Recovering 3D Facial Shape via Coupled 2D/3D Space Learning

Annan Li[1,2], Shiguang Shan[1], Xilin Chen[1] , Xiujuan Chai[3], and Wen Gao[4,1]

[1]Key Lab of Intelligent Information Processing of CAS,
Institute of Computing Technology, CAS, Beijing 100190,China
[2]Graduate University of CAS, 100190,Beijing,China
[3]System Research Center, NOKIA Research Center, Beijing,100176,China
[4]Institute of Digital Media, Peking University, Beijing,100871,China
{anli,sgshan,xlchen,wgao}@jdl.ac.cn ext-xiujuan.chai@nokia.com

Abstract

This paper presents a method for recovering 3D facial shape from single image via learning the relationship between the 2D intensity images and the 3D facial shapes. With a coupled training set, the intensity images and their corresponding facial shapes make up two vector spaces respectively. But only the correlated components in both spaces are useful for inference, so there must be embedded hidden subspaces in each space which preserve the inter-space correlation information. Thus by learning the projection onto hidden subspaces based on Maximum Correlation Criteria and optimizing the linear transform between the hidden spaces, 3D facial shape is inferred from the intensity image. The effectiveness of the method is demonstrated on both synthesized and real world data.

1. Introduction

Shape recovery is a classic problem in computer vision. A popular approach to solve this problem using single image is shape-from-shading (SFS), which is usually with the assumption of Lambertian reflectance and a single point light source at infinity [1]. Unfortunately, for recovering facial shape from real world images, SFS has proved ineffective. Because the information in a single image is inadequate for recovering accurate facial shape. For this reason, approaches using multiple images such as photometric stereo [2] perform much better than SFS.

To tackle this problem, real facial shapes obtained from laser scanner are used as prior knowledge for facial shape recovery. Kemelmacher and Basri [3] used only one reference 3D facial shape as constraints in SFS framework. Their method is simple and effective, but the accuracy of the recovered facial shapes is dependent on the selection of the reference shape. Hassner and Basri [4] propose a similar example based approach using more facial shapes. They divide faces into shape-appearance patch pairs. Facial shapes are synthesized from the most referenced patch pairs. Besides the example based approaches, more methods build statistical model on facial shapes [5, 6, 7, 8, 9, 10] . Atick et al.[5] first used statistical 3D face model to enhance SFS. Under classical SFS framework, facial shapes are recovered from a set of eigenheads derived from principle component analysis. Dovgard and Basri [6] propose a statistical symmetric SFS approach that applied the statistical SFS framework of Atick et al. on the self-ratio image developed by Zhao and Chellappa [11]. The self-ratio image has two merits. First, it preserves the symmetry of human faces. Secondly, it is an albedo free formulation of shape and intensity. However human faces are not strictly symmetric objects, asymmetry of human faces results in errors in recovered shapes. Smith and Hancock [7] propose a statistic SFS approach that represents facial shape in surface normal domain. Surface normal directions are transformed into Cartesian points using azimuthal equidistant projection. Principle component analysis is applied to construct a statistical model of surface normal, and then this model is embedded in their SFS framework. Recently, they developed a new statistical model of surface normal using principle geodesic analysis. With a robust statistic model, this deformable model can be fitted to facial images with self-shadowing [8]. Different from the above approaches that use simple single point light source, Zhang et al. [9] combined statistical 3D shape model with the spherical harmonic illumination model [12]. With this illumination model, facial shape can be reconstructed under unknown lighting condition. In the facial shape reconstruction literature, the 3D morphable model (3DMM) proposed by Blanz and Vetter [10] is considered state-of-the-art. Statistical models are

978-1-4244-2154-1/08/$25.00 2008 IE

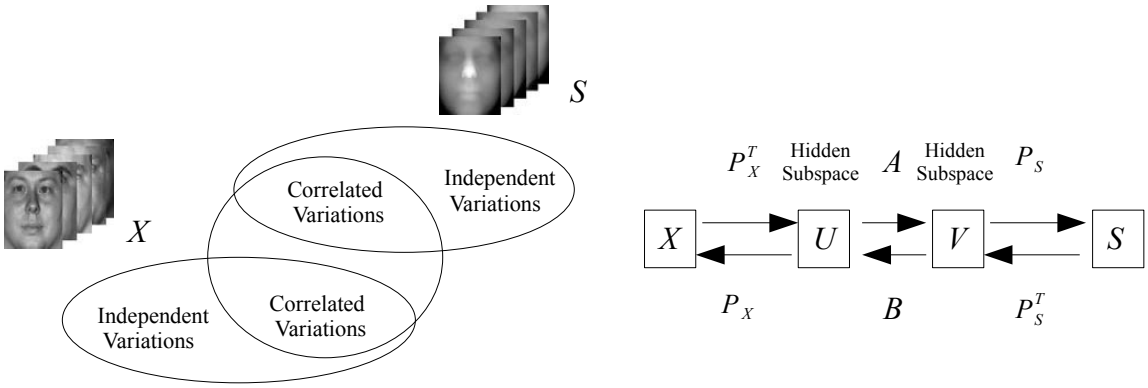

Figure 1. Framework of coupled 2D/3D space learning.

built for shape and texture separately. With a complex optimization process, facial shape and texture can be recovered across pose and illumination variations.

One commonality of the above-mentioned approaches is that they either use a single statistical model of facial shapes or two separate statistical models for shapes and appearances. In the training set they used, each facial shape can be coupled with a 2D appearance image. However, this useful information is ignored in these approaches. Recently researchers begin to realize this problem. To make use of the coupling information, Casteln et al. [13] proposed a coupled statistical model for facial shape recovery. The main idea of this coupled model is coefficients sharing. In the linear subspace an intensity image can be represented as a linear combination of images in the training set. For each intensity image has a corresponding 3D shape in the training set, the facial shape can be recovered using the same combination coefficients and replacing the intensity images by corresponding 3D shapes. However, the best fit coefficients in the intensity space are not necessarily the best fit coefficients in the shape space. To solve this problem, they performed similar coupled statistical model on the coefficients obtained from the original coupled model [14]. This approach is similar to the active appearance model proposed by Cootes et al. [15]. But in the active appearance model both the 2D shape and appearance to be fitted are known, while for facial shape recovery only the intensity image is known. The problem of coefficients mismatch still exists.

This paper aims at exploring the statistical relationship between 2D face appearance and 3D facial shape. We believe that in the appearance space and shape space only the components correlated to the other space are useful for inference. This explains why the best fit appearance coefficients different from the best fit shape coefficients. Discriminating the correlative components from independent components can improve the inference from appearance to shape. Based on similar observation on images of different resolution, Lin and Tang [16] proposed a coupled space learning framework for 2D face hallucination. We borrow this framework to explore the statistical correlations between 2D intensity appearance and 3D shapes of faces. With a set of coupled intensity images and facial shapes, two subspaces are constructed respectively. Assuming that the correlative components make up a hidden subspace, we learn the projections onto the hidden subspace. Then 3D facial shape can be recovered by projecting the intensity image to the shape space via the hidden subspace that maximizes the correlations between appearance and shape spaces.

The remaining part of this paper is organized as follows.In Section 2, we describe the coupled 2D/3D space learning framework in detail. Section 3 shows experimental results of facial shape recovery. Finally, we draw conclusions in Section 4.

2. 3D Facial Shape Recovery via Coupled 2D/3D Space Learning

In this section, we describe the coupled statistical model on 3D facial shapes and 2D appearances and how it can be used for facial shape recovery.

2.1. Framework of Coupling 2D Appearance Space and 3D Facial Shape Space

Given a 2D intensity facial image, our aim is to recover its surface height based on a coupled training set. Let $\{(X, S)\}$ be the training set, which is composed of the vectorized intensity images and their corresponding surface height of the same faces. Here, $X = \{x_1, x_2, x_3 \dots x_n\}$ denotes the intensity face image set composed of n subjects, and $S = \{s_1, s_2, s_3 \dots s_n\}$ is their corresponding facial shapes represented as surface height. And we denote

the dimensions of the image and shape spaces as d_x and d_s respectively. Based the observation that in one space only the components correlated to the other space contribute to inference, it is reasonable to assume that there is an intrinsic hidden space reflecting the inter-space correlations and the intensity and shape spaces are some transformed versions of the hidden space. Denote the hidden space as H and the vectors in hidden space as h, the transforms from hidden spaces to observed spaces as T_X and T_S.

$$x = T_x h + m_x \quad , \quad s = T_s h + m_s \tag{1}$$

Here, m_x denotes the mean vector of intensity vectors x and m_s denotes the mean vector of the facial height s. Assume the dimension of the hidden subspace is d, then the dimension of matrix T_X is $d_x \times d$ and the dimension of matrix T_S is $d_s \times d$. For analysis the composition of T_X and T_S, perform singular value decomposition on T_X and T_S, we have

$$T_X = U_X D_X V_X^T \quad , \quad T_S = U_S D_S V_S^T \tag{2}$$

Here, the dimension of U_X is $d_x \times d$, and the dimension of U_S is $d_s \times d$, while D_X, D_Y, V_X, V_Y are all $d \times d$ matrices. Considering equation (1), we have

$$\begin{aligned} U_X^T(x - m_x) &= D_X V_X^T h \\ U_S^T(s - m_s) &= D_S V_S^T h \end{aligned} \tag{3}$$

For right part of the equation, projecting $d \times d$ matrix D_X and V_X on d dimension vector h, the dimension of the final vector is still d. So, the right part of this equation can be considered as a rotated and scaled version of h. For the whole equation, it means that using matrix U_X vector $x - m_x$ can be projected to a subspace, which is the rotated and scaled version of the hidden subspace H. Lets denote this embedded subspace as U. Applying similar analysis to s, we can get a similar embedded subspace denoted as V. Then we have two d dimensional embedded subspaces associated with X and S respectively. To emphasize the projection role of U_X and U_S, we denote them as P_X and P_S. Consider that d_x and d_s may be different, directly solving the transform between two spaces is difficult and unstable. To avoid this problem, we add $d \times d$ matrix A and B as transform between U and V. Then the whole framework of Coupled Space Learning can be illustrated as Figure 1, and the transformation between appearance and shapes can be described as the following equations.

$$\begin{aligned} s - m_s &= P_S A P_X^T (x - m_x) \\ x - m_x &= P_X B P_S^T (s - m_s) \end{aligned} \tag{4}$$

Under this framework, the problem left is how to determine P_X, P_S, A and B. In the rest part of section 2 we describe the solutions in detail.

2.2. Maximize Correlations between Shapes and Appearances

Statistical dependency between shape space and appearance space is the foundation for inferring shape from intensity image. For two linear subspaces under Gaussian distribution assumption, statistical dependency is equivalent to correlation. In the coupled 2D/3D space model illustrated in Figure 1, there are two linear subspaces. Each subspace can be decomposed into two embedded subspaces. One is corresponding to the correlated variations, while the other is corresponding to the independent variations. Only the former is useful for inference. As described in section 2.1, using P_X and P_S a vector can be projected to a rotated and scaled version of this ideal embedded subspace. The closer the projected space to the ideal space, the better inference we can get. In another word, the projection P_X and P_S we pursue are those can maximum the correlation between U and V.

Denote the vectors in the projected hidden space as U and V as $\{u_1, u_2, u_3 \dots u_n\}$ and $\{v_1, v_2, v_3 \dots v_n\}$ respectively. They can be computed as follows

$$u_i = P_X^T(x_i - m_x) \quad , \quad v_i = P_S^T(s_i - m_s) \tag{5}$$

Under the assumption of Gaussian distribution, $x \sim N(m_x, C_x)$ and $s \sim N(m_s, C_s)$. For a pair of vectors u_i and v_i, their correlation can be measured in terms of covariance as

$$E[(P_X^T(x_i - m_x))(P_S^T(s_i - m_s))^T] = P_X^T C_{xs} P_S \tag{6}$$

Here, $C_{xs} = E[(x_i - m_x)(s_i - m_s)^T]$ is the covariance matrix between x_i and s_i .

Unfortunately the value of C_{xs} may be negative. It is not a symmetric semidefinite matrix. However it is the magnitude not the sign of the value represents the intensity of the correlation. For mathematical convenience, we can use the square of covariance as $Correlation\ Intensity$:

$$CI(P_X, P_S) = (P_X^T C_{xs} P_S)^2 \tag{7}$$

For hidden subspace U and V, the total correlation intensity can be derived as

$$\begin{aligned} CI(P_X, P_S) &= tr((P_X^T C_{xs} P_S)(P_X^T C_{xs} P_S)^T) \\ &= tr(P_X^T C_{xs} P_S P_S^T C_{sx} P_X) \\ &= tr(P_S^T C_{sx} P_X P_X^T C_{xs} P_S) \end{aligned} \tag{8}$$

For the given training set $\{X, S\}$, the total covariance matrix can be computed as $C_{xs} = \frac{1}{n}\tilde{X}\tilde{S}^T$ and $C_{sx} = \frac{1}{n}\tilde{S}\tilde{X}^T$. Here, $\tilde{X} = [x_1 - m_x, x_2 - m_x, x_3 - m_x \dots x_n -$

$m_x]$ and $\tilde{S} = [s_1 - m_s, s_2 - m_s, s_n - m_s, \ldots s_n - m_s]$ are the mean offset sample matrix.

Based on the above analysis, and deriving $Maximum\ Correlation\ Criteria$, the P_X and P_S we pursue can be described as

$$(P_X, P_S) = \arg \max_{P_X, P_S} CI(P_X, P_S) \tag{9}$$

Here,

$$\begin{aligned} CI(P_X, P_S) &= tr(P_X^T \tilde{X} \tilde{S}^T P_S P_S^T \tilde{S} \tilde{X}^T P_X) \\ &= tr(P_S^T \tilde{S} \tilde{X}^T P_X P_X^T \tilde{X} \tilde{S}^T P_S) \end{aligned} \tag{10}$$

2.3. Implementation Algorithm for 3D Facial Shape Recovery

Note that for equation (9) we have two unknown variations to be solved. Directly solving the objective function can not offer result for both P_X and P_S. Hence, an iterative framework is used. Using P_S to solve P_X and using P_X to solve P_S, repeat the process iteratively, then final solution can be obtained. The detailed procedure is described in Table 1:

1. Initialize $P_X^{(0)}$ and $P_S^{(0)}$to be identity matrices.
2. Iterate the following steps, at the t-th step:
   (a) Compute $S_X^{(t)} = \tilde{X} \tilde{S}^T P_S^{(t-1)} P_S^{(t-1)T} \tilde{S} \tilde{X}^T$
   (b) Update P_X by $P_X^t = \arg\max_{P_X} P_X S_X^{(t)} P_X^T$
   (c) Compute $S_S^{(t)} = \tilde{S} \tilde{X}^T P_X^{(t)} P_X^{(t)T} \tilde{X} \tilde{S}^T$
   (d) Update P_S by $P_S^t = \arg\max_{P_S} P_S S_S^{(t)} P_S^T$
   (e) Compute the correlation intensity $CI^{(t)}$ by equation (10)
3. Stop when $CI^{(t)} - CI^{(t-1)} < \epsilon$.

Table 1. Algorithm for solving P_X and P_S

Here, S_X and S_S are positive semidefinite matrixes. For step (b) and (d), P_X and P_S can be solved by performing eigenvalue decomposition on S_X and S_S, by using the d eigenvectors associated with largest eigenvalues as the column vectors of P_X and P_S respectively.

When P_X and P_S are determined, we can use them to solve A and B. First, we use P_X and P_S to project vectors in space X and S into the hidden subspace U and V according to equation (5). As illustrated in Figure 1, transforms between spaces U and V are bidirectional. This reflects the essence of coupling. It means that there is a balance between transform A and B, they should not be optimized separately. Transform accuracy should be measured not only for a single transform, but also for the backward transform. Based on this rationale, augmented matrixes are used to perform the negotiation between A and B. With an iterative framework similar to that solving for P_X and P_S, the algorithm is described as:

1. Initialize A and B by linear regression:
   $A^{(0)} = \arg\min_A ||V - AU||_F^2 = (VU^T)(UU^T)^{-1}$
   $B^{(0)} = \arg\min_B ||U - BV||_F^2 = (UV^T)(VV^T)^{-1}$
2. Repeating the following steps:
   (a) Compute the augmented matrix using $B^{(t-1)}$:
   $U_{aug} = [U, B^{(t-1)}V], V_{aug} = [V, V]$
   (b) Update A by $A^{(t)} = \arg\min_A ||V_{aug} - AU_{aug}||_F^2$
   (c) Compute the augmented matrix using $A^{(t)}$:
   $V_{aug} = [V, A^{(t)}U], U_{aug} = [U, U]$
   (d) Update B by $B^{(t)} = \arg\min_B ||U_{aug} - BV_{aug}||_F^2$
3. Stop when $||A^{(t)} - A^{(t-1)}||$ and $||B^{(t)} - B^{(t-1)}||$ are bellow some specified threshold.

Table 2. Algorithm for solving A and B

The whole procedure of coupled 2D/3D space learning can be summarized as two steps. First, based on the coupled training set, optimize projections P_X and P_S by maximizing the correlations intensity. After that, project the training set into the hidden subspaces using P_X and P_S, then optimize the transform A and B between the hidden subspaces.

When P_X, P_S, A, and B are all determined, facial shape can be recovered from input intensity image according to equation (4).

3. Experiments

In this section we give experiments of facial shape recovery on both synthesized data and real world data. For training the coupled statistical model, we use the USF Human-ID 3D face database [17] .This database contains shapes and textures of 138 human heads. The facial shapes are obtained from laser scanner. Each shape is associated with intensity texture of the same person. We use 100 heads for training the coupled statistical model, while the rest 38 heads are used for testing.

The facial shapes and the textures that originally represented in cylindrical coordinates in the database are converted to Cartesian grid of resolution 90×120 in frontal view. The facial shapes are represented as surface heights. The intensity images for training are synthesized using the textures and surface heights with frontal point light (0, 0, 1). To remove the influence of hair, the surface heights and intensity images are normalized by a facial region mask.

Figure 2 shows some results of facial shape recovery on the USF database using the remaining 38 heads. From left to right, the first and fifth column in Figure 2 are original intensity images with facial region mask. In the second and

sixth column, facial shapes recovered via coupled 2D/3D space learning are shown, while their corresponding ground truths are given in the third and seventh column. The recovered facial shapes and the ground truths are illuminated with frontal point light. Finally, error maps of height difference between recovered shapes and ground truths are given in the fourth and eighth column. To analysis the shape recovery accuracy, we compute the average height difference error as $||recovered\_surface - ground\_truth||/ground\_truth$. For the 38 heads the total average reconstruction error is 2.09%.

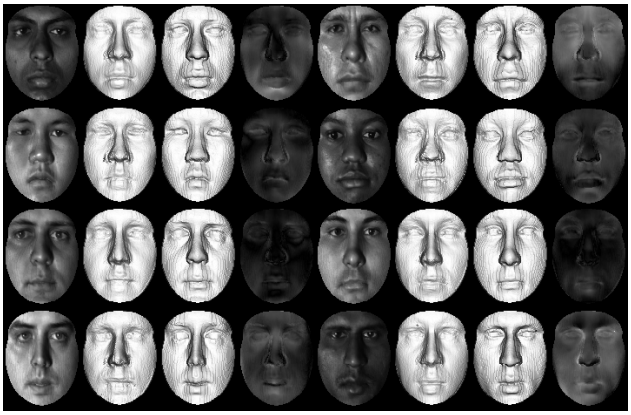

Figure 2. Example results of facial shape recovery on USF database

Besides the synthesized data, we also do experiments on the real world images. We use the same coupled statistical model obtained by training the 100 heads in the USF datasets. The images for testing are selected from the Extended Yale Face database B [2, 18] and CMU PIE database [19]. Images in the YaleB database are captured under different illumination conditions with pose variations. We select the images in frontal pose illuminated by a point light source of zero degree azimuth and zero degree elevation. The facial shape recovery results on YaleB are shown in Figure 3. From left to right, in the first two columns are input intensity images and the recovered facial shape. The rest four columns show the rendered faces using recovered shapes and their corresponding actual images under similar view points.

Similar experiments are performed on PIE database. We use images captured by camera c27 with frontal illumination for testing. Some results are shown in Figure 4. From left to right, first two columns are normalized input images and recovered facial shapes with frontal light. The rest four columns are rendered faces and their corresponding actual images viewed from camera c02 and c05.

The experimental results on both synthesized and real world data shows that facial shape can be recovered by projecting the intensity image to the shape space via the coupled hidden subspaces that maximize the correlations between intensity images and 3D facial shapes. The statistical dependency between 2D appearance and 3D shapes is well represented by the coupled space learning framework.

Figure 3. Example results of facial shape recovery on YaleB database

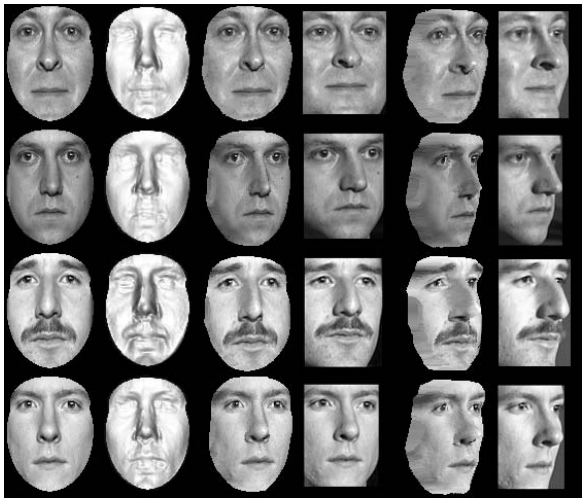

Figure 4. Example results of facial shape recovery on PIE database

4. Conclusion and Discussions

In this paper we explored the relationship between 2D appearance and 3D shapes by using a coupled 2D/3D space learning framework. 3D facial shape can be inferred from intensity image using the coupled model that maximizes the inter-space correlations. The experimental result shows that this framework can well represent the statistical dependency between intensity images and facial shapes.

One of the merits of this approach is that it needs no albedo information. For a single intensity image, the albedo and the shape are both unknown, which is a difficult obstacle for the shape recovery methods based on physical illumination models. The main drawback of the proposed approach is its sensitivity to illumination variations. It requires that the illumination condition of the training set should be similar to the test images. Recent study has proved that illumination variations of face images can be approximated by a linear combination of several images under different illumination conditions [2, 12]. The problem of illumination variations can be overcome by extending the training set with some lighting changes. So it is reasonable to extend this method to illumination variations in the future work.

Acknowledgement

This paper is partially supported by National Natural Science Foundation of China under contract No.60332010, No.60772071, and No.60673091; Hi-Tech Research and Development Program of China under contract No.2006AA01Z122 and No.2007AA01Z163; 100 Talents Program of CAS; and ISVISION Technology Co. Ltd.

References

[1] R. Zhang, P. Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999. 1

[2] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001. 1, 5, 6

[3] I. Kemelmacher and R. Basri, "Molding face shapes by example," *European Conference on Computer Vision, LNCS 3951*, vol. I, pp. 277–288, 2006. 1

[4] T. Hassner and R. Basri, "Example based 3d reconstruction from single 2d images," *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, p. 15, 2006. 1

[5] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8, no. 6, pp. 1321–1340, 1996. 1

[6] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3d structure recovery of faces," *European Conference on Computer Vision*, pp. 99–113, 2004. 1

[7] W. A. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1914–1930, 2006. 1

[8] W. A. Smith and E. R. Hancock, "Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 71–91, 2008. 1

[9] L. Zhang, S. Wang, and D. Samaras, "Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model," vol. 2, pp. 209–216, 2005. 1

[10] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003. 1

[11] W. Zhao and R. Chellappa, "Symmetric shape from shading using self-ratio image," *International Journal of Computer Vision*, vol. 45, no. 1, pp. 55–75, 2001. 1

[12] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003. 1, 6

[13] M. Castelan and E. R. Hancock, "A simple coupled statistical model for 3d face shape recovery," pp. 231–234, 2006. 2

[14] M. Castelan, W. A. Smith, and E. R. Hancock, "A coupled statistical model for face shape recovery from brightness images," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1139–1151, 2007. 2

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," vol. II, pp. 484–498, 1998. 2

[16] D. Lin and X. Tang, "Coupled space learning for image style transformation," pp. 1699–1706, 2005. 2

[17] "USF HumanID 3D Face Database,University of South Florida, Tampa, FL, USA." 4

[18] a. J. H. Kuang-Chih Lee and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005. 5

[19] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1615–1618, December 2003. 5