

Increased Resolution 3D Face Modeling and Recognition From Multiple Low Resolution Structure From Motion Models

Chris Boehnen and Patrick J. Flynn
Department of Computer Science and Engineering
University of Notre Dame

Abstract—We present an approach to combine multiple noisy low density 3D face models obtained from uncalibrated video into a higher resolution 3D model. The approach first generates ten 3D face models (containing a few hundred vertices each) of each subject using 136 frames of video data in which the subject face moves in a range of approximately 15 degrees from frontal. By aligning, resampling, and merging these models, we produce a new improved 3D face model containing over 50,000 points. An ICP face matcher employing the entire face achieved a 75% rank one recognition rate, which falls within the documented range of performance similar to whole-face 3D matcher results [2] that use more advanced laser scanners for data acquisition. The simplicity of our hardware requirements reduces cost, complexity, and may enable the use of “other people’s video” for 3D face modeling and recognition.

I. INTRODUCTION

UTILIZING three-dimensional scanning devices such as laser scanners to acquire data is common for 3D biometrics. Active scanners can produce highly accurate 3D face models comprising 100,000 points or more [1]. However, these scanners are expensive and complex. The use of face video for three-dimensional face modeling is therefore of interest; robust performing methods may also be able to make use of existing video data of interesting individuals.

Structure from motion (SFM) is a method for producing 3D models from a calibrated or uncalibrated video stream utilizing equipment that is inexpensive and widely available. A straightforward application of SFM to face modeling is relatively easy to imagine, but implementation and robust operation presents significant challenges. The low quality (fidelity) of SFM-constructed 3D face models is a key problem, and arises from the small number of trackable face points (low density) and reconstruction quality based upon those tracks.

We propose a method for combining multiple noisy low density SFM face models of the same subject into a higher definition 3D model useful for biometric recognition or other applications. In this paper, we utilize a small amount of video data (136 frames) with a relatively small amount of movement (approximately 15 degrees from frontal) with no “scripted” movement direction or path needed. This may

allow our approach to be applicable to uncalibrated video from a variety of sources. After creating multiple individual models of the same subject using SFM we combine them into a single 3D model with a high vertex count. This is accomplished via reference face correspondence, resampling, and merging. To test the modeling technique, we conducted 3D face recognition experiments. A simple whole-face ICP matcher achieved 75% rank one recognition performance using the proposed approach. By comparison, the use of a single SFM model instead of the combined high-resolution model yielded a rank one recognition rate of 38%. Previous work [2] using a similar ICP-based whole-face matcher with data acquired by a Minolta laser scanner data yielded rank-one recognition rates between 63% and 91%.

In this paper, we briefly survey prior work on 3D face modeling from video and 3D face recognition. Then we present our approach for generating models using SFM, merging them through correspondence, resampling, and averaging, and performing recognition. The paper concludes with discussion and ideas for additional work.

II. PREVIOUS WORK

The production of a 3D model from a 2D video stream for biometric recognition has been explored in previous work. However, the quality of resulting models from approaches such as structure from motion is unsatisfactory, as has been noted in previous work [7]. Here we present a brief overview of 3D face production from video followed by an overview of 3D face recognition approaches.

Production of 3D faces from video can be performed using morphable models, stereography, or structure from motion (SFM). Many variations of these approaches are face specific and require the identification and localization of the face and facial features in the 2D image.

Vetter and Blanz’ morphable model approach [3] can be used to produce 3D face models from single 2D images. As such, it is not a traditional video approach because it does not utilize temporally connected frames, but is capable of operating on video. The method trains a PCA space utilizing 3D face images, and determines a PCA reconstruction that closely matches the input 2D image. Medioni et al. [4] utilized synthetic stereo to model faces in a 3048 x 4560 video stream. By tracking the pose and location of the face, they initialize a synthetic stereo rig based upon the different poses between two frames. They create multiple point clouds from

Manuscript received May 15, 2008.

Chris Boehnen is with the University of Notre Dame (phone: 574-631-8320; fax: 574-631-9260; e-mail: chris@boehnen.com).

Patrick Flynn is with the University of Notre Dame (e-mail: flynn@nd.edu).

different stereo pairs and then integrate them into a single model. Their initial paper did not contain recognition test results. Cheng and Lai [5] tracked 2D face location and features in video. A generic 3D face model was fit to the 2D face at each frame and modified the generic face based upon silhouette changes. Resulting 3D model quality was not analyzed and no biometric recognition was performed. Park and Jain [6] extended Cheng and Lai's approach by tracking 72 facial feature points using an Active Appearance Model. They modified a generic face model to fit the location of these 72 points over time. Thus, they improved upon the earlier work by including shape information from the interior (not on the silhouette).

Roy-Chowdhury et al. [7] merged a 3D face SFM model with a generic face model. This was necessary due to the low quality of the initial SFM face model, as is common. After obtaining an initial 3D face estimate they merged the result with a generic 3D face using an energy minimization approach. No biometric experiments were performed.

The majority of 3D face recognition techniques can be classified as either ICP-based [2][8], 3D PCA-based [9], or depth map based [10]. Chang et al. [2] used a whole-face ICP region matching approach. They align a probe to each member of the gallery utilizing ICP, and use the RMS matching error as a biometric match score (lower is better). Depending upon expression and other factors, they found 63% to 91% rank one recognition rate.

Russ et al. [9] utilize a 3D Principal Component Analysis (PCA) based approach for face recognition. The approach determines correspondence utilizing a reference face aligned via ICP to determine a unique vector input into PCA. They use the coefficients from PCA to determine identity as in 2D PCA face recognition.

Kakadiaris et. al [10] converted the 3D model into a depth map image for wavelet analysis. This approach performs well and is one of the few mainstream approaches that does not utilize ICP as the basis for each match score computation, but does for the depth map production. While this approach does utilize ICP, it is distinctive from traditional ICP based 3D face biometrics.

III. FACE MODELING USING STRUCTURE FROM MOTION

Here we discuss our data collection process, a SFM-based 3D face model production approach, and a model combination technique that produces a high-resolution output from several low-resolution input face models.

A. Data Capture

We captured face video using a Sony HDR-HC3 HD consumer video camera capable of capturing video at 1080×1440 interlaced resolution at 30 frames per second. Subjects were located approximately 3 feet in front of the camera and were instructed to rotate their faces in small circles approx 20° from a frontal pose to produce the motion necessary for a SFM approach. We did not attempt to script the movement path or speed: this yielded a set of video clips with large

variation in head movement between subjects. We captured two 2.5-second clips of 24 subjects in a single day.

B. Coarse Model Construction using SFM

The proposed technique for generating coarse 3D face models from individual video sequences employs a well known point tracking technique to supply a Kalman filter-based shape estimator.

1) Face region identification

Before we can select individual points to track, we have to determine the face region from which we will pick those points. We utilize background subtraction to determine the foreground of the video, which consists of the face. To simplify the background subtraction step, we used a green cloth as a backdrop. The green cloth is not necessary to our technique, but does improve segmentation. After performing background subtraction, a face region is identified as follows. The largest region with a large background subtraction result is eroded with a morphological operator to remove noise. The resulting region is modeled as an ellipse centered at the region centroid with major and minor axes chosen to contain the region in its entirety. These steps are illustrated in Figure 1.

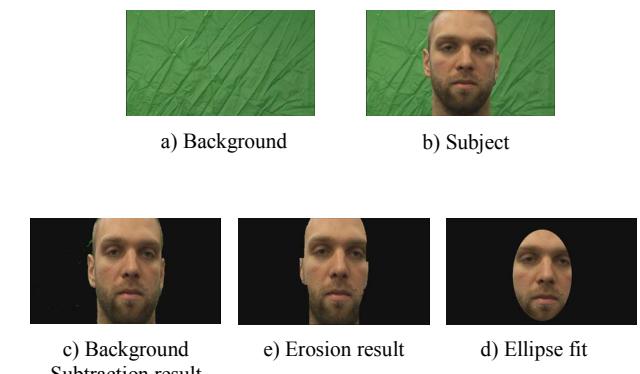


Figure 1: Face Region Identification Overview

2) Extraction of point tracks

Point tracks are supplied to the extended Kalman filter that performs SFM. Desirable tracks endure through the entire clip and are fixed to a 3D face position. Thus, a texture based point tracker, such as the interest operator defined by Shi and Tomasi [11] is appropriate; we used the implementation available in OpenCV. The Shi-Tomasi approach performs an eigenanalysis of a local neighborhood to identify points with significant texture content. The result of this detection step appears in Figure 1.d.

Optical flow is the process of determining a 2D correspondence for pixel locations between two 2D images [12]. If P_t represents the points being tracked at time t on image I_t , then optical flow is the process of determining the points P_{t+1} for image I_{t+1} . Estimation of the optical flow for a point generally involves a search in the new image for an intensity neighborhood that matches the neighborhood of the point being tracked in the original image. Design parameters for implementation of optical flow include the size of the

search neighborhood, the criterion for matching, and the search resolution (in some applications, subpixel localization is desirable and feasible). Lucas and Kanade [12] solved this minimization problem using a Taylor series. The approach is well documented and a complete explanation is outside the scope of this paper. We employ the OpenCV implementation of Bouguet's pyramidal modification of the Lucas-Kanade optical flow method [12].



Figure 2: Tracking Seeds Identified by Green Arrows

3) Factorization

For our purposes, the process of utilizing the motion tracks to estimate a rigid 3D model can be viewed as a factorization step. Although faces are not rigid, the short temporal extent of the video clips used and the use of an iterative factorizer that can accumulate errors in specific error variables is assumed to handle the nonrigidity problem adequately.

Our approach utilizes the factorization method proposed by Jebara et al. [13]. That work proposed an Extended Kalman Filter (EKF) framework for the iterative reconstruction of geometry and other quantities as variables of the filter, given the observations embodied in the motion tracks. The EKF state vector x has $7+n$ variables, where n is the number of tracked points:

$$x = (t_x, t_y, t_z, \beta, w_x, w_y, w_z, \alpha_1, \alpha_2, \dots, \alpha_n) \quad (1)$$

In the model, t_x , t_y , and t_z are translations, w_x , w_y , and w_z are components of a quaternion update to the rotation matrix, β is the inverse of focal length, and the quantities α_i are used to reconstruct the z -coordinate of points being tracked (the x and y coordinates can be reconstructed from initial image-plane coordinates and the calibrated perspective model).

The internal calculations require the inversion of a $(7+n) \times (7+n)$ matrix at each iteration. The authors of [13] claim this to be a trivial computational task computationally. This is true when n is small but not true when n becomes large. Jebara et al. do not report having tested their approach on more than 70 points, and their discussions do not include more than 16 to 30 points. They report reasonable factorization utilizing as few as 20 points and 25 frames. They tested their approach on faces and buildings utilizing 16 to 30 user-defined points over 70 to 100 frames. They

categorized their results as fast and stable, requiring low calibration. The key advantages claimed of this approach over alternatives are speed, lack of calibration required, ability for zoom changes, ease of use, and robustness of model.

C. SfM Model Production

We partitioned a video sequence of each subject into ten clips, each 36 frames long and offset from its predecessor by ten frames from the same 150 frames of video. This partition is shown in Figure 3. This allows us to use a small amount of data to create a large number of models. Processing each clip independently using the SfM technique produces ten 3D models of each subject.

The motion tracks we produce contain approximately 400 points each. The texture-based selection technique generally causes the points to be randomly and evenly distributed throughout the face. No specific feature coordinates (e.g., eyes, nose tip, point of chin) are used or required.

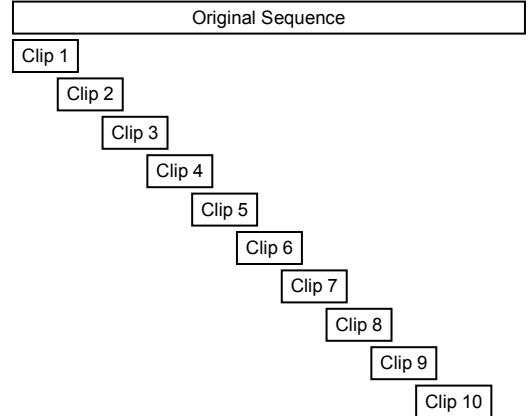


Figure 3: Overlapping Clip Data For Face Model Synthesis

D. Triangulation

Factorization yields a 3D point cloud. However, the goal is to be able to produce a 3D model in the form of a polygonal mesh. To do this, we triangulate the point cloud. Although 3D triangulation methods exist, these methods are more complex to implement, computationally expensive, and prone to errors. We implement a 2D Delaunay triangulation based upon the x and y coordinates and then, after determining the 2D triangulation, lofting the mesh by adding the z coordinate to each vertex. Figure 4 shows raw meshes resulting from triangulation of SfM-generated 3D data from two subjects.

E. Smoothing

Many 3D scanners routinely use smoothing routines to minimize local noise that can be present in 3D data. The coarse 3D models produced by factorization and triangulation do contain noticeable noise. We used a Laplacian smoother to remedy this noise. This smoother averages each node's position with the average positions of the neighboring nodes in the triangulation.

The tracking, factorization and postprocessing steps

described above are repeated for each of the ten clips from the original video sequence, yielding ten 3D face models from each 2.5 second video sequence. Examples of the outputs of these steps for two different subjects are shown in Figure 4c and d. These ten models are then passed to the high resolution approach from the next section to be merged into one.

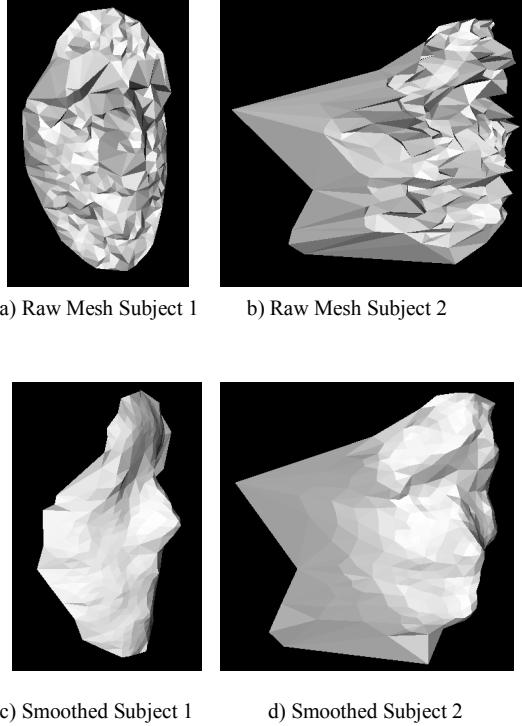


Figure 4: Factorization output with and without smoothing of two different subjects

IV. MERGING MULTIPLE MODELS AND RESAMPLING

Our SFM modeling approach produces ten 3D face models with approximately 400 vertices each from a single 2.5 second video sequence. No previous work exists on converting multiple noisy sparse density 3D models into one. This is a large part of the contribution of this paper.

Alignment methods, such as ICP, are designed for aligning two models, not three or more. Aligning ten models to one another simultaneously is a new area of research. The task of merging multiple sparse density models once aligned is a non-trivial task that has not previously been explored. Dense point cloud methods have previously been proposed with varying degrees of success. We propose a new method for merging that employs a vector representation of the mesh using a correspondence method developed previously [9]. An essential feature of the correspondence method is the use of a reference face surface obtained from training data, whose resolution determines the resolution of the model resulting from the correspondence and merging step.

A. 3D Face Model Vector Representation

To simplify the averaging process we begin by

representing each face model as a vector representing the face structure. The representation employs a reference face whose construction is described below. The following steps are applied to each of the ten face models resulting from SFM.

1. The input face centroid is aligned to the reference face centroid.
2. The input face is aligned with the reference face using ICP.
3. The reference face vertices yield corresponding input face vertices through a normal vector search.
4. The corresponding vertices are stored in a vertex vector whose indexing is tied to the reference face, thus maintaining alignment of all input faces to the reference face and to each other.

The process is depicted conceptually in Figure 5.

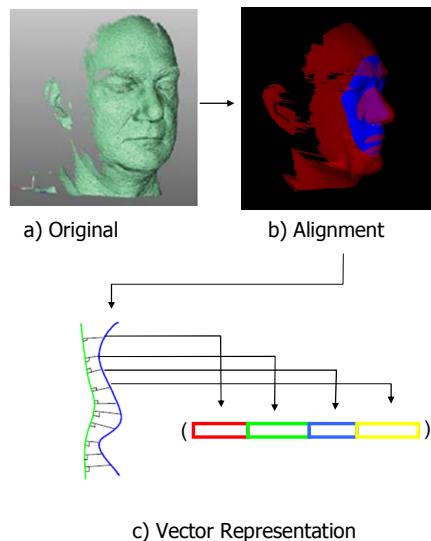


Figure 5: 3D Vector Production

1) Reference Face Selection

The reference face is a polygonal mesh generated manually from smoothed, registered, merged and resampled range images of 50 distinct subjects (the resampling process employs the same vector representation used below). The same reference face must be used to maintain consistent vector representations of all gallery and probe face meshes. The vertex density on the reference face determines the point density on all probe and gallery faces, since the vector representation consists of a closest point on the aligned probe or gallery mesh to each reference face vertex. Previous work [9] suggests that reference face selection has a negligible impact on performance and that a 3D model of a unique subject can be smoothed and used.

2) Alignment to the Reference face

We align the SFM-produced 3D face model to the reference face by translating the SFM face so that its centroid is coincident with the reference face centroid and applying ICP to rotate (and, if needed, further translate) the SFM model into alignment with the reference face mesh. We show a sample of the result of these alignments between a reference

face and our SFM data in Figure 6.

3) Correspondence and Vector Placement

After aligning the two meshes, we traverse the vertex list of the reference face in a specified order. The closest corresponding point in the input face mesh for the current point on the reference face is saved in an output vector. The order in which the output vector is populated is consistent; thus, the vector representation maintains registration among all models and the reference face as well. The correspondence search is a “normal search” technique and can synthesize a point not coincident with an input mesh vertex; this effectively provides a resolution enhancement ability. Previous work [9] concluded that this type of normal search is superior to a nearest neighbor search. Although the resampling process increases the vertex count, the quality of the output is the same as that of the triangulated SFM-produced model. The merging step described below provides the quality improvement.

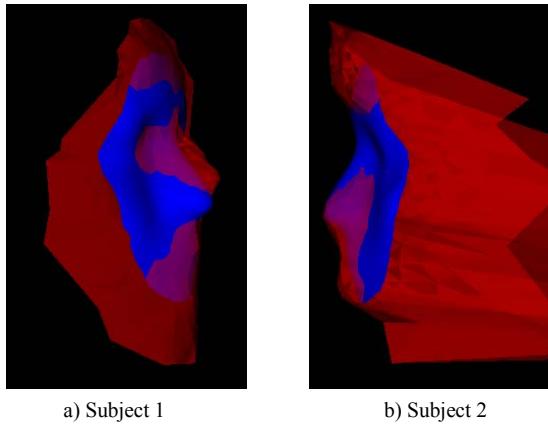


Figure 6: Sample Reference (blue) to Input (red) Face Alignment

We show results of the reconstructed surfaces in Figure 7, 8 and 9. While the sparse data representation is still apparent, the edges have been rounded off.

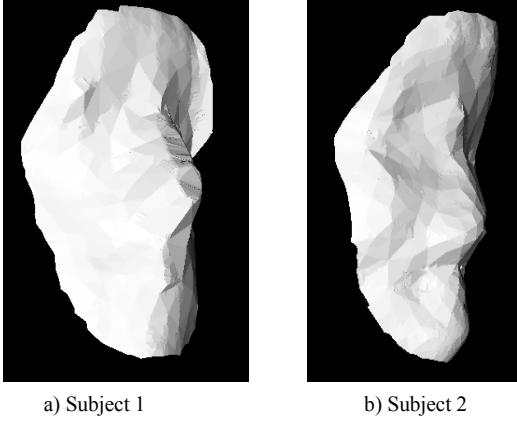


Figure 7: Correspondence Vector Representation

4) Model Merging

The ten SFM models are now represented as registered vectors. Merging of these models is performed by averaging

the contents of corresponding vector locations. The averaging process results in a higher fidelity representation that lacks the faceted appearance common to the SFM models. Traditional approaches for merging would employ vertex set merging followed by remeshing, which can lead to large variations in vertex density. By using the reference face to index vertex geometry, finer control over resolution can be maintained.

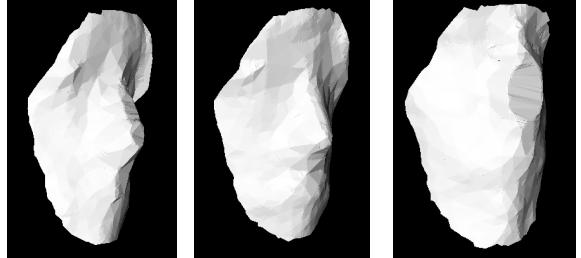


Figure 8: 9 Sample 3D SFM Vector Models Before Averaging (Subject 1)

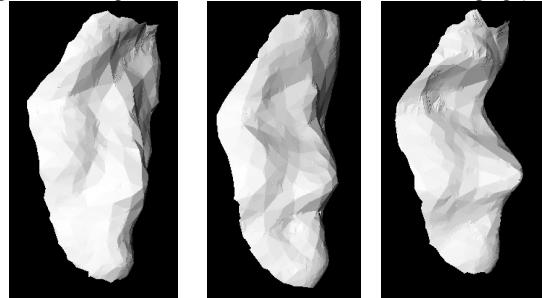


Figure 9: 9 Sample 3D SFM Vector Models Before Averaging (Subject 2)

The merging step tends to smooth concave areas, especially around the nose. We believe this to be the result of occlusion and the sparse density of the SFM models.

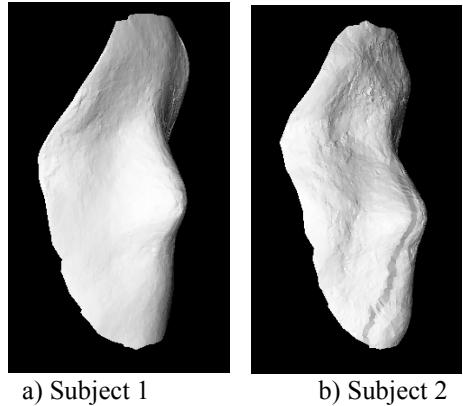


Figure 10: Average Reconstructed Face Models

B. Overcoming Alignment Failures

SFM models are aligned to the reference face by making the centroids coincident and then invoking ICP. Occasionally, imperfect alignments can result, as seen in Figure 11. By averaging out the results the proposed approach is able to overcome a failed alignment. More advanced alignment techniques might eliminate these errors and future work may address this issue. However, in our experience, a single failed alignment among the ten SFM

models does not appreciably degrade biometric matching performance.

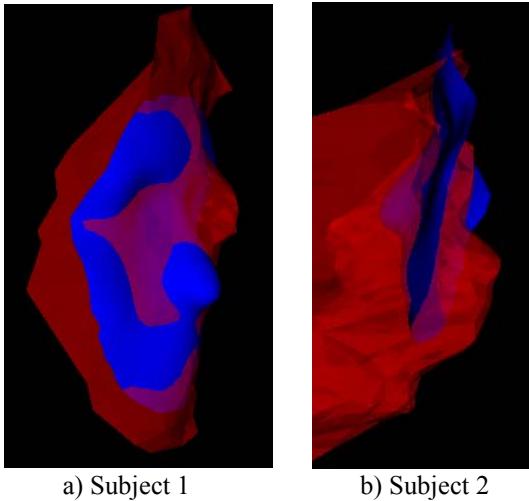


Figure 11: Failed Alignments For Correspondence

V. BIOMETRIC RECOGNITION

To perform recognition, we utilize a whole-face ICP alignment procedure, and use the RMS ICP registration error as a matching score (lower is better). We align the probe face to each face in the gallery utilizing ICP. Then we measure the distance between every point on the probe face to the gallery face and average the results together to determine a match score. We hope to apply more advanced recognition techniques in the future, but utilize this simple technique here because our focus is on model generation and alignment.

A. Biometric Performance

To increase the size of our gallery, we created a unique gallery for each probe consisting of every model but the probe. Utilizing 24 subjects, we produced 48 distinct face models. This allowed for a 47 model gallery for each probe. Our method achieves a 75% rank one recognition rate. We show the rank recognition performance for the raw SFM and our proposed approach in Figure 12.

Our 75% performance includes models of poor quality (generally the result of particularly fast or slow movement), which caused problems for factorization. Some previous SFM approaches have manually adjusted parameters for improved performance. No manual adjustments were used in this experiment. As a result, three models corresponding to three subjects failed to be matched correctly. Had these three subjects been excluded from our experiment for model failure we would have achieved an 85% rank one recognition rate. By contrast, utilizing a single SFM-produced low resolution face model achieved a 38% rank one recognition. Our combination approach proposed here almost doubles the recognition performance to 75%.

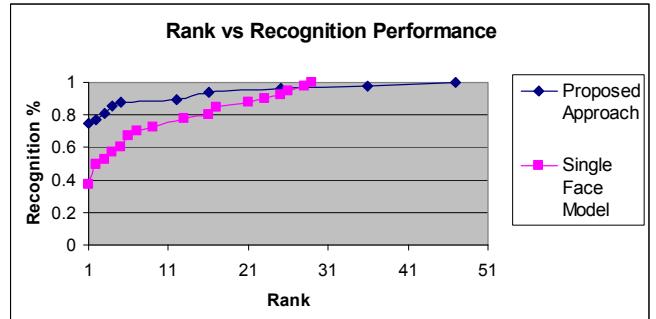


Figure 12: Biometric Performance For Proposed High Resolution Approach

By comparison, previous work using a traditional single region ICP comparison [2] achieved between a 63% to 91% rank one recognition rate using high quality Minolta data from a laser scanner. Our 75% rank one performance lies in the middle of performance for this type of biometric approach and utilizes simpler and less expensive hardware. Further, it can potentially operate on a variety of data sources not originally intended for use in biometrics (e.g., news footage).

VI. CONCLUSION

We have shown a new method for producing a higher resolution 3D face model using multiple low definition 3D face models from a video stream using SFM. The proposed approach for acquiring 3D face data is advantageous because it does not require the use of more complex and expensive active acquisition devices such as laser scanners and can potentially be applied to video from a variety of sources such as news footage.

Utilizing our approach combining multiple models we were able to almost double the performance over using a single SFM model to a 75% rank one recognition rate. For the single region ICP comparison we performed, this puts us in the middle of biometric performance for previously published approaches [2] that utilized laser range scanners for input data.

While experimentation with more advanced 3D biometric comparisons is needed, our model production shows the potential to match the performance of more complex and expensive 3D acquisition approaches. Combined with our ability to utilize uncalibrated video we open up the application of 3D face biometrics to video sources traditionally only possible utilizing 2D methods, especially utilizing “other peoples’ video”.

VII. FUTURE WORK

In the future, we hope to experiment with larger datasets, producing higher point SFM models, and improved methods for biometric comparison. Larger datasets could allow us greater experimentation as well as an indication of performance on a larger subject base. Higher SFM point models should allow for a greater averaged model accuracy as well. Finally, improved biometric methods should improve

performance as well potentially increasing the accuracy into the realm of much more expensive approaches involving laser scanners.

VIII. ACKNOWLEDGEMENTS

Biometrics research at the University of Notre Dame is supported by the Central Intelligence Agency, the National Science Foundation under grant CNS01-30839, by the US Department of Justice/National Institute for Justice under grants 2005-DD-CX-K078 and 2006-IJ-CX-K041, by the National Geo-spatial Intelligence Agency, and by UNISYS Corp.

References

- [1] Chris Boehnen, Patrick J. Flynn, "Accuracy of 3D Scanning Technologies in a Face Scanning Scenario," *Proc. 3DIM 2005*, Ottawa, pp. 310-317, 2005.
- [2] Kyong I. Chang, Kevin W. Bowyer, and Patrick J. Flynn, "Multiple Nose Region Matching for 3D Face Recognition under varying facial expression." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1695-1700, 2006.
- [3] Blanz, V. and Vetter, T. 2003. "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.
- [4] Gerard Medioni, Douglas Fidaleo, Jongmoo Choi, Li Zhang, Cheng-Hao Kuo, Kwangsu Kim, "Recognition of Non-Cooperative Individuals at a Distance with 3D Face Modeling", *Proc. IEEE Workshop on Automatic Identification Advanced Technologies*, Alghero, 2007.
- [5] Cheng, C. and Lai, S. "An Integrated Approach to 3D Face Model Reconstruction from Video," *Proc. IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Vancouver, 2001.
- [6] U. Park and A.K.Jain, "3D Model-Based Face Recognition in Video," *Proceedings of 2nd International Conference on Biometrics*, Seoul, pp. 1085 - 1094, 2007
- [7] Chowdhury, A.R.; Chellappa, R.; Krishnamurthy, S., "3D face reconstruction from video using a generic model," *IEEE International Conference on Multimedia and Exposition*, Switzerland, pp. 449-452, 2002.
- [8] P. Besl and N. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239– 256, 1992.
- [9] Russ, T., Boehnen, C., and Peters, T. 2006. "3D Face Recognition Using 3D Alignment for PCA," *Computer Vision and Pattern Recognition*, New York, pp. 1391-1398, 2006.
- [10] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, and T. Theoharis, "3D face recognition," *Proceedings of the British Machine Vision Conference*, pp. 200–208, 2006.
- [11] J. Shi and C. Tomasi, "Good features to track," *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, pp. 593-600, 1994.
- [12] Lucas, B. D. and Kanade, T. "An iterative image registration technique with an application to stereo vision," *Proc. Seventh International Joint Conference on Artificial Intelligence*, Vancouver, pp. 674-679, 1981.
- [13] A. Jebara T. Azarbayejani and A. Pentland, "3D Structure from 2D Motion," *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 66-84, 1999.