A Taxonomy of Face-models for System Evaluation

V. N. Iyer[1] , S. R. Kirkbride[1,2] , B. C. Parks[1] , W. J. Scheirer[1,2] , T. E. Boult[1,2]
{viyer,skirkbride,bparks,wscheirer,tboult}@vast.uccs.edu
[1]University of Colorado at Colorado Springs and [2]Securics, Inc

Abstract

Generating statistically significant datasets for face matching system evaluation is a laborious and expensive process. Capturing variables such as atmospheric turbulence and other weather conditions especially with respect to face recognition at a distance exacerbate the problem further. It is even more difficult to work on system issues for long-range systems that impact the collection phase such as automated control loops for gain, focus or zoom, as they directly impact the collected data. And since system performance is confounded with variations in subject selection, pose, lighting, expression, etc., formal evaluation of second order effects are difficult without extremely large collections.

This paper describes a taxonomy of face-models for controlled experimentation that overcome these challenges. We show that a gap has existed in experimental design and how a range of model-based approaches can partially fill that gap. Methods for generating 3D models that can be easily manipulated to create variations in pose are presented. Additionally described are techniques for validating and capturing model-based data for use in developing and testing outdoor long-range face matching systems.

1. Introduction

Biometric evaluations must contend with significant uncertainty and variations in their subjects, resulting in the need to collect large datasets and apply statistical comparisons to draw conclusions about system performance. The time/cost/accuracy tradeoff in experimental design is non-trivial. The core of good science includes experimental designs that allow one to test theories and evaluate algorithms. Most areas of hard science design "controlled experiments", where almost everything is held constant and one or a few items are varied, allowing one to better interpret the results. In the words of Nobel Laureate Lord Rutherford, "[i]f your experiment needs statistics, you ought to have done a better experiment." (quoted in [2]). While the advances in quantum mechanics show even physics needs statistics, the issues about designing controlled experiments still remain true. Statistics are needed to account for the inherent uncertainty that will exist in all measurements, but the greater the control over unintended variations during an experiment, the greater the power (in a statistical sense) of the resulting data. Quite literally, control means power.

A secondary, often unstated, goal of biometric evaluations is to suggest performance in operational use. While we can do collections in the lab, increasing our control over variables, performing large scale collections under many conditions is difficult. This is exacerbated by the fact that most biometric systems are strongly impacted by collection, user behaviors and environmental conditions. Thus, for biometrics, experimental control often reduces how it generalizes and applies to real problems.

The relationships between these dimensions and experimental setups are shown in Figure 1, which depicts the correlation of results to operational scenarios on the horizontal axis, and degree of experimental/scientific control that is exercised on the vertical axis. The colors show estimates of the risks of drawing erroneous conclusions when trying to use that data as a predictor for other settings, with red being high risk, orange medium, yellow moderate and green low.

An ideal experiment would be shown in the upper right of Figure 1 and would provide both high levels of scientific control and high levels of operational relevance. For biometrics, this ideal is unrealizable because the environments and subjects cannot always be controlled and because environments and subject change – sometimes intentionally. An important question for our field becomes: what experimental designs can be realized, are there any gaps, and what are the cost vs. performance tradeoffs for those designs?

The classic biometrics experimental processes are depicted via exemplars along the bottom "row" of Figure 1, showing increasing generalization to operational use. Small lab experiments have some control but little operational relevance. They are commonplace for testing but are rarely considered sufficient for publication as it is too easy for them to produce significantly misleading results. Strongly controlled datasets, such as PIE [18], provide increased control and study enough parameters of interest to improve

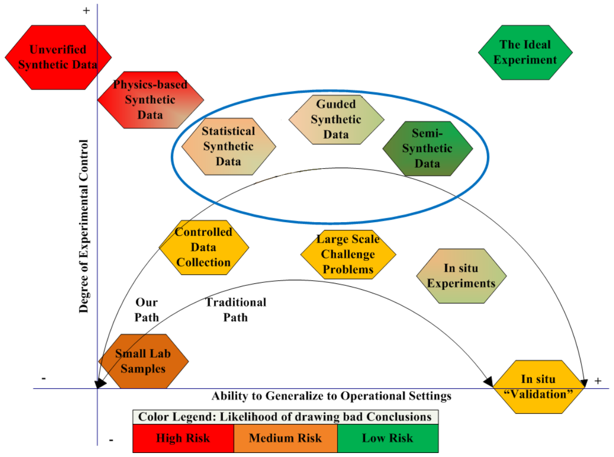
+
Unverified Synthetic Data
The Ideal Experiment
Physics-based Synthetic Data
Guided Synthetic Data
Statistical Synthetic Data
Semi-Synthetic Data
Degree of Experimental Control
Controlled Data Collection
Large Scale Challenge Problems
In situ Experiments
Our Path
Traditional Path
Small Lab Samples
-
In situ "Validation"
+
Ability to Generalize to Operational Settings
Color Legend: Likelihood of drawing bad Conclusions
-
High Risk
Medium Risk
Low Risk

Figure 1. The graph above shows the relationships between different types of datasets: real vs. synthetic. We show that the semi-synthetic approach most closely captures qualities present in the ideal experiment.

their overall generalization. While PIE is small in the number of subjects, its strong experimental design allows for pose to be decoupled from illumination and expression, which has made it extremely popular for evaluation (accumulating nearly 800 citations).

Challenge problem datasets such as FERET and more recently FRGC/FRVT [13] have provided the community with ground truthed data sets with much larger numbers of images in a broader range of conditions, but with much lower control over the collections. They too have become critical elements of evaluation for face matching algorithms and are widely used. Testing closer to operational use, though less widely shared in publication, is still important. Pilot testing, or *in situ experimentation*, provides more control and oversight with more data but often is not conducted with the final system, and the behaviors of pilot subjects are often subject to selection bias and artificial actions. The most relevant testing, however, for exactly one setting is *in situ* validation, which generally has only small samples of ground truth and little scientific control, but has high operational relevance. It may be worth noting that except for *in situ* validation, the remaining techniques all use human "models" whose activities are often controlled and behaviors very cooperative. This can be very important when evaluating the realism related to the actual operational scenario, especially if real systems might have adversarial subjects trying to defeat the system.

Looking at just the bottom row of Figure 1, it is apparent there is a gap in experimental ability – a gap we believe can be filled with model-based experiments. We briefly introduce our taxonomy of model-based evaluations, and then briefly review previous work in related areas. Model-based classes are shown in Figure 1, along the top "row", and represent, from left to right, the least operationally relevant to the most operationally relevant.

Unverified synthetic models At the top-left of the figure, which is below zero on our scale of operational relevance, are pure synthetic models. These models may be artist rendered or simple mathematical models. The models may look fine visually but have no underlying physical or statistical basis and no validation.

Physics-based models are based on structure and materials combined with properties formally modeled in physics. They have have been used for muscle attachments for facial movement and the texture within irises. In a different element of experimental controls, physics models are commonly used for modeling synthetic imaging systems including sensor or lens modeling. Being derived from physics models improves the relevance, but they are only as good as their underlying assumptions, which are often simplified to make the modeling tractable.

Statistical models use estimates of parameters to supplement or enhance synthetic models or physical-models. Because they are tied to measurements from real biometric data they are somewhat valid and tend to have greater predictive power for operational relevance for the the population of the data.

Guided models are individual models based on individual people. There is no attempt to capture properties of large groups or actual physics across models. For guided models, a new model is created for each person, with populations addressed by building many such models. For faces, guided models are composed of 3D structure models and skin textures and hence the models capture many artifacts that are not formally parameterized. Accuracy in model construction will likely be important. The 3D face models can be combined with (physics-based) graphics rendering to generate samples under different conditions.

Semi-synthetic models use measured data, such as 2D images or 3D facial scans as the model. But rather than modeling the imaging system, they are incorporated into a real system for evaluations. We call these semi-synthetic since the underlying data is no longer really a synthetic model, but a re-rendering of measured data instead. Like guided models, semi-synthetic models are derived from individual biometric data and hence can capture important biometric properties that are never explicitly modeled, e.g. distributions of skin textures, facial geometry, and facial hair.

An important aspect of any type of synthetic model is how it is validated. Previous work has looked at the gross shape of match and non-match distributions, but biometric system performance lives in the tails of the distributions. Physics-based models and statistical models would need large scale experiments to validate their relationship to any actual biometric distribution. *A more meaningful and* *natural way to validate biometric models is to "replicate experiments" conducted with real biometric data. This is straightforward for guided and semi-synthetic models and is one of their significant advantages over other types of synthetic models.*

We are, of course, not the first to consider the use of synthetic biometric data for evaluation. Various surveys on synthetic biometric data for evaluation can be found in [11, 10]. These discuss issues related to physics-based biometrics modeling, but they do not seriously discuss issues that arise in using these models for system evaluation, nor how to validate the models

A common rationale provided for the use of synthetic data for evaluation is that it allows for the generation of larger data sets. However, there are no large public synthetic data sets for face and little published use of it in such testing. Rather, the most common use of models is for improving recognition systems: using 3D models for missing data (pose correction for face models [7]), 2D-3D model reconstruction followed by 3D model matching [19], and image face-morphing between views or individuals). These are important works and almost every one uses good face modeling; however, our focus herein is the use of models for biometric system evaluation, which previous work does not address.

A second critical difference between our work and past work is our focus on how the models help control unwanted variations in biometric system evaluations. Prior work that used synthetic data for control did not directly validate models, but rather used synthetic models for testing with control, and then included a small amount of testing with real data for further validation. For example, [1] studied synthetic hand models from many camera views, but only used real data for limited testing. In long-range face evaluation, no such mixture of controllable synthetic and real data has been studied – probably because quality faces are more difficult to properly simulate than synthetic gestures.

Taking this problem to more difficult conditions, there is a growing interest in long-range system evaluation. Yao et al. in [21] attempt to create a data set for long-range recognition. One problem with this data set is fairly obvious – if one wants to test another lens, another camera, another focus-control algorithm, another weather condition, or any other system or environmental parameter other than the matching algorithm, a new collection would be required. A more subtle issue is that, because the data set still contains variables that other studies such as [4, 5, 3] have shown to significantly impact recognition in controlled environments, the conclusions that can be reached are limited because there is insufficient data to reject any hypothesis that does not have a very large effect.

The question at hand is how one can control the same variables studied in [4, 5, 3] in order to gain the power of having a controlled experiment again, including a methodology that will work at long-range and possess significant data realism.

The rest of this paper is organized as follows. In Section 2, we present details on previous attempts at semi-synthetic modeling, which were developed for long-range biometric evaluation during the DARPA HID program circa 2000. These models have significant advantages for face system evaluation. In Section 3 we present our improvements on this concept using 3D guided models for evaluation. In both cases we describe the data and models, the system validation process, and some results. Our focus is on the use of models and the process, not post-processing algorithms. Section 4 discusses data sets captured using our methodology. Experiments and results from various recognition algorithms on data captured outdoors are in Section 5. Finally, we conclude and discuss future work in Section 6.

2. Previous Attempts to Solve The Long-range Face Evaluation Problem

When collecting long-range face data, several problems exist including weather and atmospheric effects (distortion cased by thermal aberrations in the atmosphere). The original concept of *photoheads* was created in [9, 8]. These are semi-synthetic experiments developed during the DARPA HID[1] effort. For these experiments, a long-term setup with 2 cameras at distances of 94ft and 182ft is used for the capture system. The display system consists of a waterproof 800x600 LCD designed for marine use to display a subset of the FERET [14] data set. This controlled setup allows for data capture in various weather and atmospheric conditions. The semi-synthetic use of images controls for local pose, expression and facial illumination over a long period of time. Second, since in depth analysis has been conducted on FERET [14], results can be compared across algorithms.

As described in [8], initial validation was was performed at 15ft and replication of experiments was performed using the CSU Face Identification Evaluation System [6]. The most controlled long-range experiments followed a protocol called "self-matching at a distance", where the same image is used for probe (sample submitted to the system for matching) and gallery (enrollment data). This test allows for a clear indication of variables causing any degradation and showed that, at that time, commercial recognition algorithms were not sufficient for even the short distance to the near camera. These experiments led to the development of algorithms to improve general matching performance in [16]. Multiple observations that were initially non-intuitive, including performance variations at different times of day and weather effects (for example, light rain was better than sunny) were studied in detail.

[1] http://www.itl.nist.gov/div898/itperf/humanid.htm

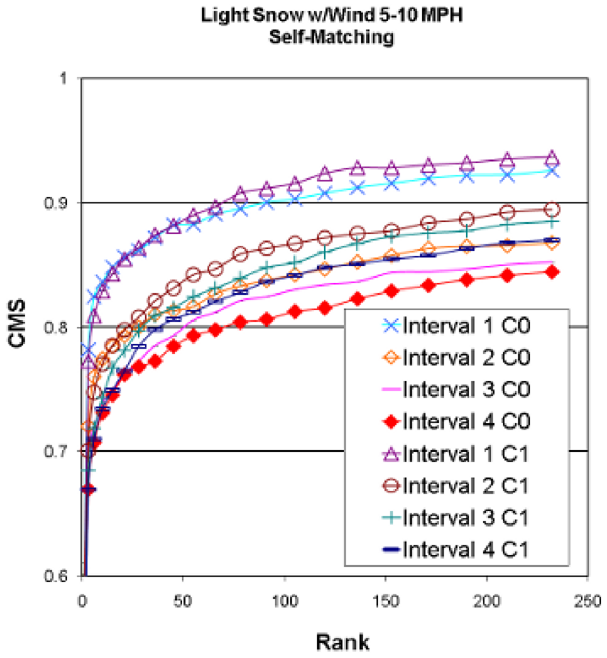

Figure 2. A collection of self matching data over a variety of degrees of light snow precipitation from camera C0 at about 100ft and C1 at about 200ft. Surprisingly, C1 is often better.

Figure 2 is a plot of four time intervals during snow conditions from both the near and far cameras. Normally the close camera outperforms the far camera for this data; however, in this plot, the far camera (labeled C1) consistently outperforms the near camera's (labeled C0) match scores while being nearly twice as far away. Can you think of why?[2] The point here is not to claim that these are novel findings. Rather, we are showing that a unique analysis is only possible with the semi-synthetic data approach of [8].

Another surprising conclusion was that morning was better than midday for recognition performance. The cause was thermal atmospherics, commonly referred to as the "mirage effect". These effects include atmospheric blur and geometric distortions. At long distances these effects can potentially distort the precise features of a face and impact matching performance. To this end, effective deblurring techniques have gained significant interest. Specifically [17] looks at blur caused by atmospherics. Synthetic blur techniques can use models of the point spread function (PSF) of atmospheric blur [12], but that is a type of pure synthetic modeling, as the blur models are not physically or statistically validated.

Current face data sets do not provide data relevant to these problems. Clearly there is a need within the community to do facial recognition research on long-range data sets. We are building such data sets at 3x the distance of the semi-synthetic data in [8] (distances between 100m and 200m). These data sets will allow researchers to further analyze the complex problems that atmospheric blur, geometric distortions, adverse weather conditions, and distance pose to current face matching algorithms. The paradigm using guided and semi-synthetic data could be used by others to design their own model-based experiments.

3. Guided-Synthetic Photoheads

Our extension of the photohead concept moves from 2D static space to 3D dynamic space. Re-capturing 2D images, while valid, has limitations. Factors such as pose and expression are limited to the pictures available within the dataset. 3D models of an entire human head are far more flexible. Easily scripted animation for pose and motion give repeatable results each time and add another dimension of measurement. Additionally, lighting changes within the image are repeatable for every head giving uniformity to each image as well. Clearly, by generating accurate 3D representations of human faces we can prove our concept. The following is an overview of this new photohead system.

3.1. Display Engine

For displaying images, many programs exist with the functionality to display images at timed intervals. While there are programs available to render 3D objects, none seem to exist that can display multiple models for specified lengths of time. In light of this we designed a custom 3D rendering program, which we implemented using the OpenGL, GLUT and DevIL libraries. Multiple 3D file formats were included for support including Wavefront Object and VRML. In addition to displaying a series of 3D objects, each for a specified amount of time, we allowed for script-able size scaling, rotation, and movement along the X,Y, and Z axes. Also, as seen in Figure 3, we incorporate a bar code into the scene, which is used to identify a captured image with an offline decoding program written using the OpenCV library. Additionally, the bar codes display numbers for manual human decoding, if necessary.

3.2. Guided Model Generation

To generate our guided models, we used the commercially available software package Forensica Profiler from Animetrics, Inc.[3] It provides a robust photo mapping system consisting of major and minor facial key-points. A frontal image as well as left and right profile images are used to create the 3D model of the face. Following the blueprint used by [8] we modeled the well known dataset PIE [18]. The right profile and frontal images for each of the 68 subjects are drawn from the same cameras in the lights subset of PIE. An example of a generated model and its source pictures can be found in Figure 3.

[2] The cause is a larger depth of field for the near camera, causing more flakes to be viable obscurities rather than be just blurred out.

[3] http://www.animetrics.com/products/Forensica.php

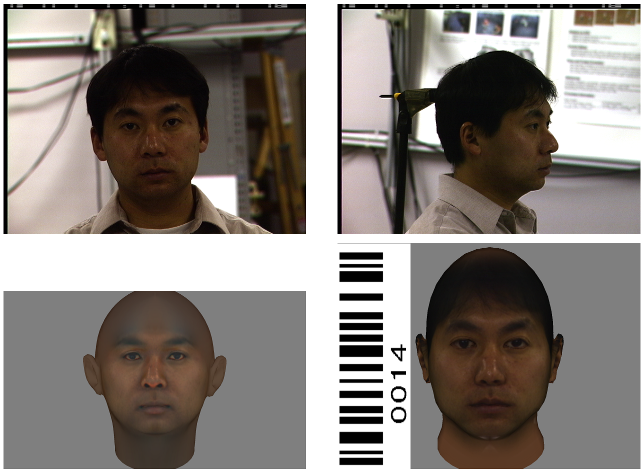

Figure 3. The first row images are examples of the "real data" used to generate the semi-synthetic models derived from the PIE database. The bottom contains a screen shot of a FaceGen model (left), and a screen shot of a Forensica model (right). The bar code in the Forensica image is used as a labeling scheme to automatically identify each person after each collection. While the hair/head region may look unrealistic, only face regions are used for recognition.

3.3. Model Generation Challenges

Model generation went through multiple iterations before we arrived at our current solution. The initial attempt was with 3D data contained in the FRGC [13] dataset. The commercial software product FaceGen, produced by Singular Inversions[4], was used as our second attempt. An example model generated by the FaceGen software is shown in Figure 3 as well as a normalized version in Figure 4. When attempting to test both FaceGen and FRGC 3D models at close capture distance we blamed observed failure mainly on the display setup. We neglected to question two important assumptions: first, that our display program was rendering the models accurately; and second, that the models were really accurate representations of their human counterparts. We found, after analyzing the two assumptions individually, that both were incorrect. After making adjustments to our display program to increase its realism capabilities, we tested screen shots of models from FRGC and FaceGen in recognition tests. Both obtained rank 1 recognition rates below 50% for a self-matching protocol where close to 100% recognition was expected. Thus, a better set of models was needed, leading us to the superior Forensica software.

3.4. Model Validation

Although the Forensica models look to the human eye like the person in the source imagery, we recognize that this subjective metric is not enough. To further verify our models, we conducted recognition tests using two different recognition cores, which are described briefly in Section 5.

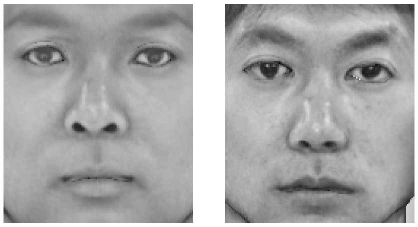

Figure 4. Screenshots from the 3D Photohead display program. Both images are models of the same person. The left image was created using FaceGen, while the the right image was created using Forensica. The full screenshots for these images can be seen in Figure 3. Both are preprocessed with SQI lighting normalization.

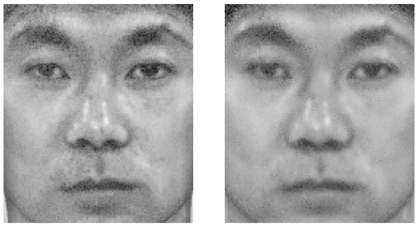

Figure 5. Animetrics model from Figure 4 re-imaged and preprocessed with SQI lighting normalization The left image is indoors at 81M, while the right image is outdoors at 214M.

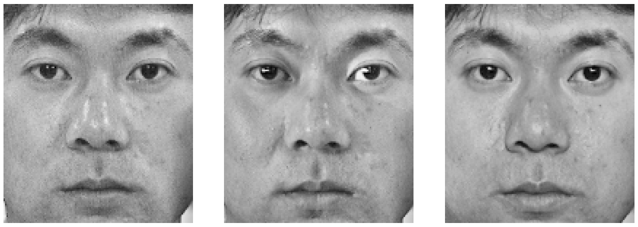

Figure 6. Gallery used to to run experiments with the same preprocessed SQI lighting normalization. These are three normalized PIE images, not models! Compare to Figures 4–5.

For simulated recognition, screenshots of the head model facing forward from the photohead display program were used as probes. The gallery, as seen in Figure 6, used 3 images: the one image from the PIE gallery set and two images from the PIE lights subset not used to generate the mode. We conducted this simulation to ensure that the screenshot was not simply matching back to the same picture due to similarities within the texture of the 3D model. With a perfect rank 1 recognition rate using a V1-like recognition algorithm [15], the models were shown to be accurate representations of the real people.

For our validation, we sought to minimize and eliminate as many variables as possible. We moved from a simulation to a controlled experiment that tested both the capture system and display system indoors at 80m. By conducting the first set of experiments indoors, lighting and atmospherics were minimized. Additionally, we were also able to scale the synthetic data to simulate longer and shorter distances, which eliminated other possible optic distortions. In this case recognition rank 1 results from the V1-like recogni-

[4] http://www.facegen.com/

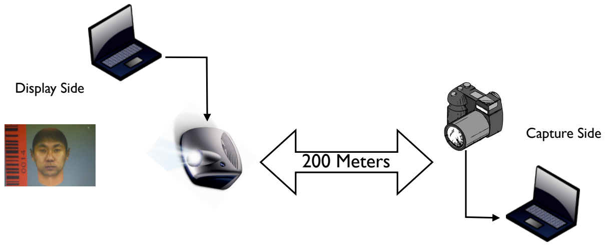

Figure 7. The diagram above shows a high level overview of the photohead system. The capture system consists of a Canon 7D camera fitted on a Sigma 800mm lens with a Canon 2x adaptor, which downloads data to a laptop. The display system consists of a BENQ SP820 projector and custom display apparatus.

tion algorithm were still 100%. Thus, our 3D models and display apparatus demonstrate that we have accurate representations of the PIE dataset [18].

3.5. Display and Capture Setup

Another challenge that we had to overcome was finding an appropriate imaging medium to capture all real world factors present in long distance face recognition. The main obstacle in this challenge was finding the correct display configuration. Most commercially available displays are not bright enough to compensate for the additional light noise from the sun. With a goal of 200m for captures, almost 4x the max distance camera in [8], it became clear that a bigger display and brighter projection device was needed to accurately display and capture models outdoors. To provide flexibility, a portable design was implemented. The entire apparatus from display to capture can be set up in less than 15 minutes by two people.

Figures 7 & 8 depict our current setup, which evolved over a year's worth of experiments with different outdoor displays. Mounted at the front of the display box is a BENQ SP820 projector. It is an extremely bright projector rated at 4000 ANSI Lumens. It is capable of a 1400x1050 resolution, but to achieve maximum vertical refresh rates of 85Hz, it must be used at 1024x768. Even with exceptional brightness it cannot, on its own, overpower the natural light outside. Standing only a few feet away from the screen, the human eye still had trouble viewing the projected image. To combat this, a sun blind was built around the screen to block out external light.

On the capture side, the camera used for data acquisition is a Canon EOS 7D. It is fitted to a Canon 2X adapter and a Sigma 800mm F5.6 EX APO DG HSM lens. The 7D is hooked up to a laptop running custom capture software built using the Canon SDK. Using the integrated face finder in the camera, we locate a face on the display apparatus each time a new one appears and capture an image.

4. Guided-Synthetic Data Sets

We created two data sets from guided-synthetic models: a data set derived from PIE and a data set derived from images specifically created for this work. Below is a description of the three data sets produced from our photohead methodology. Details of how to obtain these data sets will be included in a final revision of this paper.

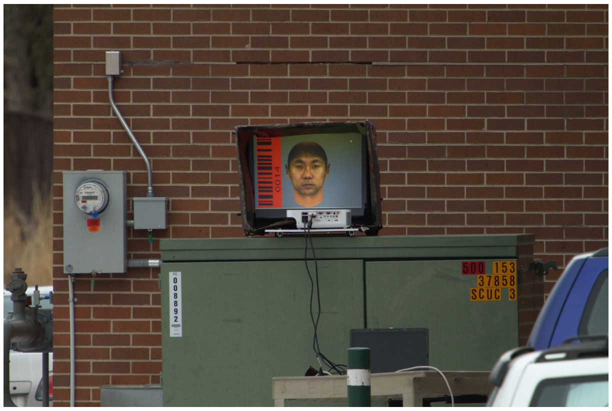

Figure 8. Captured frame example, taken from 214m.

With our sets come several different types of data. The format for the captured images is a Canon proprietary raw format called CR2. They have also been converted to PNG formats for experimentation.The format for the 3D models is Wavefront object file format. The textures for the models are in PNG format. Of course, with the 3D models, any pose variation desired can easily be supported. The poses which we have included in our data are identical to those in the original PIE data set [18]. Each data set consists of a probe set and a gallery set. We ensured that the images used to create the semi-synthetic data in the probe set were different from the images in the gallery set.

To prevent damage to the equipment we needed fair weather conditions to collect data. This had two effects on our data set collections. First, the collections occurred in very bright conditions, which showed that our methodology would work in a worst case lighting scenario. Second, it limited when we could capture. For these reasons the 3D data sets do not have a large a range of data collected over various times of day and weather conditions as is present in [8]. The data sets mainly consist of sunny to partly cloudy conditions captured generally between 11am-4pm.

The internal 3D data was also limited to the same factors as the 3D version of PIE in terms of capture time. The internal set was modeled in same fashion as PIE. Frontal and profile images of 10 individuals were taken under similar lighting conditions with a basic handheld camera. Due to the small assortment of images we recognize that this data set's size may not prove very useful for large scale analysis. However, this dataset is not encumbered by any licensing restrictions, in contrast to the PIE-3D dataset and FERET datasets in Section 2, allowing us to release it in the future for analysis by the community.

PIE 3D has 68 probe images for each re-imaged or screen shot collection. Screen shots are in PNG format. Re-imaged captures are in CR2 and PNG format. The internal 3D probe sets consist of 10 images each. Screen shots are also in PNG format and the re-image captures are in the same formats as the PIE 3D probes.

The original PIE gallery as well as 2 images from the lights subset of PIE that were not used to generate the 3D models were used in our gallery. They are in their original format of PPM. The images used for the internal gallery are in high quality JPEG format. These images were not used to generate any models, but are taken by the same hand-held camera and in similar light settings as the images that were used for modeling.

5. Experiments

With our experiments, we show that using quality guided-synthetic data is a feasible evaluation technique for face recognition algorithm development. To this end we used the PIE data set as a foundation to prototype and lay the groundwork for future large-scale data set generation and captures in the same way that captures were conducted in [8]. We needed to address several issues in the initial captures. A good deal of mobility was needed in the display and capture systems because of the location we were using to validate our tests. This mobility, as well as factors detailed in Section 4, limited the size and conditions in which the data was collected.

With a future goal of long term captures, software was developed to collect an image when a face is detected. However, for these experiments a human conducted the capture of the data and focus of the lens. Manually capturing the data did afford us a few advantages that a full automated capture would not be able to do. We were able to precisely focus the lens to ensure the best image possible, allowing a degree of compensation for atmospheric and distance effects.

Section 3.5 describes the capture and display systems. In this section we describe the validation process on the data we collected. Our goal in validating the data is twofold. First, we intended to develop a scientific system by which our results would be repeatable on a large scale across various data sets. Second, we wanted to show that using a guided-synthetically generated probe captured in a real life scenario would generate the same recognition results as would a real life probe.

We ran a series of tests on the collected data through the use of two different recognition cores. One core, described in detail in [15] as "V1-like", constructs a feature vector for each input image composed of Gabor responses and leverages the power of a multiclass Support Vector Machine for its underlying classification model. In order to utilize this technique, several preprocessing steps were required, as the recognition core does not include any face detection or lighting normalization. Thus, we used the CSU Face Identification Evaluation System [6] to perform geometric normalization based on ground truth that we provided, coupled with Self Quotient Image (SQI) lighting normalization as described in [20]. These procedures were performed on both gallery and probe sets prior to running recognition core.

Data	Iso	Distance	V1	Comm.
FRGC Screen Shots	N/A	N/A	42.11	-
FaceGenScreenShots	N/A	N/A	47.76	-
AnimetricsScreenShots	N/A	N/A	100	-
PIE-3D-20100210B	500	81M	100	-
PIE-3D-20100224A	125	214M	58.82	100
PIE-3D-20100224B	125	214M	45.59	100
PIE-3D-20100224C	250	214M	81.82	100
PIE-3D-20100224D	400	214M	79.1	100
Securics-1-02242010	125	214M	20	100
Securics-2-02242010	250	214M	33.33	100
Securics-3-02242010	400	214M	30	100

Table 1. 3D recognition percentages. The commercial algorithm was not available for use with FRGC and FaceGen. Given the perfect results on AnimetricsScreenShots and PIE-3D-20100210B, we felt there was no reason to run additional algorithms. Screenshots were used for a self-matching test to verify the models, with expected results of 100%. This was to give us a baseline for distance captures. If for instance we used the FaceGen screenshots for distance captures it would have been un-realistic to expect recognition rates higher than 47%. Our main goal was not to create a hard dataset to break algorithms, but rather to validate that the photohead methodology worked.

In order to increase the accuracy of the SVM's convergence, the gallery for both the PIE-3D and internal datasets comprised at least three images per subject, of which exactly three were chosen for the experiments based on consistency of illumination and pose to ensure a well-behaved gallery. None of the gallery images overlapped with those used to generate the models to eliminate the chance that the recognition core over trained on non-face conditions of the particular image used (lighting, reflections, background, etc.). For the PIE-3D set, the official PIE gallery, along with images 27_16 and 27_20 from the 'lights' subset of PIE, were chosen to serve as gallery. For the internal data set, several photographs were taken both inside and outside to fit our pose and illumination criteria.

The second recognition core used is a leading commercial face recognition system. This system does include face detection and its own normalization techniques, so none of the preprocessing described above was necessary. The results for both recognition cores are shown in Table 1. The V1-like recognition core achieves rank-1 recognition of up to 80%, adequately demonstrating its stability and proving that it is a worthy candidate for improvements specific to long-distance face recognition problems, while the commercial recognition system tested establishes a benchmark with a consistent 100% rank-1 recognition score.

6. Conclusions and Future Work

Through various steps of validation, we have proven that semi-synthetic data is a viable alternative to data collected from real people. We have recognized the influential characteristics of both semi-synthetic and real data. These characteristics are: quality input data, robust modeling software, a dynamic display system, and accurate capture system. These characteristics have led to the development of a robust modeling system allowing us innumerable configurations with one real life set of data. The size of our dataset is only limited by the number of models we are able to produce. In addition to isolating these characteristics we expanded the photohead methodology into a 3D embodiment.

Future work will include studying pose and motion in a controlled and repeatable setting using guided-synthetic models. Specifically, since our photohead program has the ability to vary pose and lighting, we would like to create a guided synthetic version of PIE [18], as screenshots and eventually at distance. This would further validate guided-synthetic models as an accurate alternative to real people. Additionally, we would like to conduct long term tests in various weather conditions using guided-synthetic models, as has previously been done with semi-synthetic data [8].

7. Acknowledgments

This work was supported by ONR STTR (Award Number N00014-07-M-0421) and ONR MURI (Award Number N00014-08-1-0638) and NSF PFI (Award Number 650251). We would also like to thank the following people: Justin Schiff for his work on the photohead 3D display software. Sean Cazzell, Patrick Caldwell and Anthony Magee for their help with data captures and building of display hardware. Dana Scheirer for model generation. Chris Olsowka and Jon Parris for model generation, ground truth and data captures.

References

[1] V. Athitsos and S. Sclaroff. An appearance-based framework for 3d hand shape classification and camera viewpoint estimation. In *FGR '02: Proc IEEE Int. Conf. on Automatic Face and Gesture Rec.*, pages 45–51, 2002. 3

[2] N. Bailey. *The Mathematical Approach to Biology and Medicine*. Wiley, 1967. 1

[3] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Comput. Vis. Image Underst.*, 113(6):750–762, 2009. 3

[4] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui. FRVT 2006: Quo vadis face quality. *Image and Vision Computing*, 28(5):732 – 743, 2010. Best of F&G 2008. 3

[5] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 performance. In *FG*, pages 1–8, 2008. 3

[6] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper. The CSU Face Identification Evaluation System Users Guide: Version 5.0. *Technical report, CSU*, 2003. 3, 7

[7] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE TPAMI.*, 25(9):1063–1074, 2003. 3

[8] T. Boult and W. Scheirer. Long range facial image acquisition and quality. In M. Tistarelli, S. Li, and R. Chellappa, editors, *Handbook of Remote Biometrics*. Springer, 2009. 3, 4, 6, 7, 8

[9] T. E. Boult, W. J. Scheirer, and R. Woodworth. FAAD: face at a distance. In *SPIE Conf.*, volume 6944, Mar. 2008. 3

[10] D. Buettner and N. Orlans. A taxonomy for physics based synthetic biometric models. *IEEE Wksp Automatic Identification Advanced Technologies*, pages 10–14, 2005. 3

[11] N. Orlans, D. Buettner, and J. Marques. A survey of synthetic biometrics: Capabilities and benefits. In *Proc. Int. Conf. Artificial telligence*, pages 499–505, 2004. 3

[12] G. Pavlovic and A. M. Tekalp. Maximum likelihood parametric blur identification based on a continuous spatial domain model. *IEEE TIP*, pages 496–504, 1992. 4

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE CVPR 2005 Volume 1*, pages 947–954, Washington, DC, USA, 2005. IEEE Computer Society. 2, 5

[14] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295 – 306, 1998. 3

[15] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? In *IEEE CVPR*, 2009. 5, 7

[16] T. Riopka and T. Boult. The eyes have it. In *WBMA '03: Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, pages 9–16, 2003. 3

[17] O. Shacham, O. Haik, and Y. Yitzhaky. Blind restoration of atmospherically degraded images by automatic best step-edge detection. *PRL*, 28(15):2094–2103, 2007. 4

[18] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) Database. In *Proceedings of the IEEE F&G*, May 2002. 1, 4, 6, 8

[19] M. Vaillant, G. Zang, J. Aliperti, N. Santhanam, S. Doucette, B. Hoffman, and M. I. Miller. Computational anatomy for generating 3d avatars and boosting face recognition systems. In *IEEE CVPR 05 Wksp on FRGC*, pages 150–156, 2005. 3

[20] H. Wang, S. Z. Li, Y. Wang, and J. Zhang. Self quotient image for face recognition. In *IEEE International Conference on Image Processing*, volume 2, pages 1397–1400, 2004. 7

[21] Y. Yao, B. R. Abidi, N. D. Kalka, N. A. Schmid, and M. A. Abidi. Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement. *CVIU*, 111(2):111–125, 2008. 3