

FACE RECOGNITION USING VIDEO CLIPS AND MUG SHOTS

**Gaile G. Gordon
Marquess E. Lewis
TASC**

**55 Walkers Brook Drive, Reading, MA 01867
email: GGGordon@tasc.com, MELewis@tasc.com
617-942-2000**

ABSTRACT

This paper presents two automated identification systems based on 3D face recognition developed at TASC. They have the capability to process mug shots, consisting of frontal and profile views, as well as video. Both systems are discussed in terms of their advantage with respect to viewing angle changes. The mug shot system, FaceMatcher, is based on low level pattern recognition of geometrically normalized subimages. The addition of profile imagery is shown to dramatically lower the recognition error rates of a frontal view based system. The video based system, FaceMatcher3D, explicitly extracts the 3D pose of the head in each video frame. Low level features, such as corners, are identified and tracked in the video stream. The feature tracks are processed by a shape from motion algorithm which produces estimates of 3D geometry and pose. The geometry and pose estimates are considered together with facial structure constraints, temporal constraints, and initial pose estimates to refine knowledge of the specific face structure and its pose. This information can be used to extract frontal and profile views from the sequence and then to geometrically normalize them for comparison with the database. The capability to perform face recognition from video presents important advantages over other biometric identification methods, such as finger prints or retinal scans, because identification is now passive. Passive identification can be used not only in traditional access control, but also in surveillance applications.

1 Introduction

Variation in viewing angle is a key issue for face recognition applications of any kind. This paper presents two systems which treat view variation for different data scenarios. The first system addresses applications using still images taken with some knowledge of view, e.g. conditions common in frontal view matching systems. The second system addresses surveillance applications, in which no particular view can be assumed. In this case, view invariance is of even greater importance.

In single view matching systems misalignment with respect to view is a common source of error. For instance, for frontal view matchers, the tilt of the chin or the rotation of the head right or left might not be exactly the same in both views. Most controlled environment matching algorithms have a tolerance for slight misalignments in view, however, similarity scores will always be a function of view alignment. Even for perfect matches in identity, poor matches in view will lower recognition scores. The larger the database becomes, the more likely that classification errors will result.

The first system discussed, FaceMatcher, uses independently taken frontal and profile views (the exact rela-

tionship between the views is not known) . This system copes with view error by providing relief information from the profile view as additional evidence about the identity of the person. This is an example of coping with view variation in an application in which it is not possible to extract the actual pose of the head in the images used. It is shown that this information sharply reduces the recognition errors over the frontal only system.

In surveillance applications, before any recognition algorithm can be used, the head must be isolated, and the direction of view must be explicitly computed. Computing the view parameters is equivalent to computing the relationship between the camera and a head centered coordinate system. We refer to this information as the *3D pose parameters* of the head. Once the pose parameters are known, many options for the treatment of view become available, including selection of aspects which best match those previously stored, and specific geometric normalization to compensate for view misalignments.

The pose of any three dimensional rigid object has six degrees of freedom including three translational parameters and three rotational parameters. Without specialized range sensors, the passive computation of these pose parameters for the head requires *multiple 2D views* [1]. In the second system, FaceMatcher3D, pose is computed from a video sequence of the head in motion. Examples are shown of automatic pose extraction from real video sequences.

2 Recognition from Mug Shot Pairs

For humans, facial appearance is a very common and reliable method of judging identity. It is for this reason that most existing biographical databases contain facial photographs of some kind. In comparison to other information likely to be stored in biographical databases (e.g. height, weight, hair and eye color), facial images make the most sense for use in verification of identity because of their relative invariance over time, and ease of passive data acquisition.

The largest factor determining the appearance of the face

is its three dimensional shape. Since very little information about the relief of the face is discernible reliably from the frontal view, the profile view offers important and virtually independent information about the shape of the face. The system, FaceMatcher, described in this section demonstrates that the addition of the profile information reduces the recognition error rates by approximately 40% [2]. The addition of this independent information will effectively decrease the impact of any systematic error in the frontal view comparison, however, in these applications, view alignment is a key source of error.

2.1 Algorithm Summary

The recognition process begins by extracting a model representation from the input images. The model contains the location of two key feature points in each image and a set of five normalized image chips. The feature points are used to establish the location of the face within the image and to perform geometric normalization. They include the *pupil centers* from the frontal view, and the *tip of the nose* and the *tip of the chin* from the profile view. The normalization process scales and rotates the images such that the key feature points have a constant distance and orientation for all images of the same view. The image chips are extracted from the geometrically normalized frontal and profile images and include *left and right eye, nose, mouth, and central profile*. Feature points are extracted automatically using morphological operators, pattern matching, and other low level image analysis algorithms[2]. An example of an extracted face model is shown in Figure 1.

The comparison of a target model with a reference model is performed based on pattern matching of the image chips. A number of different algorithms could be used for this purpose. The system currently uses a variation of normalized cross correlation. The images are adjusted to have zero mean which yields a normalized cross correlation match measure given by [3]:

$$\text{Zero mean: } \frac{(\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n fg) - uv}{\sigma\tau},$$

where f and g refer to the images being compared, mn is the area of the image region compared, u , and v are

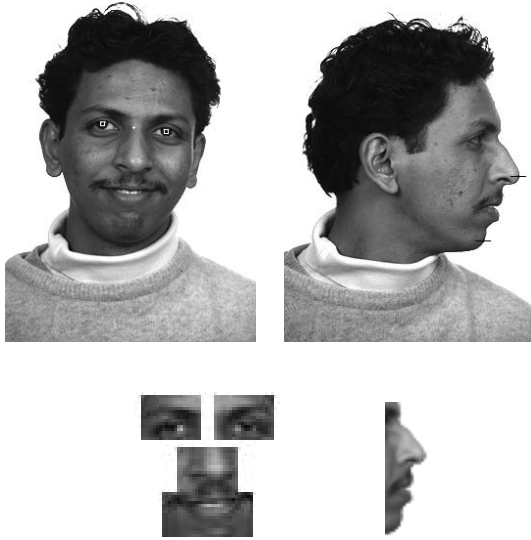


Figure 1 : Original frontal and profile images with marked features (top), and normalized subimages which are stored in model (bottom).

the mean values of the two images, and σ and τ are their standard deviations. This formulation increases the tolerance of the match to simple overall changes in intensity. An additional term, \mathcal{V} , representing the variance of noise, is added to the denominator:

$$\text{Noise adjusted: } \frac{(\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n fg) - uv}{\sigma\tau - \mathcal{V}}.$$

This decreases the tendency for high match scores in areas of very low image variance.

The overall similarity score is a linear combination of the five correlation scores. The output of the process is a ranked list of subjects from the reference database, ordered by similarity to the query subject.

2.2 Recognition Results

The FERET program (ARL/ARPA) database was used for algorithm evaluation. The data set for a given subject includes two frontal and two profile (left and right) images, the minimum data required for recognition testing. For each subject, one of the two frontal view images was paired with the left profile image to build a reference model. The second frontal view image was paired with

the right profile image to build a query model. There is a wide variety of image scale and lighting and a good representation of age, ethnic background, and gender in the database. Many of the subjects have glasses, several have moustaches and/or beards. Although there are photographs included from each of three different time periods, there is no overlap of subjects between the three sets. A total of 194 subjects were used for testing.

It is expected in many face recognition applications that several of the top candidates, not just the single top candidate, are presented for further human consideration. In the presentation of results we use the concept of rank threshold, R , which is the percentage of the best database entries to be presented as the result of the comparison. If the “ideal” match is contained within the first R candidates in the ranking list, we consider this a successful recognition result. We report system performance by giving error rate as a function of rank threshold. This provides error rates for all choices of R .

Our testing had two goals. The first was to quantitatively show the value of the addition of the profile data on recognition error rates. The second goal was isolate the effect of incorrect feature detection from the rest of the comparison process.

To evaluate the value of the profile data, a direct comparison was made between the system performance when using only frontal images and the system performance when using both frontal and profile images for the same subjects. The results (Figure 2) show that the addition of profile information substantially reduces error rates. Using $R = 5\%$ as a reference point, recognition error rates are 45% lower in the profile/frontal case, yielding a correct recognition rate of 85%.

Our second goal was to separate the errors in detection of the normalization features (pupil centers, nose and chin tip) from the performance of the underlying comparison method. The feature extraction module is the most complex part of the system. Its capabilities determine essentially what types of images the system can ingest. Any automated feature extraction system will contribute to the general error rate of the system as a whole.

The effects of feature localization on the system perfor-

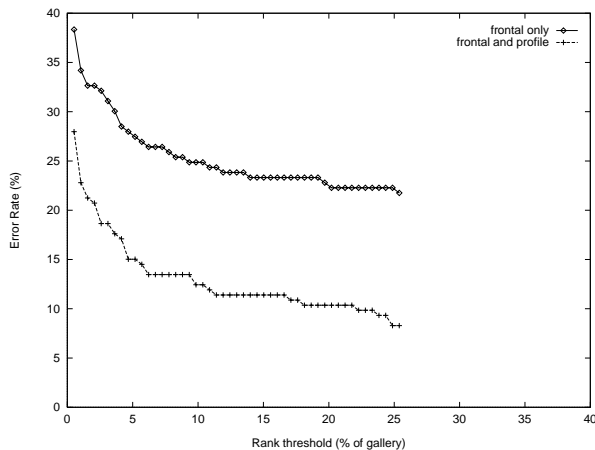


Figure 2 : Comparison with frontal only system.

mance were examined through the use of an additional model database constructed from the same images, however, the four normalization features were located interactively in each data set using the computer mouse. Figure 3 shows error rates vs R threshold for both the automated feature location database and manual feature location database. Using the same $R = 5\%$ reference point, the error rates using interactive feature location were below 2%, corresponding to a correct recognition rate of *above 98%*. In general, errors in automatic feature detection increased system recognition error rates by approx. 10% across most values of R .

The residual error in the system (2–5% at $R = 5\%$) were due to view variation and large expression changes.

This analysis shows the critical role of feature point detection in the recognition system. In the video based system which is described in the following section, a great deal more information is available for the feature location modules. Other improvements to the Face-Matcher system are also being addressed which should improve the normalization feature extraction including probabilistic evidence fusion and improved initial head scale estimation.

It is important to note, however, that applications which can afford a small amount of interaction (selection of 4 points per data set) can eliminate this source of error and greatly simplify the system at the same time.

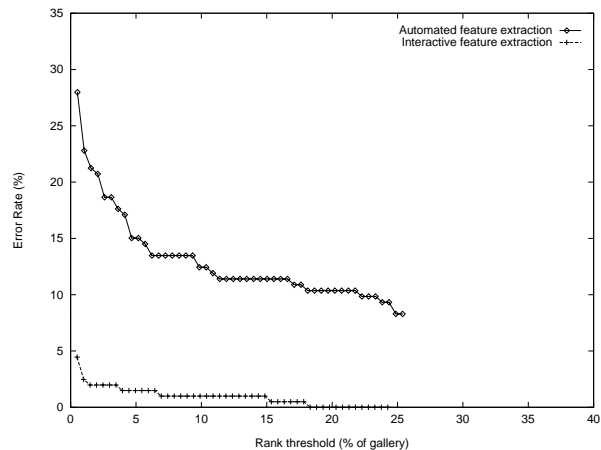


Figure 3 : Performance with and without feature extraction error for normalization features.

3 Pose extraction from Video Sequences

Images taken in unstructured environments will include many viewing angles. In order to effectively compare faces on the basis of image pattern data, the viewing angle (or equivalently, head pose) must be explicitly taken into account. What data is required to compute head pose? Pose extraction from a single 2D image is only possible if a well defined 3D model of the object exists [4, 5, 6], which it does not in most face recognition applications. Pose computation from three dimensional data, however, is a simple matter of geometry. Passive computation of three dimensional data from two dimensional data of a rigid object is possible through the use of stereo or structure from motion techniques [7, 5, 8]. The FaceMatcher3D system uses video sequences and structure from motion techniques because only a single uncalibrated sensor is required. This method is also attractive because video cameras are already available in many locations for the purposes of surveillance (e.g. ATM's, many passport control gates).

3.1 Algorithm Summary

The pose extraction algorithm includes three key steps:

- Feature selection

- Feature tracking
- Computation of Structure and Pose

The third step, a structure from motion algorithm, is addressed first because its requirements drive the design of the other steps.

The structure from motion algorithm used in the Face-Matcher3D system is based on the factorization method presented by Tomasi and Kanade [8]. The algorithm takes as input a set of feature tracks. Each track consists of an array of 2D image locations, one for each frame, showing the path which a 3D point on the surface of a rigid object takes from frame to frame. From this information a solution is provided for both the 3D location of the points (structure) and the relationship between the object and the camera (pose).

The job of *automatically* selecting and tracking the points for input to the factorization algorithm is complex. The features tracked must correspond to surface points on a rigid object. Image features corresponding to occluding boundaries, shadows, and motion independent from the head, e.g. torso or other objects in the scene, violate this assumption and will act as a source of error.

The basic feature selection method [9] identifies local image patterns containing edges in more than one direction. These are identified by examining the eigen values of the matrix

$$G = \sum_W \begin{bmatrix} x'^2 & x'y' \\ x'y' & y'^2 \end{bmatrix},$$

where W is a small local image region (15x15 pixels). Two strong eigen values of G are a good indication of strong gradients in multiple directions and hence corner-like features. A threshold on the minimum eigen value of G is used as a specific selection criteria.

This low level feature selection algorithm has been augmented with knowledge of the application domain. An initial pose selection algorithm provides an estimate of likely frontal views in the image stream. This estimate is used to locate the head initially, and limit the feature selection to the central head region.

Once the features have been selected they are tracked by modeling the motion of the feature window as a simple translation. The displacement of the feature window is computed by minimizing the error residue between frames based on this model [10]. Although the translation model is more simplistic than the actual image change expected over time (e.g. it doesn't include rotation or scale changes) it is quite effective over the small interframe distances typical in video rate sequences.

Two further steps are taken to minimize the inclusion of points not representative of rigid surface motion. The first is an affine criteria used to monitor longer term changes in the feature window [9]. Including affine transformations allows for a more accurate and realistic tracking model. To examine the likelihood that a feature window has been tracked correctly, we can compare it's original values with those of its final tracked position. The affine transform which best models the relationship between the two windows is computed. This transform is applied to the original window, and the result is compared with the window in the actual tracked position. Low similarity is indicative of occlusion. A threshold can be used to remove these features from further consideration.

In addition to the affine similarity criteria, nonrigid features are identified by comparing the actual observed tracks with those estimated by the computed pose and structure [1]. Features with large position residuals are removed and the pose and structure are recomputed. This method is effective at removing features not representing true surface points, e.g. image windows containing edges from more than one object.

3.2 Results

The pose extraction algorithm has been tested using the FERET Program video database, which was collected by TASC. Figure 4 shows an example of features which were automatically selected and successfully tracked through a 22 frame sequence.

These feature tracks were processed by the structure from motion algorithm to produce a pose estimates at each frame, as well as the 3D locations of the points.

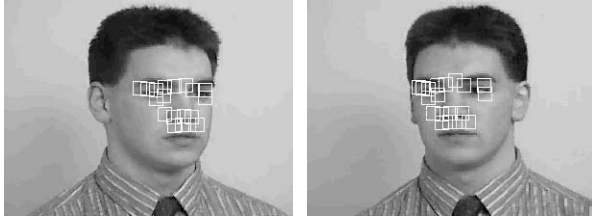


Figure 4 : First and last frames of a 22 frame sequence showing tracked feature windows.

As is represented in Figure 4, the movement represented in the sequence is a rotation to the left about the Y axis (vertical) with very little motion about the X axis (through the ears) or Z axis. The pose extracted matches this expectation, indicating final rotation angles of 0.7 degrees about X, 37.9 degrees about Y, and 0.5 degrees about Z.

A model was generated from the extracted 3D points using 2D Delaunay triangulation to provide the topology of the points. Image data from a single frames was texture mapped onto the polygonal model for visual evaluation. This model is shown in Figure 5 at two different viewing angles. These views demonstrate the accuracy of the model. In particular, the eye sockets are set back with respect to the nose, and the overall face shape is gently curved.

Because of the difficulty measuring the true pose of a human head, additional experiments have been performed for the purposes of testing pose accuracy. Video sequences were taken of inanimate objects on rotating platforms. These experiments indicate ± 2 degree accuracy in pose computations.

4 Summary

FaceMatcher and FaceMatcher3D provide automated face recognition technology valuable in a wide range of input scenarios. The FaceMatcher system is for use with still images in applications where an approximate knowledge of view angle is available. Testing of this system on the FERET Program 200 person database demonstrates a clear quantitative advantage in consid-

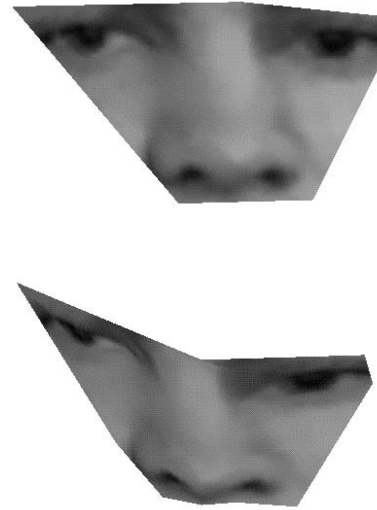


Figure 5 : Two different views of the 3D model generated from the above tracks.

ering profile data in addition to the classic frontal view. Testing also shows that applications which can tolerate a small amount of interaction during the matching process (clicking on two standard points in each image) can simultaneously eliminate a major source of error in recognition and greatly simplify the installed system.

FaceMatcher3D has the capability of 3D pose extraction from video sequences of a moving subject. This information enables both extraction of key poses from video sequences, and matching from video to video. The extraction of key pose frames is valuable both in creating still databases from surveillance videos, or for matching live data against a database containing only still images. The 3D model extracted by FaceMatcher3D also provides valuable information to adjust the extracted imagery to form a better match in view than is available directly from the video frame data.

Acknowledgements

This work was supported by the FERET program of the Army Research Lab and ARPA under contract DAAL01-93-C-0118.

References

- [1] G. G. Gordon, "Ferret: Using multiple images for high-performance 3d face recognition," Technical Report TR 7745-1, TASC, Inc., Reading, MA, September 1995.
- [2] G. G. Gordon, "Face recognition from frontal and profile views," in *Proceedings of the International Workshop on Face and Gesture Recognition*, (Zurich, Switzerland), pp. 47–52, June 1995.
- [3] A. Rosenfeld and A. C. Kak, *Digital Picture Processing, Second Edition*. Vol. 2, Orlando, Florida: Academic Press, Inc., 1982.
- [4] W. E. L. Grimson, *Object Recognition by Computer*. Cambridge, MA: MIT Press, 1990.
- [5] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT Press, McGraw Hill, 1986.
- [6] M. A. Fischler and R. C. Bolles, "Random sample consensus: Paradigm for model fitting with application to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1978.
- [7] O. Faugeras, *Three-Dimensional Computer Vision*. Cambridge, MA: MIT Press, 1993.
- [8] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.
- [9] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994.
- [10] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.

Author Biographies

Gaile Gordon

Gaile Gordon is a member of the technical staff in the Visual Exploitation Technologies Department at TASC. Her area of expertise centers on image analysis, segmentation, and description for a variety of applications and data modalities. Current work has focused on content based image search, face recognition in intensity and depth images, machine inspection, and analysis of 3D surface data. Dr. Gordon received B.S. and M.S. degrees in Computer Science and Engineering from MIT in 1985 and 1986 respectively. She received an M.S. and a Ph.D. in Computer Science, in 1989 and 1991 respectively, from the Division of Applied Sciences at Harvard University. Her Ph.D thesis presented several approaches to the problem of face recognition based on curvature descriptors calculated from range data. Dr. Gordon has presented papers at national and international conferences in the field of computer vision. She was an invited participant to the first NSF Summer Institute in Japan, and the NATO Advanced Study Institute on the topic of Active Vision.

Marquess Lewis

Marquess Lewis has ten years experience in the exploitation of imagery for geometric information and developing geometric models of sensor systems. Mr. Lewis' academic background is in classical photogrammetry with B.S and M.S. degrees from the SUNY College of Environmental Science and Forestry. As a Member of Technical Staff at TASC for 6 years, Mr. Lewis has been involved in the development of algorithms and software for generating geometric models of objects imaged with hand-held cameras, processing SPOT stereo satellite imagery, and performing rectification of airborne scanner imagery. Prior to joining TASC, Mr. Lewis was employed by DBA Systems and developed software and algorithms for geopositioning workstations. Recent areas of interest include video exploitation using structure from motion and modeling highly dynamic sensor systems.