Face Inpainting with Local Linear Representations

Zhenyao Mo, J.P. Lewis, Ulrich Neumann
Computer Graphics and Immersive Technology Lab
University of Southern California

Abstract

A number of investigators have had success using domain specific prior knowledge to produce improved superresolution images of faces ("hallucinating faces"). These efforts address the scenario where a face image is obtained from a low-resolution camera. A related but less studied problem occurs when the missing information is the result of occlusion rather than low camera resolution, as in the case when a person is wearing sunglasses. Recently Hwang and Lee [14] introduced the first algorithm for solving this reconstruction "inpainting" problem. In the current work we report results of a psychological study that provides independent evidence regarding the validity of the face reconstruction task, and we demonstrate an improved reconstruction approach using a positive, local linear representation. The positive, local mixture operates on real-world images without manual intervention in many cases, and provides demonstrably lower reconstruction error than is obtainable with a global representation.

1 Introduction

When a stranger removes his or her sunglasses, we may occasionally be surprised by their appearance – or their appearance may be unsurprising. The fact that we can have expectations on a part of the face that we have not seen suggests that there are correlations between the appearance of that region and other visible face regions, and that, as human observers, we have some knowledge of these correlations. In fact this is true: roughly

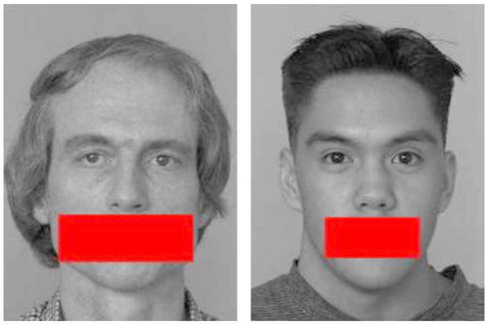

Figure 1: People have considerable knowledge of the conditional probability densities of parts of the face. For example, most people easily predict that the individual on the right has thicker lips than the person on the left.

BMVC 2004 doi:10.5244/C.18.37

eight out of ten people correctly estimate that the individual on the right in Fig. 1 has "thick" rather than "thin" lips, and nearly everyone believes that this individual has thicker lips than than the person on the left (see Section 2).

In this paper we give experimental evidence that humans have considerable knowledge of such conditional probabilities for several parts of the face. We then describe a simple algorithm that represents the underlying correlations between different face regions and allows a "best guess" reconstruction of an obscured region. A 'cascaded similarity' scheme allows known regions of a novel face to be represented as positive linear combination of similar training faces. The same positive local linear combination is adapted to provide successful reconstruction of unknown face regions, and it out-performs the reconstruction obtained from a standard least squares (positive/negative, global) linear representation in many cases. This algorithm complements the successful face superresolution or "hallucination" algorithms but addresses a different problem scenario in which large regions of a face are obscured.

After reviewing related work, the remainder of the paper is organized as follows: Section 2 discusses human performance on the face reconstruction task. Section 3 describes our reconstruction approach, and Section 4 evaluates our results.

1.1 Related Work

A number of authors have exploited domain knowledge in producing superresolution images of faces [1, 15, 18, 11, 12]. The original "face hallucination" work [1] gave a MAP formulation of face superresolution in which a per-pixel prior on the unknown image gradient is obtained from a nearest-neighbor search of image feature vectors in the high-resolution training images. In [11, 12], the eigenvectors of a training set of high-resolution faces provide the prior for superresolution reconstruction of faces directly in the low-resolution face eigenspace – expansion to a pixel representation is bypassed, providing computational efficiency for some applications. In [18] good results were obtained on the same face hallucination problem using a non-parametric sampling approach similar to recent texture synthesis methods to provide the unknown high-frequency detail. While these approaches assume dense and regularly spaced pixel information, Hwang et. al. [15] take a slightly different approach by reconstructing faces from scattered sample points.

Although the face superresolution problem has been considered by a number of authors, a related face enhancement problem remains relatively unexplored. In this problem scenario an entire region of the face is unknown, as would be the case if the eyes are obscured by sunglasses, or if the mouth is obscured by the head of a closer person or some other object. This is termed the "inpainting" problem in the computer graphics literature. Recently (and concurrently with our own work) Hwang and Lee [14] introduced the first algorithm to address this problem. Following [16, 15], Hwang and Lee [14] represents a face image with two eigenbases, one for the shape correspondence between the particular face and a mean or reference face, and a second for the shape-normalized face texture. Given a face with obscured regions, the shape and texture of known regions are fit with a linear combination of their respective eigenvectors, and the same linear combination is then used to recreate the unknown region. The examples in [14] are primarily images of synthetic 3D face models with identical controlled lighting.

In this paper we also adopt the appearance/morphable model representation [8, 16], and use the fit over known regions to extrapolate the unknown regions. Our work is distinguished in its use of a convex local representation that provides improved performance in real-world cases (section 3), and in our inclusion of a psychological study (section 2) that provides supporting evidence regarding validity of the face reconstruction task.

2 Human Knowledge of Face Proportions

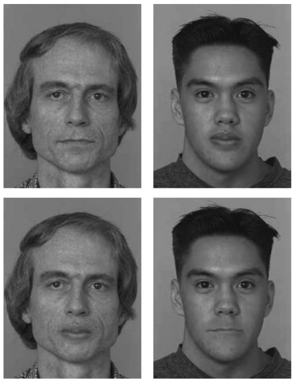

Figure 2: Using Photoshop to carefully interchange the mouths of two individuals produces unusual faces. (The luminance of the mouths has been adjusted to better blend into the target faces).

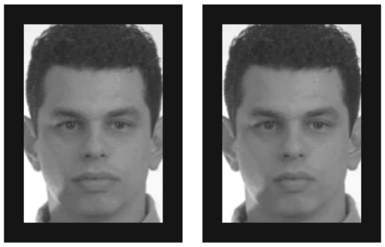

Figure 3: A web experiment presented pairs of face images in which the proportions of some part (in this case the mouth) are subtly manipulated. Subjects generally had a clear preference for one of the images as having a more "natural" appearance, and their preferences were consistent with a computed probability density of face feature proportions.

The validity of the machine face reconstruction task is an open question, since one might object that a demonstrated reconstruction is merely the result of a small or biased set of training images, and that many other reconstructions are equally plausible. Psychological experiments can provide independent evidence regarding this question. The results summarized in this section show that, on the contrary, the conditional densities of important face regions (eyes, nose, mouth) possess little uncertainty.

We first noted the possible extent of correlations between different face regions by creating face images with obscured regions (Fig. 1) and asking people to describe their guess of the appearance of the obscured region. For example, when we asked people to characterize their guess of the missing mouth on the individual on the right in Fig. 1 using the terms "thin" versus "thick", a majority of people correctly state that the individual probably has thick lips. When we showed both pictures in Fig. 1 and asked which of the two individuals has thicker lips, nearly everyone gave the correct answer without hesitation. Similarly, in Fig. 2 we show the result of interchanging the mouths of two individuals using the Photoshop image editing tool. The resulting faces are unusual, again suggesting that human observers have implicit knowledge of conditional face probabilities.

These observations were later formalized in a web study in which 30 people were asked to identify which of a set of computer manipulated face images appeared most natural. All the images were altered by a thin-plate spline warp (to produce equivalent image resampling effects in all cases), with the proportions of either the eye, nose, or mouth regions being altered by various amounts. Image pairs were randomly selected from five graduated alterations (with duplication prohibited), so both obvious and subtle alterations were tested. Subjects viewed up to 600 pairs of such images and were asked to pick the image of each pair that appeared most natural (Fig. 3). Subjects were instructed that they could stop at any point. Partial results are valid since the task is identical in each trial. We retained all results in which the subject completed at least 50 pairs. A total of 8340 pair votes were obtained, with each person completing an average of 278 pairs.

The subjects' ratings were compared with a kernel density computed from the eye, nose, and mouth widths of faces in the training database (Fig. 4). The Gaussian kernel width was set to 1/10th of the standard deviation of the data along each dimension (anisotropic). Subjects' rating were in agreement with the kernel density in 74.0% of the trials, meaning that if a subject rated face A as being more natural than face B, the kernel probability also assigned A a higher probability, $K(A) > K(B)$. After removing pairs with an insignificant difference in probability ($|K(A) - K(B)| < .001$), the percentage of agreeing pairs increased to 87.7%. (Our web-based study also may have provided lower quality data than would be obtained in a study with paid subjects under controlled conditions, so these percentages may be lower than their true values.)

In this experiment people are effectively rating the plausibility of the manipulated face region conditioned on the appearance of the rest of the face. The experiment presented a variety of subtly altered faces without obvious alteration artifacts (Fig. 3), yet subjects chose faces consistent with a kernel density computed from the data even when presented with other choices. This result confirms that face regions are correlated (Fig. 4, left) rather than independent (Fig. 4, right), and that the reconstruction results shown here (and in [14]) are not necessarily an artifact of a small or biased training set. Again note that all images were manipulated, so subjects could not identify the "real" images simply by evidence of alteration.

3 Reconstruction Algorithm

3.1 Face Model

Following [16, 8] a face image is encoded as a linear combination of shape+texture vectors. The face *shape* consists of the x and y coordinates of 94 feature points (Fig. 5). Feature points are used to define a warp that provides a dense shape correspondence between different face images. We use thin-plate splines to define the x and y warps since the thin-plate spline minimizes an approximate curvature and thus provides smooth interpolation though the points. The face *texture* is the face image warped into the shape of the mean face. A total of 400 faces from the FERET database [20] were used as training data. The feature points (shape) for a novel face can sometimes be located automatically with a gradient descent fit ([8, 16]; we use a procedure similar to [8]), though this procedure can fail due to unusual hair, lighting, or other factors.

3.2 Local Linear Face Representation

Linear appearance models in computer vision have been pioneered independently by a number of authors including [2, 9, 13, 16, 8]. While linear representations are simple, in general they can be expected to be accurate only over small regions. In the case of faces, the probability density of some face proportions is not convex and is highly structured (Fig. 4), so there is some danger that a linear combination of very different faces may

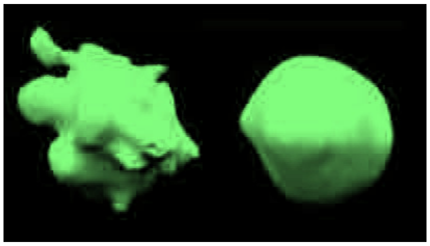

Figure 4: (Left) kernel density visualization of the distribution of the eye width (X axis) versus nose width (Y axis) versus mouth width (Z axis) for 400 faces. (Right) shape of the density resulting from the product of the marginal densities of the eye, nose, and mouth widths taken separately. The true distribution (left) suggests that a linear mixture of faces may produce an unusual or unrealistic result if the faces in the mixture are separated by low probability "holes" in the density; this fact is not evident when looking at the distribution of individual features (right). Human subjects have considerable implicit knowledge of the true distribution (Section 2).

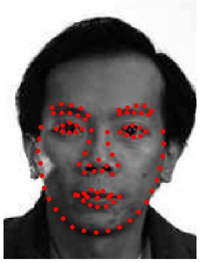

Figure 5: Training faces are labeled with 94 feature points. Several points between the lips are coincident in this closed-mouth image.

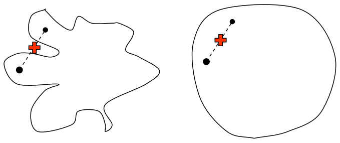

Figure 6: Extrapolation from distant data is safe if the data has a simple density (right) but not with the complex, non-monotonic density on the left.

result in a face image that falls in one of the "holes" in the distribution (Fig. 6), i.e., a face with unusual and perhaps unrealistic proportions. Mixing a large number of faces can also result in the "blurring" that is frequently visible in eigenface reconstructions.

The limitations of a global linear model have been addressed by fitting a collection of local linear models (e.g. [5, 3]). We adopt this local linear approach, but rather than using pre-defined clustering, we instead dynamically select the cluster best suited to represent a given face. This is done by a novel nearest-neighbor scheme (described next); it might also be considered as a form of outlier rejection in fitting the local linear model. (Note that "local" refers to the distance in the face representation space, not in the face image.)

3.3 Cascaded Similarity Threshold

The selection of nearest neighbors brings up several issues: how is "similarity" defined, and how is the threshold chosen? While the choice of a threshold can be a worthwhile design parameter, it seems somewhat arbitrary to include one data item but reject another that may be arbitrarily close but lying on the other side of the threshold. Although the ultimate choice of basis for a similarity measure is not always explicitly discussed in the literature, some applications require a similarity measure that is reasonably correlated with the human rankings of face similarity. For example, if witnesses describe a sought individual as looking "like a younger brother of Paul McCartney", a face recognition

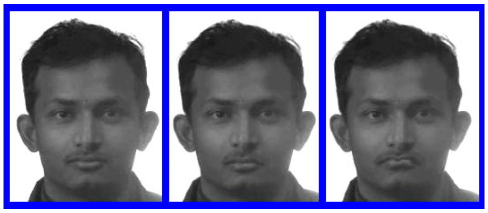

Figure 7: Positive-only versus unconstrained reconstruction with extensive training data: left, original image; middle, positive-only reconstruction of the mouth region; right, reconstruction with unconstrained (positive and negative) weights.

algorithm can contribute to the search only if its similarity measure correlates with human similarity judgments. For face data the common sum squared error (L_2) norm may not relate well to human similarity judgments.

Based on these considerations, we use a cascaded *sequence* of psychologically relevant weak similarity measures to select the nearest neighbors (this resembles slightly the cascaded use of weak classifiers in boosting [10]). Specifically, we use three similarity criteria,

- skin texture variance
- eye, nose, and mouth locations
- eye, nose, and mouth sizes

These particular measures were selected based on psychological evidence that humans use these features in discriminating faces [6, 19]. Given a face, a set of weakly similar faces is first selected using the skin texture variance measure, using an L_2 distance. The similarity threshold is set conservatively, to include all similar faces as well as many dissimilar ones. This set is then reduced by applying the location and size measures in sequence. While additional measures could be employed as well (and considerable experimentation would be required to find a set that best corresponds to human ranking of face similarity), using the intersection of several independent similarity measures is likely to be an improvement over relying on only one such measure. The similarity thresholds are design parameters that must be chosen by the experimenter.

3.4 Positive Mixture

When a linear representation is used, the desirability of restricting the weights to be positive has been noted by several researchers [7, 4]; [7] provides a discussion. Intuitively, a convex combination of prototypes (corresponding to weights $w_k > 0$, $\sum w_k = 1$) should result in a well behaved representation – without the positive constraint an exemplar might sometimes be represented by a combination of prototypes with large positive and negative weights in a canceling combination, leading to amplification of noise and poor reconstruction (Fig. 8). Although an unconstrained linear representation often works, as shown in [14], in fact it can result in poor reconstruction in some cases (Fig. 7). The number of such possibly undesirable representations increases as the number of training images increases (i.e., the as linear representation tends towards being underdetermined or overcomplete),

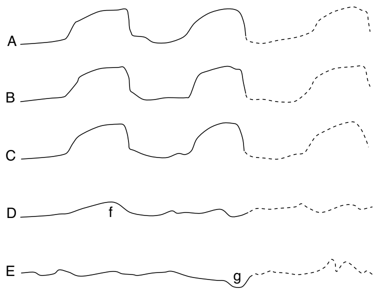

Figure 8: A signal A can be approximated as a convex sum of a library of similar signals such as B, C. The same weighted sum may allow successful reconstruction or extrapolation (dashed lines). If unconstrained weights are allowed, the signal can also be represented as a weighted combination of many dissimilar signals such as D, E, by magnifying possibly incidental features such as f, g with large positive and negative weights to match features of A. The large weights also magnify noise, and reconstruction may be impaired.

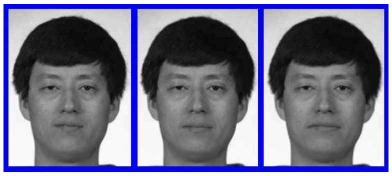

Figure 9: Restricting a linear representation to only near neighbors often improves extrapolation. Left, original image; middle, reconstructed mouth extrapolated from similar faces only; right, mouth extrapolated from a linear mix of all faces. Also see Table I.

since the number of ways of representing an exemplar to within a specific accuracy increases.

We use the non-negative least squares (NNLS) algorithm [17] to solve for the positive mixture representation of a given face. Target faces are first translated and scaled to approximately match the mean position and size of the selected local neighborhood prototype images, using a least squares fit. This makes it more likely that the target face will lie within the convex hull of the prototypes.

3.5 Reconstruction

Given a target face with some obscured regions, using the procedure described in section 3.3.3 we select a subset of the training faces that (evaluated in the unobscured regions only) are relatively similar to the target face $\hat{\mathbf{f}}$. Positive representation weights are chosen,

$$\arg\min_w ||\hat{\mathbf{f}} - \sum w_k \mathbf{f}_k||_{\text{visible}} \quad w_k \geq 0$$

again evaluating the match only in the unobscured regions.

Face shape and texture are correlated [8], so we fit the shape and texture in one step rather than independently. In order to equalize the relative contributions of the shape and texture components, the actual quantity minimized is

$$||\Delta S|| + \frac{\sigma_s^2}{\sigma_t^2} ||\Delta T||$$

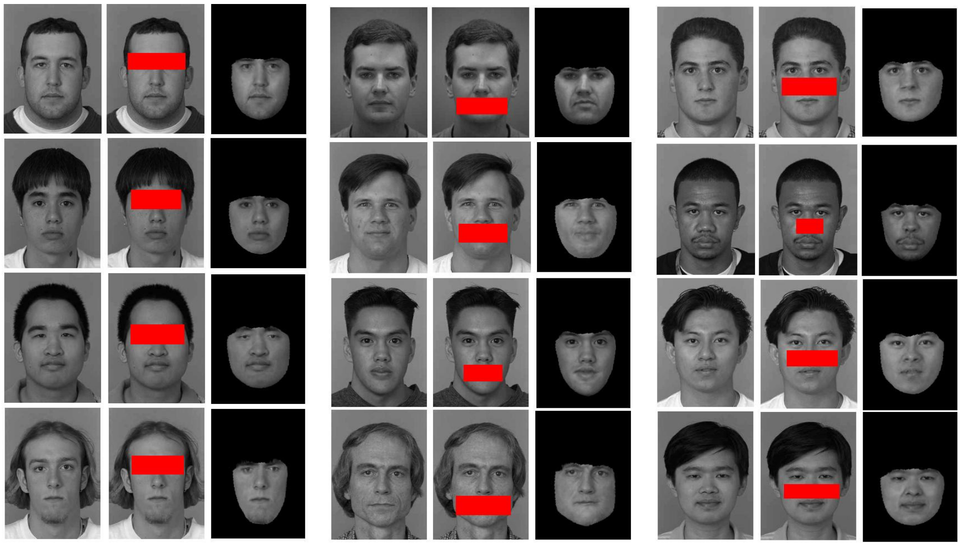

Figure 10: a),b),c) From left to right: original faces, target faces with obscured region, and reconstruction results, for eye, mouth and nose regions.

with ΔS and ΔT begin the shape and texture errors in the visible regions and σ_s^2, σ_t^2, being the variances of the shape and texture training data, respectively. The texture is then warped into the neutral face shape, allowing it also to be approximated as a linear combination of the same prototypes. Once known parts of the face are fit, the same linear mixture $\sum w_k \mathbf{f}_k$ is used to extrapolate the unknown region(s).

4 Results and Conclusion

Some results are shown in Fig. 10. The left column of each group shows the original face images, the middle column shows the target images with eyes (Fig. 10a), mouth (Fig. 10b), or nose (Fig. 10c) regions obscured, and the right column shows our reconstruction. These target images were not members of the training set. We see that the procedure usually produces reasonable results, with some exceptions (in this figure the reconstructed nose of the individual in the second row is too narrow, perhaps reflecting an inadequate number of people with similar features in the database).

Local linear reconstruction also produces a demonstrable improvement in reconstruction compared with "global" reconstruction based on the entire training set: Table I shows the reconstruction error for 600 face image outside the training set with obscured eye, nose, and mouth regions (200 cases each).

Table 1:

	mean error global fit	mean error local fit	local better percent cases
mouth	1.19%	0.47%	74.0%
nose	1.16%	0.43%	70.5%
eyes	1.36%	0.88%	67.5%

In this table, the mean error columns report the average shape error between the extrapolated image and the true image (not part of the training set), calculated as

$$\frac{||S_{\text{new}} - S_{\text{old}}||}{||S_{\text{old}}||}$$

and the third column shows the percentage of cases where the local representation results in a smaller reconstruction error than the global representation. Local linear reconstruction also outperforms simple nearest neighbor reconstruction. For comparison, the mean nearest neighbor shape errors for the mouth, nose, and eyes are 6.67%, 5.79%, and 10.62% respectively.

4.1 Conclusion

In this paper we summarize results of a psychological experiment showing that the conditional correlations between different face regions in a face database are not merely the result of the limited size of the database. We then demonstrate a model-based algorithm that uses the correlations implicit in the face database to provide improved reconstruction of obscured regions in novel faces. The reconstruction uses a positive-only mixture of training faces, and the mixture is restricted to faces relatively similar to the target face with similarity being defined by the intersection of several cascaded weak measures.

The choice of a positive-only (and typically convex) combination of basis functions eliminates the possibility of large and canceling weights that could result in noise amplification or poor extrapolation in our reconstruction task. Excluding dissimilar faces from the representation restricts it to a small region of face space where a simple linear combination can be successful and helps prevent the interpolated result from falling in a low probability "hole" in the face space. The resulting reconstruction algorithm provides demonstrably lower reconstruction error on a set of 600 face images that were not part of the training set.

References

[1] S. Baker and T. Kanade. Hallucinating faces. In *Proc. Fourth ICAFGR*, page 83, 2000.

[2] Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV (1)*, pages 329–342, 1996.

[3] Bowden. *Learning non-linear models of shape and motion*. PhD thesis, Brunel University, 2000.

[4] Christoph Bregler, Lorie Loeb, Erika Chuang, and Hrishi Deshpande. Turning to the masters: motion capturing cartoons. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pages 399–407. ACM Press, 2002.

[5] Christoph Bregler and Stephen M. Omohundro. Surface learning with applications to lipreading. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 43–50. Morgan Kaufmann Publishers, Inc., 1994.

[6] R. Brunelli and O. Mich. SpotIt! an interactive identikit system. *Graphical models and image processing: GMIP*, 58(5):399–404, 1996.

[7] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. Technical report, Department of Statistics, Stanford University, February 1996.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. *Active Appearance Models*, volume 1407 of *Lecture Notes in Computer Science*. 1998.

[9] M. Covell and C. Bregler. Eigen-points. In *Proc. IEEE International Conference on Image Processing*, pages vol 3 p 471–474, 1996. Lausanne, Switzerland, Sept 16-19 1996.

[10] Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. Artif. Intel.*, 14(5):771–780.

[11] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes III, and R. M. Mersereau. Eigenface-based super-resolution for face recognition. In *IEEE Int. Conf. on Image Processing (ICIP) vol. 2*, pages 845–848, 2002.

[12] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes III, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606, 2003.

[13] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Computer Vision and Pattern Recognition*, pages 403–410, 1996.

[14] B-W. Hwang and S-W. Lee. Reconstruction of partially damaged face images based on a morphable face model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):365–372, 2003.

[15] B-W. Hwang, S-W. Lee, V. Blanz, and T. Vetter. Face reconstruction from a small number of feature points. In *International Conference on Pattern Recognition (ICPR)*, pages 842–845, 2000.

[16] Michael J. Jones and Tomaso Poggio. Model-based matching by linear combinations of prototypes. Technical Report AIM-1583, MIT Artificial Intelligence Lab, 1996.

[17] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, N.J., 1974.

[18] C. Liu, H-Y. Shum, and C-S. Zhang. A two-step approach to hallucinating faces: Global parametric model and local nonparametric model. In *Computer Vision and Pattern Recognition (CVPR'01) - Volume 1*, pages 192–198, 2001.

[19] D. McNeill. *The Face: A Natural History*. Little Brown, London, 1998.

[20] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J.*, 16(5):295–306, 1998.