Experimental results prove the potential power of this method in face super-resolution and its application to face recognition.

1 Introduction

Super-resolution methods produce higher-resolution images of a scene given one or more images of the same scene with lower resolutions. Many methods have been proposed for super-resolution which can be categorized into reconstruction-based (eg. [9, 16, 19]) and example-based (eg. [1, 7, 10]) methods. The example-based methods can further be categorized into general methods [6, 7] and domain-specific methods [1, 10, 20] which aim to super-resolve images of a given domain (eg. text or human face). In this work we are interested in the latter case where the problem is restricted to human face super-resolution.

Reconstruction-based super-resolution methods try to fuse information from multiple observations (low-resolution images) of the same scene to reconstruct a higher resolution representation of the scene. These approaches have had some success but are not applicable to cases where only a single observation is available. In such cases, example-based methods are used which utilize a number of exemplar images to learn or model the relationship between low-resolution and high-resolution images. Freeman *et al.* [6, 7] used a Markov Random Field (MRF) to probabilistically model the relationship between low and high resolution patches. By blurring and down-sampling a set of high-resolution images, the authors

BMVC 2009 doi:10.5244/C.23.119

generated a training set of sharp and blurred image pairs which is used to learn the compatibility functions of the MRF. They used this framework for general image super-resolution. However, as the authors also stated in their paper [7]:*"Without restriction to a specific class of training images, it's unreasonable to expect to generate the correct high resolution information"*. Nevertheless, in applications where the images belong to the same class (*e.g.* faces, or text), example-based super-resolution can considerably improve the results. Pickup *et al.* [14] used domain-specific texture samples to build a texture prior in the form of a p.d.f. which is used in a MAP framework for super-resolution. In this paper we are specifically interested in super-resolution of *facial* images.

Wang and Tang [20] proposed an eigentransformation approach in which the input image is represented as a linear combination of the training LR images using a PCA transform. The HR image is generated by replacing the LR images with their HR counterparts while keeping the mixture coefficients. Capel and Zisserman [4] propose another example-based, domain-specific method for faces based on PCA. Each face image is broken down into 6 regions and each region is represented with PCA. The PCA representation is then used in 3 different ways for super-resolution: an ML estimation constraining the solution to lie in the face sub-space spanned by the PCA components, a MAP estimate with the prior defined over the face sub-space and a MAP estimate with the prior defined over the whole image space which encourages the solution to lie *near* the face sub-space.

Liu *et al.* [12] proposed a two-step method for face super-resolution. This method is based on the fact that the HR image is a combination of common (global) face properties and individual characteristics. The global and individual properties are captured by a global parametric model (F^g) and a local non-parametric model (F^l) respectively. The final HR image (F) is assumed to be a combination of these two models: $F = F^g + F^l$. In this framework, F^g is obtained by a global parametric model and F^l is obtained by a local non-parametric MRF model. Li and Lin [11] have further improved this approach by applying PCA to both LR and HR images in the first step in order to incorporate the estimation of the noise model and also using a MAP framework instead of the MRF model for finding the optimal local face.

Baker and Kanade [1] present another example-based face super-resolution method. The registration parameters between images are assumed to be known *a priori* and only allow for horizontal and vertical translation. In a MAP formulation, an observation model in the form of equation 1 is used to define the likelihood term.

$$f_j(m,n) = \sum_{p,q} W(m,p,\Delta x_j)W(n,q,\Delta y_j)F(p,q) + \eta(m,n,j) \tag{1}$$

where f_j is the j^{th} LR image, Δx_j and Δy_j are the horizontal and vertical translations of the j^{th} input respectively, and F is the sought HR image. In equation 1, the weights $W(.)$ are a function of how much the LR pixel $f(m,n)$ and the HR pixel $F(p,q)$ overlap, and η is the pixel noise which is assumed to be *i.i.d.* Gaussian. A gradient prediction algorithm is used to define a recognition-based prior on the horizontal and vertical gradients of the HR image using a set of HR training images and their corresponding LR images from the training set. The parent structure vector [5] of each pixel of the input LR image is then compared with that of all pixels in the same spatial position in the LR database to find the training pixel with the closest parent structure. The corresponding training HR patch is then selected and the horizontal and vertical gradients in this selected patch are taken as a *predicted* value for the gradients of the sought HR image in the corresponding spatial location. The prior term is then defined to encourage the gradient values in the super-resolved image to be close to these predicted values by assuming Gaussian error for the gradient prediction algorithm.

The above methods generate an HR image of a single facial modality (i.e. at a given expression, pose, and illumination). In [10] Jia and Gong propose a *generalized* method to handle multi-modal face super-resolution across different poses, expressions, and imaging conditions. This example-based method utilises samples of different modalities in the training set. A tensor structure is used to incorporate information and interactions of the training images of different modalities at different resolutions. The super-resolution problem is formulated as a MAP estimation of HR images in multiple different modalities given an LR image of a single modality. Although this approach succeeds in handling different modalities, its performance depends on the availability of training data from different modalities. If for instance only frontal sample images (or even only a few poses) are available in the database, this method cannot super-resolve LR images with significantly different poses.

Most example-based methods need to register the input face properly with the faces available in the sample set. However, due to the 3D structure of the human face 2D registration methods may have limitations in registering the input with the samples in cases where the pose of the input image is significantly different from those available in the sample set. An alternative is to use a 3D model.

Yu *et al.* [21] used a generic 3D face model in a reconstruction-based framework. The input to their algorithm is a low-resolution video of a face with varying pose. After estimating the initial pose and illumination from the first frame, the generic 3D model's texture is updated sequentially by error back-projection in each frame.

While the approach in [21] is a reconstruction-based approach that relies on multiple input frames, we propose to use a 3D morphable model in an example-based framework which can also be used for single-frame super-resolution.

2 Proposed Approach

In example-based face super-resolution, the quality of the super-resolved face depends crucially on how representative the training set is and how well a specific super-resolution method can generalize it and utilize the available information. Due to these issues, example-based methods are limited in handling variations in the subject's pose or other imaging conditions. This is because it is impractical (virtually impossible) to represent all possible poses in a training database.

A powerful tool introduced by Blanz and Vetter [2] which can describe and synthesize human faces under a large range of poses and imaging conditions is the 3D morphable model. A 3D morphable model is a vector space representation of 3D faces. We used a morphable model built by Tena [15]. The model is built using 69 3D face scans. Each raw scan comprises a 3D mesh and a 2D texture map. The (x,y,z) coordinates of the vertices of the 3D mesh of each face scan and the corresponding texture values are concatenated into a *shape* (S) and a *texture* (R) vector respectively. Assuming these vectors are in full dense correspondence for all face scans, PCA is used to compress the data:

$$S_{mod} = \bar{S} + \sum_i \alpha_i s_i \qquad R_{mod} = \bar{R} + \sum_i \beta_i r_i \tag{2}$$

where $\bar{S}$ and $\bar{R}$ are the mean shape and texture vectors respectively and s_i and r_i are the i^{th} eigenvectors of the shape and texture covariance matrices respectively. Also, α_i and β_i

are the mixture coefficients known as the model shape and texture parameters respectively.

Given a single 2D face image as input and a set of landmarks, the parameters of the 3D morphable model can be estimated such that they represent the 3D shape and texture of the input face. This process is called model fitting which estimates the model parameters together with a set of rendering parameters -including lighting parameters, 3D translation, pose angels, camera focal length etc.- such that rendering the model with the estimated parameters will produce an image which resembles the input face image. Model fitting can be formulated as a MAP estimation of the model and rendering parameters given the input face image and the landmarks. Assuming independence between some parameters:

$$\alpha^*, \beta^*, \rho^* = \underset{\alpha,\beta,\rho}{\operatorname{argmax}}\, P(\alpha,\beta,\rho|f,L) = \underset{\alpha,\beta,\rho}{\operatorname{argmax}}\, P(f|\alpha,\beta,\rho)P(L|\alpha,\beta,\rho)P(\alpha)P(\beta)P(\rho) \tag{3}$$

where f is the input face image, L is a set of landmarks marked on f, α and β are the model shape and texture parameters respectively, and ρ is the set of rendering parameters.

In order to use the real facial texture from the input image (as opposed to the texture *estimated* by model fitting), the shape and rendering parameters estimated during the model fitting can be used to extract the facial texture from the input image (where available) and map it to a predefined texture coordinate frame. This coordinate frame is independent of the initial subject pose in the input image. The texture mapping process yields a pose- and shape-normalised 2D facial texture. Figure 1 illustrates this process.

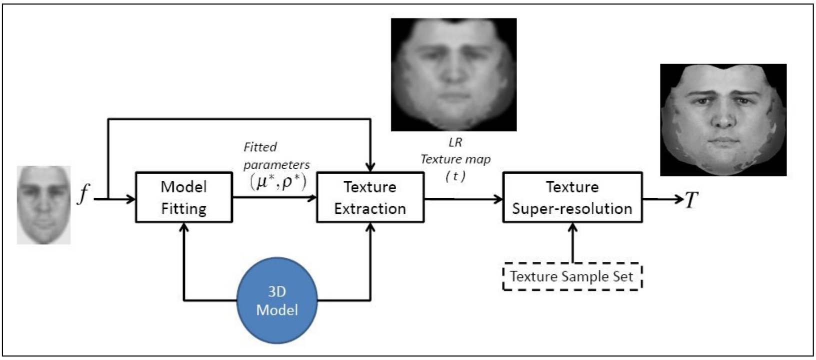

Figure 1: Our proposed method for 3D-assisted facial texture super-resolution

It is on this texture image that we propose to perform example-based super-resolution. That is, we first build a sample set by fitting a 3D morphable model on our training facial images and extracting texture from them. Then, given an input LR image, we super-resolve its texture by fitting the model on the input LR image, extracting the texture and super-resolving the extracted texture. The super-resolved texture can then be used together with the estimated shape parameters and arbitrary rendering parameters to render a new HR version of the face in the same or a different pose. Although the estimated shape parameters represent a low-resolution shape, the final rendered image will have a high-resolution appearance since it is rendered by mapping HR texture.

The above proposed method effectively means that the face image registration in conventional face super-resolution will be replaced by *model fitting* and *texture extraction*. Thus the accuracy of the model fitting is of crucial importance. Fitting a 3D morphable model on a very low-resolution and small face yields unacceptable results but we found good performance by fitting the model to a bilinearly interpolated version of the input LR image. By fitting the model on a bilinearly interpolated face we gain sub-pixel accuracy in the model fitting in some sense.

Our approach is fundamentally different from the approach of Yu *et al.* in [21] in that by mapping the facial texture from the input image to a texture domain, our approach allows the use of example-based methods for super-resolution which enables it to perform single-frame super-resolution, whereas the approach in [21] is a reconstruction-based method that relies on multiple input frames for super-resolution.

2.1 Formulation

Let μ be the set of model parameters ($\mu = \{\alpha, \beta\}$), ρ be the rendering parameters reflecting the imaging conditions, T be an HR texture map and f be the input LR face image. The sought HR texture map, T^* is the texture map that maximises the following marginalised probability:

$$T^* = \underset{T}{\operatorname{argmax}} \sum_{\mu,\rho} P(T,\mu,\rho|f) = \underset{T}{\operatorname{argmax}} \sum_{\mu,\rho} \{P(T|\mu,\rho,f)P(\mu,\rho|f)\} \tag{4}$$

Although more appropriate ways of finding T^* from equation 4 are possible, in this paper we make some simplifying assumptions in our formulation in order to show that even in the simplest case our approach can yield impressive results. The first assumption is that $P(\mu,\rho|f)$ peaks at the optimum values of the model and rendering parameters and it has a dense distribution around these values. In other words, we are assuming that the distribution of $P(\mu,\rho|f)$ can be estimated by $\delta(\mu-\mu^*)\delta(\rho-\rho^*)$ where μ^* and ρ^* are the optimum model and rendering parameters respectively which are given by fitting the model on f:

$$\{\mu^*,\rho^*\} = \underset{\mu,\rho}{\operatorname{argmax}}\, P(\mu,\rho|f) \tag{5}$$

Given the above assumption, equation 4 reduces to

$$T^* = \underset{T}{\operatorname{argmax}}\, P(T|\mu^*,\rho^*,f) \tag{6}$$

where μ^* and ρ^* are given by 5. Thus, the whole process of optimising equation 4 can be thought of as two separate steps of model fitting (second term) and texture super-resolution (first term).

The second simplifying assumption is that T can be completely described through an intermediate *low-resolution texture map*, t, which is extracted from the LR input face using the fitted parameters. Thus the distribution of $P(T|\mu^*,\rho^*,f)$ can be replaced with $P(T|t)$:

$$T^* = \underset{T}{\operatorname{argmax}}\, P(T|\mu^*,\rho^*,f) = \underset{T}{\operatorname{argmax}}\, P(T|t) = \underset{T}{\operatorname{argmax}}\, P(t|T)P(T) \tag{7}$$

where:

$$t = \text{TEXTURE\_EXTRACT}(\mu^*,\rho^*,f) \tag{8}$$

Following the above discussion, our algorithm consists of three steps:

1. **Model Fitting**: The optimum values of the 3D morphable model together with the optimum values of rendering parameters are obtained by fitting the model on the input image (equation 5).

2. **Texture Extraction**: The texture from the LR input face is extracted and mapped to the texture domain (equation 8).

3. **Texture Super-resolution**: Super-resolution in the texture domain is formulated as a MAP estimation problem (equation 7).

In this paper we use the approach of [3] for the first step. The model fitting problem is posed as a MAP estimation (equation 3) of the model and rendering parameters, given the input image and a set of feature points marked on it. Using a Bayesian approach, a cost function is defined that includes a term to penalise the error between the input image and the image rendered by the model, another term to penalise the error between the location of the marked feature points and the location of their equivalent points on the 3D model when projected to the image plane, as well as 3 terms that constrain the values of the shape, texture and rendering parameters. This cost function is then minimised with respect to α, β, and ρ using gradient descent.

For the texture mapping step we use the approach of [15]. In order to map the texture from the input image to a texture map (step 2 of the algorithm), a 2D map is needed on which the polygons of the model are shown with their proportions preserved. This is achieved by a 2D representation of the generic 3D model using the *isomap* algorithm [18] which preserves the geodesic distances between the vertices. The texture coordinate frame is then defined by overlaying a grid on this 2D map. Next, the polygons of the fitted model are projected to the image plane and tested for visibility using a z-buffer test. If a triangle is visible, the texture values of all texels which fall within that triangle on the 2D texture coordinate frame are copied from the image to the texture map defined in the texture coordinate frame. The texels on the texture map that correspond to triangles that are not visible on the image are filled with texture values estimated in the model fitting step.

As for the third step of the algorithm, we use Baker and Kanade's well-known hallucination method [1] with a minor change in the comparison of input texel with sample set texels by using not only the texel at hand but also its neighbourhood to find the closest match. The parent structure vector which is used in this framework is built by stacking together the laplacian image and horizontal and vertical first and second derivatives of different scales. Note that in Baker and Kanade's method the generative model is defined to estimate the process of forming an LR *image* given an HR *image* whereas here we actually need a model to describe the process of forming an LR *texture map* given a high-resolution one. However, we have assumed here that the same observation model can be used in the texture domain which essentially means that we are neglecting the actual way that the texture maps are formed in this step and treating them as if they were ordinary 2D *images*.

3 Experimental Results

In this section we present some results for face super-resolution in different poses. Furthermore, we present recognition results with and without super-resolution and compare our method to the benchmark work of Baker and Kanade [1].

The experiments in this section were performed on the XM2VTS database [13]. 95 subjects were chosen randomly as the sample set for the super-resolution algorithm. The sample set was built by fitting the morphable model on 8 frontal shots of each sample subject. The LR test data was generated by fitting the model on HR shots of the other subjects, extracting the texture, and rendering the results while setting the focal length value to 8 times smaller than what was estimated for the HR image in order to produce lower-resolution images. In all cases the zoom factor is 8, i.e. each LR image will be enlarged 8 times in each direction.

3.1 Frontal face super-resolution

Figure 2 shows the results for a frontal face super-resolution and compares our results with Baker and Kanade's method. This figure shows that super-resolving the face in the texture domain (our method) can yield results that are at least as good as the ones hallucinated in the image domain (Baker and Kanade's method) with the added advantage that, once the texture is super-resolved, the same texture can be used in conjunction with the fitted model to render novel views of the same face (Figure 3)

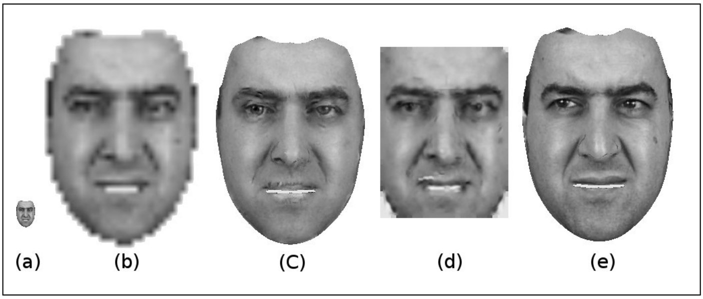

Figure 2: (a) LR image (b) Bilinear interpolation (c) Our method (d) Baker and Kanade's method (e) HR image

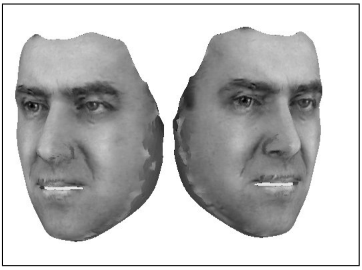

Figure 3: 2 novel super-resolved views of the LR face in Figure 2(a)

Figure 4 shows some more examples. The first row in this figure shows LR images enlarged by bilinear interpolation; the second row illustrates the results of Baker and Kanade's method while the third and fourth rows show our results and the HR images respectively.

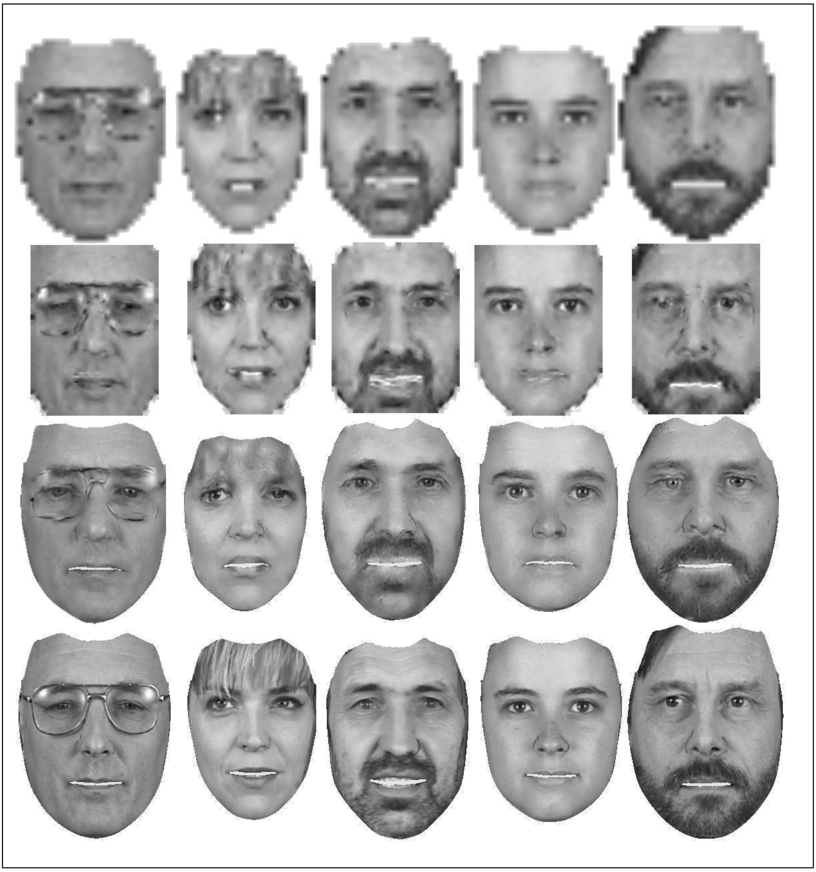

Figure 4: Super-resolution of frontal faces. Top row: Bilinear interpolation, second row: Baker and Kanade's method, third row: Our method, fourth row: HR images.

3.2 Non-frontal face super-resolution

The advantage of our method is that it can be used with no change and with the same sample set for super-resolving non-frontal faces. Figure 5 shows an example of super-resolving a non-frontal face. The super-resolved texture is rendered using the fitted model and superimposed on the bilinearly interpolated version of the LR face.

Note that the results in Figure 5 were generated using a sample set that only consisted of frontal face images. This is to show that even without using other poses our method manages to do an acceptable job in super-resolving different poses. However, the performance can be further enhanced by adding samples of other poses to the sample set.

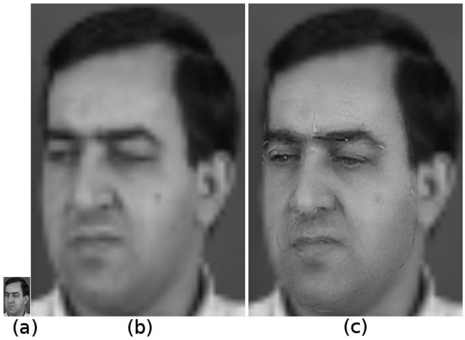

Figure 5: Non-Frontal Super-resolution

3.3 Face Recognition

One of the challenges in face recognition is low resolution of input images. The performance of face recognition algorithms is known to drop below a certain resolution. One remedy is to use super-resolution to increase the resolution of the input image before recognition. Although in [8] it is argued that better results could be obtained by simultaneous super-resolution and recognition, we resort to the traditional approach of performing super-resolution as a pre-processing step prior to recognition.

We use the face identification algorithm proposed by Shan *et al.* [17]. An ensemble of classifiers is built by dividing the image into 100 regions. Spatial histograms of Local Binary Patterns (LBP) are used in each region as feature vectors which are then projected into LDA space. A sum rule is then used to combine the similarity scores (normalised correlation) from the 100 classifiers. Nearest Neighbour(NN) is then employed for final classification. 140 subjects were chosen from the XM2VTS database (not including those already used as super-resolution sample set). The HR data for training the LDA transform was generated by fitting the morphable model to 3 training shots of each subject and rendering the model with focal length value of 1500. The test data was generated by fitting the model to another shot of the same subject and rendering it with focal length 1500 for HR test data and focal length 187.5 for LR test data.

Table 1 shows the identification rates. The high identification rate for HR test samples drops dramatically when bilinearly interpolated LR samples are used instead. This rate is improved significantly using super-resolution instead of interpolation. As table 1 shows, the identification rate obtained with our method is a few percents less than that obtained with Baker and Kanade's method. However, our approach has the capacity to super-resolve faces of arbitrary pose. In any case, we believe that the lower performance is caused by errors in fitting the model to LR images which our current work is attempting to rectify. Nevertheless, the fact that the identification rate is significantly increased with both super-resolution methods shows that the additional information injected is beneficial for recognition.

Method	Identification Rate
HR	99.28
LR+Bilinear	78.57
LR+Baker_Kande	96.43
LR+Our Method	93.57

Table 1: Face Recognition Results

4 Conclusions

In this paper we presented a framework for 3D-assisted facial texture super-resolution. A 3D morphable model is used to map facial texture to a shape- and pose-normalized texture map which is then super-resolved using an example-based approach. The super-resolved texture can then be used to render novel views of the face. The super-resolution results show that although the model fitting step can add additional error, the final super-resolved face has visually acceptable quality and it can also be used to enhance face recognition with low resolution inputs.

References

[1] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):1167 – 1183, September 2002.

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In Alyn Rockwood, editor, *Siggraph 1999, Computer Graphics Proceedings*, pages 187–194, Los Angeles, 1999. Addison Wesley Longman.

[3] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, September 2003. ISSN 0162-8828.

[4] David Capel and Andrew Zisserman. Super-resolution from multiple views using learnt image models. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:627, 2001. ISSN 1063-6919.

[5] Jeremy S. De Bonet and Paul Viola. A non-parametric multi-scale statistical model for natural images. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 773–779, Cambridge, MA, USA, 1998. MIT Press. ISBN 0-262-10076-2.

[6] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael Y. Learning low-level vision. *International Journal of Computer Vision*, 40:2000, 2000.

[7] William T. Freeman, Thouis R. Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65, 2002. ISSN 0272-1716.

[8] Pablo H. Hennings-Yeomans, Simon Baker, and B. V. K. Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *CVPR*, 2008.

[9] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graph. Models Image Process.*, 53(3):231–239, 1991. ISSN 1049-9652.

[10] Kui Jia and Shaogang Gong. Generalized face super-resolution. *IEEE Transactions on Image Processing*, 17(6):873–886, 2008.

[11] Yang Li and Xueyin Lin. An improved two-step approach to hallucinating faces. *Image and Graphics, International Conference on*, 0:298–301, 2004.

[12] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:192, 2001. ISSN 1063-6919.

[13] K Messer, J Matas, J Kittler, J Luettin, and G Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999.

[14] Lyndsey C. Pickup, Stephen J. Roberts, and Andrew Zisserman. A sampled texture prior for image super-resolution. In *In NIPS 16*, pages 1587–1594, 2003.

[15] J.R. Tena Rodriguez. *3D Face Modelling for 2D+3D Face Recognition*. PhD thesis, Centre for Vision, Speech, and Signal Processing, University of Surrey, November 2007.

[16] Richard R. Schultz and Robert L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5:996–1011, 1996.

[17] Shiguang Shan, Wenchao Zhang, Yu Su, Xilin Chen, and Wen Gao. Ensemble of piecewise fda based on spatial histograms of local (gabor) binary patterns for face recognition. *Pattern Recognition, International Conference on*, 4:606–609, 2006. ISSN 1051-4651.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. ISSN 0036-8075.

[19] Hanoch Ur and Daniel Gross. Improved resolution from subpixel shifted pictures. *CVGIP: Graph. Models Image Process.*, 54(2):181–186, 1992. ISSN 1049-9652.

[20] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 35(3):425–434, 2005.

[21] Jiangang Yu, Bir Bhanu, Yilei Xu, and Amit K. Roy Chowdhury. Super-resolved facial texture under changing pose and illumination. In *ICIP (3)*, pages 553–556. IEEE, 2007.