Face Recognition At-a-Distance Based on Sparse-Stereo Reconstruction

Ham Rara, Shireen Elhabian, Asem Ali
University of Louisville
Louisville, KY
{hmrara01,syelha01,amali003}@louisville.edu

Mike Miller, Thomas Starr, Aly Farag
University of Louisville
Louisville, KY
{tlstar01,aafara01}@louisville.edu

Abstract

We describe a framework for face recognition at a distance based on sparse-stereo reconstruction. We develop a 3D acquisition system that consists of two CCD stereo cameras mounted on pan-tilt units with adjustable baseline. We first detect the facial region and extract its landmark points, which are used to initialize an AAM mesh fitting algorithm. The fitted mesh vertices provide point correspondences between the left and right images of a stereo pair; stereo-based reconstruction is then used to infer the 3D information of the mesh vertices. We perform experiments regarding the use of different features extracted from these vertices for face recognition. The cumulative rank curves (CMC), which are generated using the proposed framework, confirms the feasibility of the proposed work for long distance recognition of human faces with respect to the state-of-the-art.

1. Introduction

Face recognition is a challenging task that has been an attractive research area in the past three decades [1]. Initially, most efforts were directed towards 2D facial recognition which utilizes the projection of the 3D human face onto the 2D image plane acquired by digital cameras. The recognition problem is then formulated as: given a still image, it is required to identify or verify one or more persons in the scene using a stored database of face images. The main theme of the solutions provided by different researchers involves detecting one or more faces from the given image, followed by facial feature extraction which can be used for recognition.

Recently, there has been interest in face recognition at-a-distance. Yao, et al. [2] created a face video database, acquired from long distances, high magnifications, and both indoor and outdoor under uncontrolled surveillance conditions. Medioni, et al. [3] presented an approach to identify non-cooperative individuals at a distance by inferring 3D shape from a sequence of images.

In this paper, we propose to use active appearance models (AAM) to provide sparse point correspondence. Given a stereo pair (left and right images), we first detect facial region using the *logRGB* space, landmark points such as eye centers, mouth center and nose tip are then extracted to initialize the AAM mesh fitting algorithm. We exploit the correspondence between the fitted mesh vertices of the left and right images to apply stereo reconstruction.

To achieve our goal and due to the lack of facial stereo databases, we built our own passive stereo acquisition setup to acquire a stereo database. Our setup consists of a stereo pair of high resolution cameras (and telephoto lenses) with adjustable baseline. The setup is designed such that user can remotely pan, tilt, zoom and focus the cameras to converge the center of the cameras field of views on the subject's nose tip. We used our acquisition system to capture stereo pairs of *30* subjects at various distances (3-, 15- and 33-meter ranges).

The paper is organized as follows: Section 2 describes stereo-based reconstruction, Section 3 discusses AAM mesh initialization and fitting, Section 4 talks about the features used for recognition, and later sections deal with the experimental results, discussions and conclusions.

2. Stereo-based Reconstruction

Stereo camera systems are usually calibrated (e.g., with a calibration pattern) before 3D reconstruction is performed. For this work, we use the known orientation between the two cameras to estimate 3D points. The theoretical explanation of the cameras relationship is illustrated in Fig. 1.

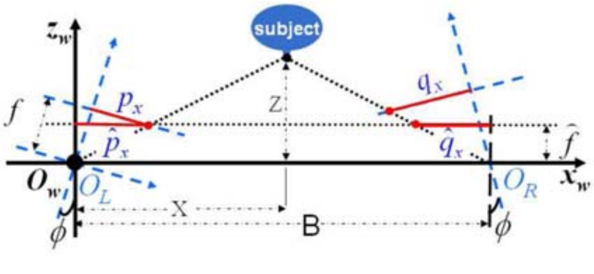

Figure 1: General stereo pair setup, where O_L, O_R are the left and right camera coordinate systems and O_w is the world coordinate system.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

Since the system parameters (i.e. baseline *B* (meter), focal length *f* (mm), pan angle *Φ* (degree), and scale factor of the cameras k_α (pixel/mm)) are known, the scene point *(X,Y,Z)* can be reconstructed from its projections p and q using the geometry shown in Fig. 1, assuming $p_y = q_y$, as follows. First we calculate the values

$$\hat{p}_x = \sqrt{p_x^2 + f^2 k_\alpha^2}\sin(\phi + \tan^{-1}(\frac{p_x}{f k_\alpha})),$$
$$\hat{f} = \sqrt{p_x^2 + f^2 k_\alpha^2}\cos(\phi + \tan^{-1}(\frac{p_x}{f k_\alpha})),$$
$$\hat{q}_x = \hat{f}\tan(\phi + \tan^{-1}(\frac{-q_x}{f k_\alpha})).$$

Then, we compute the reconstructed scene point *(X, Y ,Z)* as follows (starting from *Z* to *X* and *Y*) [4]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \dfrac{Z}{\tan(\frac{\pi}{2} - \phi - \tan^{-1}(\frac{p_x}{f k_\alpha}))} \\ \dfrac{p_y\sqrt{X^2 + Z^2}\cos(\frac{\pi}{2} - \phi - \tan^{-1}(\frac{Z}{X}))}{f k_\alpha} \\ \dfrac{\hat{f}B}{\hat{p}_x - \hat{q}_x} \end{bmatrix}$$

3. Active Appearance Model (AAM) Fitting

Matthews and Baker [5] considered the independence of shape and appearance (independent AAMs) in their AAM version. The shape s can be expressed as the sum of a base shape s_0 and a linear combination of n shape vectors s_i, $s = s_0 + \sum_i p_i s_i$, where p_i are the shape parameters. Similarly, the appearance $A(\boldsymbol{x})$ can be expressed as the sum of the base appearance $A_0(\boldsymbol{x})$ and a linear combination of basis images $A_i(\boldsymbol{x})$, $A(\boldsymbol{x}) = A_0(\boldsymbol{x}) + \sum_i \lambda_i A_i(\boldsymbol{x})$, where the pixels $\boldsymbol{x}$ lie on the base mesh s_0.

Fitting the AAM to an input image involves minimizing the error image between the input image warped to the base mesh and the appearance $A(\boldsymbol{x}) = A_0(\boldsymbol{x}) + \sum_i \lambda_i A_i(\boldsymbol{x})$, that is

$$\sum_{x \epsilon s_0} \left[A_0(\boldsymbol{x}) + \sum_i \lambda_i A_i(\boldsymbol{x}) - I\big(W(\boldsymbol{x}; p)\big)\right]^2$$

For this work, the error image is minimized using the *project-out* version of the inverse compositional image alignment (ICIA) algorithm [5].

AAM Initialization: To facilitate a successful fitting process, the AAM mesh is initialized according to detected face landmarks (eyes, mouth centers, and nose tips). Fig. 2 shows the results of the detection of face landmarks. After detecting these face features, the AAM base mesh is warped to these points.

The *logRGB* space is used to detect candidate face regions, and each candidate is scored whether a face or not by trying to detect a pair of eyes. The detection of eyes involves the concept of eigeneyes [6]. The mouth center detection [7] involves transforming the image by a linear combination of the red, green, and blue chrominance components of the RGB color space. This transformed image emphasizes the lips pixels. Estimating the nose tip involves finding the centers of mass of two nostril candidates. The mean of these centers is considered to be the nose tip. The nostril candidates are determined by taking into account its low red response in the RGB space, and their distance to the mouth and eye centers.

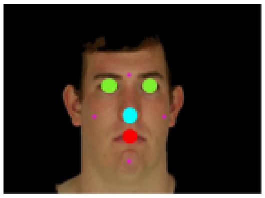

Figure 2: Detection of face features.

4. Features for Face Recognition

For face recognition, we use four approaches for using the 3D face vertices derived from AAM and stereo to identify probe images against the gallery, namely: (a) feature vectors derived from Principal Component Analysis (PCA) of 3D vertices, (b) goodness-of-fit criterion (*Procrustes*) after rigidly registering the 3D vertices of a probe with that of a gallery subject, (c) feature vectors from PCA of *x-y* plane projections of the 3D vertices, and (d) the same procedure as (b) but using the *x-y* plane projections of the 3D vertices of both probe and gallery, after frontal pose normalization. The use of the *x-y* plane projections will be explained further in the experimental results (Sec. 5).

Principal Component Analysis (PCA) [8]: To apply PCA for feature classification, the primary step is to solve for the matrix *P* of principal components from a training database, using a number of matrix operations. The feature vectors *Y* can then be determined as follows: $Y = P^T X$, where X is a centered input data. The similarity measure used for recognition is the L_2 norm.

Goodness-of-fit criterion (Procrustes): The Procrustes distance between two shapes is a least-squares type of metric that requires one-to-one correspondence between shapes. After some preprocessing steps involving the computation of centroids, rescaling each shape to have equal size and aligning with respect to translation and rotation, the squared Procrustes distance between two shapes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is the sum of squared point distances:

$$P_d^2 = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2$$

Rigid alignment involves solving a transformation $T(.)$, such that the term $\|T(\boldsymbol{x}_1) - \boldsymbol{x}_1\|^2$ is minimum [9].

Partial-to-Full Information: Primary experiments of this paper use the total number (*68*) of AAM vertices for recognition. Additional experiments are done to investigate the effect of using only a fraction of the total number of vertices for recognition. Fig. 3 illustrates the increasing number of points on the face. The first case contains only the positions of the two eyes, nose tip and mouth center. Case 5 considers the outline of the face parts. The last case is the total number of points resulting from an AAM fitting procedure.

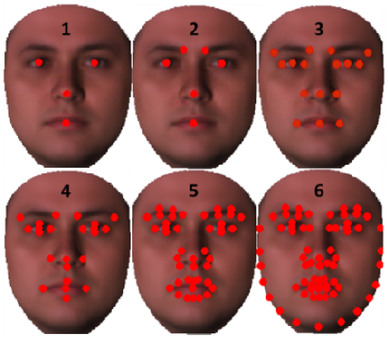

Figure 3: Visualization for test cases involving increasing the number of points used for recognition.

5. Experimental Results

To test our framework, we used our 3D acquisition system to build a human face database for *30* different subjects at different ranges in controlled environments. The subjects are asked to stand in front of our system setup for capturing. The system is adjusted to converge to the center of the cameras' field of, views on the subject's nose tip. Our database consists of a gallery at 3 meters and three different probe sets at the 3-, 15-, and 33-meter ranges. The training of the AAM model involves images from the gallery.

Table 1 shows the system parameters at different ranges. Fig. 4 illustrates the captured images (left image of the stereo pair) from different ranges. The approximate face region in pixels for the three ranges are (1500 × 1900) for the 3-meter range, (700 × 100) for the 15-meter range, and (320 × 470) for the 33-meter range.

Range (m)	Baseline B (m)	Zoom f(mm)	Focus $\approx$	Pan ϕ (degree)	Tilt (degree)
3	0.6	150	Range	$\approx 5.6°$	$\approx 0°$
15	1.76	400	Range	$\approx 3.3°$	$\approx 0°$
33	1.76	400	Range	$\approx 1.5°$	$\approx 0°$

Table I: Stereo-based acquisition system parameters

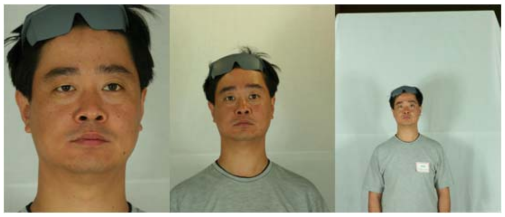

Figure 4: Left to right: Illustration of captured images at 3-, 15-, and 33-meter distances.

Figs. 5 and 6 show AAM fitting results on sample probe images. The vertices of the final AAM mesh on both left and right images can be considered as a set of corresponding pair of points, which can be used for stereo reconstruction.

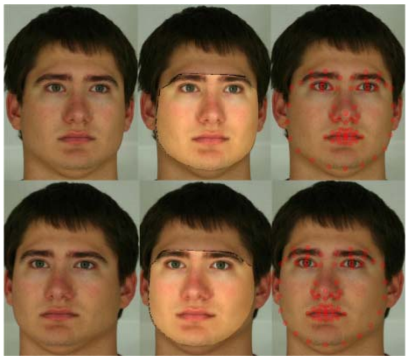

Figure 5: AAM fitting results for a probe at 3 meters. First column is the input image. Second column is the final AAM result with superimposed best-fit texture. Third column shows final AAM vertices.

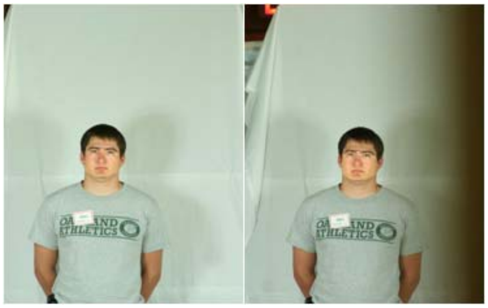

Figure 6: AAM fitting results for a probe stereo pair at 33 meters. The best-fit texture is superimposed to the input images, similar to the second column in Fig. 5.

Fig. 7 shows stereo reconstruction results of three subjects, visualized with the *x-y*, *x-z*, and *y-z* projections, after rigid alignment to one of the subjects. Notice that in the *x-y* projections, the similarity (difference) of 2D shapes coming from the same (different) subject is enhanced. This is the main reason behind the use of *x-y* projections as features in Sec. 4.

Figures 8-10 show the cumulative rank curves (CMC) curves for the four types of feature vectors mentioned in the previous section, using 3-, 15-, and 33-meter probes.

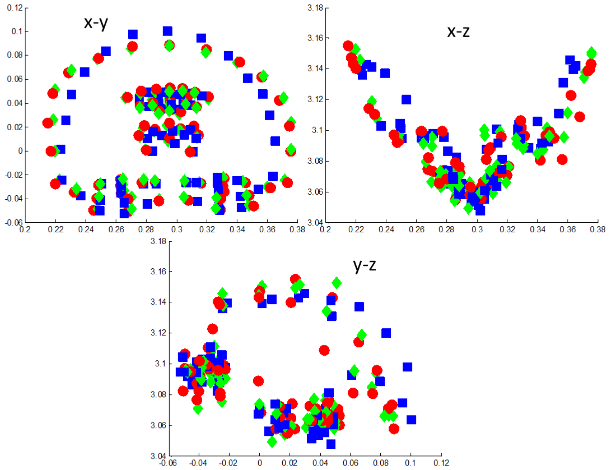

Figure 7: Reconstruction results. The 3D points are visualized as projections in the *x-y*, *x-z*, and *y-z* planes. Red (circle) and green (diamond) markers belong to the same subject, while the blue (square) maker is that of a different subject.

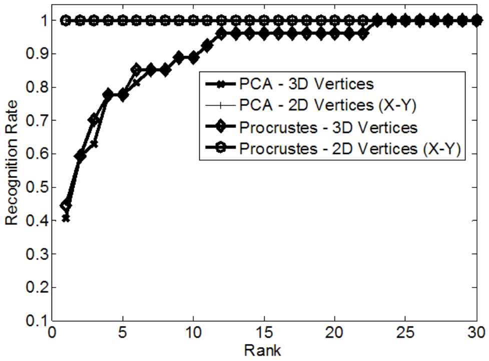

Figure 8: Cumulative match characteristic (CMC) curve of the 3-meter probe set

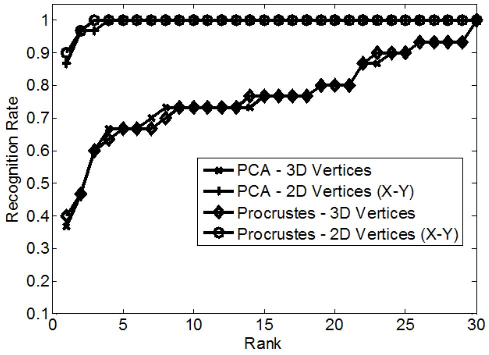

Figure 9: Cumulative match characteristic (CMC) curve of the 15-meter probe set

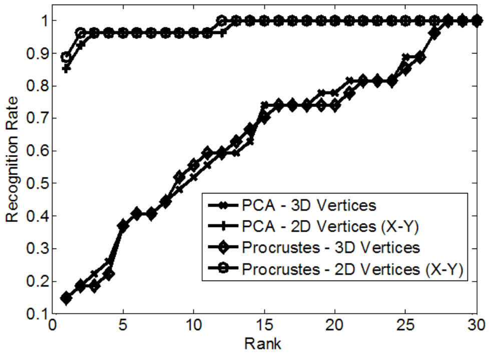

Figure 10: Cumulative match characteristic (CMC) curve of the 33-meter probe set.

6. Discussion of Results

From Figs. 8-10, we can draw three conclusions: (a) both 2D Procrustes and 2D PCA outperform both 3D Procrustes and 3D PCA, (b) goodness-of-fit criterion (Procrustes) slightly outperforms PCA in both 2D and 3D, and (c) degradation of recognition at increased distances.

The conclusion in (a) can be explained with help of Fig. 11. The diagram shows the top view of a simple stereo system. O_l and O_r are centers of projection, and p_l, p_r, q_l, q_r are points on the left and right images.

Assume that the *y*-coordinates of the four image points are equal (see Sec. 2). p_l and p_r will reconstruct P, q_l and q_r will reconstruct Q, and so on. Notice that a small change in the correspondence affects the *xyz* reconstructions hugely, i.e., their Euclidean distance between each other is huge. But when they are projected to the *x-y* plane, their 2D Euclidean distances with each other are considerably lesser. This scenario is possibly happening with the correspondence of the AAM vertices between the left and right image. What essentially happened here is using stereo as pose correction of the vertices in Fig. 5. The left and right AAM vertices, by themselves, cannot perform recognition due to presence of pose but after performing stereo reconstruction and orthographic (*x-y*) projection, the real 2D shape of the face is extracted.

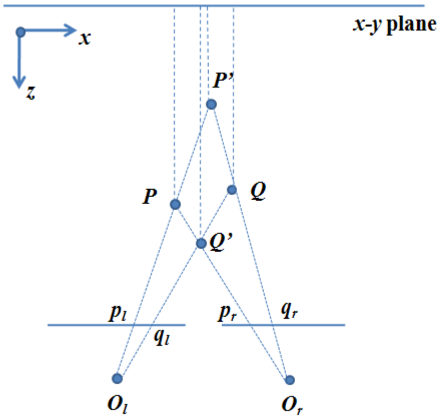

Figure 11: Simple stereo illustration. 3D reconstruction is sensitive to the correspondence problem but projection to the x-y plane minimizes the error.

The conclusion in (b) is related to the primary purpose of PCA, which is optimal reconstruction error, in the mean-square-error (MSE) sense. It is possible that projecting the original shape vector to a low-dimensional space removes the classification potential of the vectors. There is no dimensional change with rigid alignment using Procrustes; similar shapes are expected to have less Procrustes distance after rigid alignment and geometric information of faces (e.g., distance ratios between face parts) are maintained.

Results are expected to degrade with distance since the captured images are at less ideal conditions (although recognition using 2D *x-y* projections remain stable). The work in Medioni [3] deals with identification at a distance using recovered dense 3D shape from a sequence of captured images. The results at the 15- and 33-meter range (Figs. 9 and 10) are comparable (and slightly better) than their 9-meter results; however, their experimental setting may be less controlled than ours.

Fig. 12-14 shows the CMC curves for the test cases in Fig. 3. The points here are the *x-y* projections of the original 3D vertices and the recognition approach involves the *goodness-of-fit* criterion (since they perform best in Figs. 8-10).There are two conclusions that can be drawn here: (a) with the 3-meter probe set, *Case 3* is enough to perform perfect recognition and (b) with the 15- and 33-meter probe, however, Cases 3-5 outperform (or perform similarly) the full number of AAM points at *rank 1*. For the 15- and 33-meter probe sets, the points inside the mouth contour, as well as the face outline, contribute to the error. From a correspondence problem perspective, this is expected since the face outline points (as well as the inside mouth points) are found in homogenous regions of the face and are prone to correspondence artifacts that affect depth estimation (see Fig. 11).

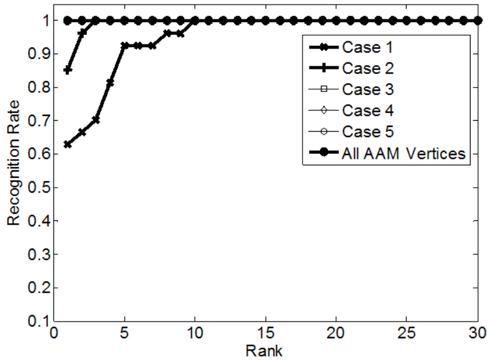

Figure 12: Cumulative match characteristic (CMC) curve of the test cases in Fig. 3, using the 3-meter probe set.

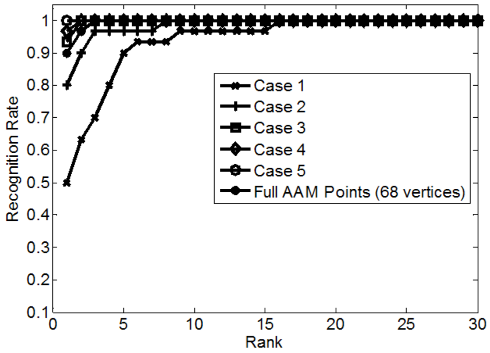

Figure 13: Cumulative match characteristic (CMC) curve of the test cases in Fig. 3, using the 15-meter probe set.

The conclusions here are valid, at least, for this database. Further work will involve testing these concepts on larger databases (e.g., increasing the number of subjects in our database).

7. Conclusions and Future Work

We have studied the use sparsely-reconstructed points from the AAM vertices of a stereo pair, in the context of long-distance recognition. Using our database of images taken at the 3-, 15-, and 33-meter distances, we have illustrated the potential of using these few vertices, as opposed to the whole set of points of the human face. Results show comparable performance compared to the state-of-the-art.

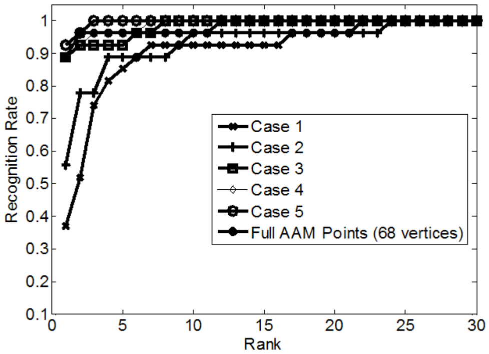

Figure 14: Cumulative match characteristic (CMC) curve of the test cases in Fig. 3, using the 33-meter probe set.

The next steps in this project are to increase the database size and capture images at further distances (with the help of state-of-the art equipment). The authors expect more challenging recognition scenarios and will incorporate additional information and techniques (e.g., texture and super-resolution) to improve the recognition algorithms. Further studies involving sensitivity analysis of the recognition performance with respect to various errors introduced to the system (e.g., inaccurate location of feature points in both left and right images), will be performed in the future.

8. References

[1] W. Zhao, R. Chellapa, and A. Rosenfeld,“Face recognition: a literature survey,” ACM Computing Surveys, 35 (2003) 399–458

[2] Yao, et al., “Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement,” CVIU 111(2), 2008

[3] G. Medioni, et al., “Non-Cooperative Persons Identification at a Distance with 3D Face Modeling,” BTAS, 2007

[4] E. Trucco and A. Verri, “Introductory Techniques for 3-D Computer Vision,” Prentice-Hall, March 1998

[5] I. Matthews and S. Baker, “Active Appearance Models Revisited,” ICCV, 2004

[6] W. Huang, Q. Sun, C. Lam, and J. Wu, "A robust approach to face and eyes detection from images with cluttered background," ICPR 1998

[7] E. Gomez, et al., “Biometric identification system by lip shape,” Security Technology 2002

[8] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection,” IEEE Trans. PAMI, 19(7), 1997

[9] T.F. Cootes and C.J. Taylor, “Statistical Models of Appearance for Computer Vision,” Technical Report, University of Manchester, UK, March 2004