Face System Evaluation Toolkit: Recognition is Harder Than it Seems

Vijay N. Iyer[1] Walter J. Scheirer[1,2] Terrance E. Boult[1,2]
{viyer,wscheirer,tboult}@vast.uccs.edu
[1]University of Colorado at Colorado Springs and [2]Securics, Inc

Abstract— **Challenges for face recognition still exist in factors such as pose, blur and distance. Many current datasets containing mostly frontal images are regarded as being too easy. With obviously unsolved problems researchers are in need of datasets that test these remaining challenges. There are quite a few datasets in existence to study pose. Datasets to study blur and distance are almost non-existent. Datasets allowing for the study of these variables would prove to be useful to researchers in biometric surveillance applications. However, until now there has been no effective way to create datasets that encompass these three variables in a controlled fashion.**

Toolsets exist for testing algorithms, but not systems. Designing and creating toolsets to produce a well controlled dataset or to test the full end-to-end recognition system is not trivial. While the use of real subjects may produce the most realistic dataset, it is not always a practical solution and it limits repeatability making the comparison of systems impractical. This paper attempts to address the dataset issue in two ways. The foremost is to introduce a new toolset that allows for the manipulation and capture of synthetic data. With this toolset researchers can not only generate their own datasets, they can do so in real environments to better approximate operational scenarios. Secondly, we provide challenge datasets generated from our validated framework as a first set of data for other researchers. These datasets allow for the study of blur, pose and distance. Overall, this work provides researchers with a new ability to evaluate entire face recognition *systems* from image acquisition to recognition scores.

I. INTRODUCTION

Recognition of human faces is no different from any other complex problem. The general problem has been divided into many smaller problems such as face and feature detection, blur mitigation, pose estimation, and matching algorithms. Researchers constrain these problems even further. For instance, a recognition core can be tested using images and ground-truth of all the feature points necessary for the algorithm. While this is useful to initially design the algorithm, the constraints allow the designer of the algorithm to assume that perfect feature points will always be supplied. Thus when a recognition core is coupled with a feature detector it may not perform as intended because the detector's interpretation of good feature points is completely different.

With current advancements in facial recognition algorithms, combining multiple solutions for individual smaller problems and usefully applying them to real world scenarios is slowly becoming a reality. Datasets and evaluation tools are an important element in advancing to this goal. While many good datasets exist, Pinto et al. [17] brings to light that some of the popular face datasets, such as LFW [11], may not accurately simulate operational scenarios that a real recognition system would be exposed to. They point out that apparent improvement in algorithms could be due to accidentally exploiting irrelevant image characteristics in a given dataset.

Using a relatively simple algorithm, described as "V1-like", Pinto et al. were able to achieve comparable or even better performance compared with complex algorithms. Their work shows that there is a need to design datasets that are closer to operational scenarios. They suggest that a trivial algorithm establishes a baseline that other algorithms can attempt to improve upon. By using a "simple" recognition core, such as their "V1-like" implementation, datasets can be evaluated for the presence of low-level regularities to prove if they provided a significant enough challenge to algorithm developers. Synthetic data is also suggested as a way to create a more realistic set of data because of its flexibility.

Datasets alone cannot address many important dimensions of face-system recognition. Pose, blur and distance are some of the variables that a system may need to handle when recognizing a subject in an unconstrained or surveillance setting. Consider an attempt to recognize people in these types of applications. Obtaining imagery of uncooperative subjects may not always yield the best pose – the subject may be in motion under low light and changing the vantage point is not always option. Thus, the ability to recognize a face under any pose at a distance with blur becomes a necessity. In a maritime environment, the problem of our sponsor, both atmospheric and motion blur impact recognition results. Atmospheric blur will always be a problem during the day – especially in warmer weather. Motion blur is amplified by the fact the subject may be on a vessel that could be moving in almost any direction due to the ocean. Since there is no way to control either type of blur, it leads to challenges in how to evaluate systems for such an environment.

In order to improve face recognition algorithms with respect to these variables one must collect a dataset that allows for the study of how they truly effect recognition. Due to a lack of long distance datasets there are not many ways for researchers to evaluate how long ranges effect recognition cores. Currently only one dataset with real subjects has been created by Yao et al. [23] – and it is not publicly available. Available sources of data for blur are also lacking. Pavlovic et al. [15] have resorted to using point spread functions (PSF) to synthetically generate models of blur to study.

On the other hand, for pose, we find quite a few datasets. PIE [19] & Multi-PIE[8] are most commonly used to evaluate pose. FERET [16] is also used though its pose variance is limited. Even with the large amount of research into this variable, multiple surveys [24], [20], [25] have concluded that pose variation is still one of the major unsolved problems for face recognition. Kroon et al. [13] also draws a similar conclusion, namely that because most algorithms are designed for only frontal poses, recognition in an unconstrained environment proves to be a non-trivial problem.

50 of the 200 models in the dataset found in [3] were used by Levine et al. [14] as a dataset to evaluate algorithms on pose. One thing to note about the study is that it also included analysis of expression and illumination in addition to pose. For expression and illumination they used the PIE and Yale [7] datasets respectively. We note that it is interesting that they chose to use a semi-synthetic dataset even though they were already using a dataset that contained subsets of pose variance. They use this data to simulate the method described in their work of capturing real subjects using a video camera. The real subjects turn their heads at different vertical tilts, while being recorded, to gain a large sample of poses. This is designed to eliminate the need for simultaneous multiple camera capture for every desired pose, as was done in datasets such as PIE. While we are not disputing the validity of the method they propose, choosing to use synthetic data over real subjects shows how it is a more flexible medium that allows for a finer degree of control over experiments.

Attempting to gather imagery in operational settings is not an easy task and introduces additional problems. The obvious problem is that placing cameras in multiple locations at long ranges and at the same distance from the subject would be harder and more expensive than an indoor short range setup. Not only would you need more sets of expensive equipment but setting it up at equal distances from the subject and at the same line of sight angle would be another obstacle. Even if cost is not an object, dealing with the synchronization of all the cameras over a wireless network would be difficult due to delay and packet loss. More importantly it does not allow researchers optimum control over their experiments, because it will be hard to decouple pose and blur from each other. Thus the question this paper aims to solve is not pose, blur or distance specifically, nor is it to evaluate the specific effectiveness of algorithms when dealing with these variables. Rather we want to solve the problem of being able to create datasets that allow researchers to effectively study pose and blur at distance.

Previous works in this area have been helpful, but they don't address our motivating problem – to be able to evaluate face data at statistically significant levels in a maritime environment. Using real subjects is difficult even in simple uncontrolled indoor settings. Adding distance, weather, water and boats into the mix and the number of samples needed to draw a significant conclusion quickly becomes intractable. Add the issue of trying to get people on/off a boat to do a collection and it becomes clear that creating a traditional dataset in a maritime setting is impractical. Synthetic datasets offer an appealing solution to this problem. Large numbers of models can be generated which addresses the statistically significant size dilemma. The models can be displayed in an exact repeatable manner, which makes each created dataset controlled except for environmental changes.

We propose using the synthetic data framework created in our previous work [12] where we generated a guided-synthetic set of 3D models. As opposed to semi-synthetic data, models from guided synthetic data use properties of an individual to create a model but it is not a direct re-rendering of the person like that of a facial scan or image. Instead a guided model uses the properties of an image/scan to generate the shape and texture of the model. This potentially results in better models than the one produced by a scan. A system defined by Blanz et al. [3] created semi-synthetic 3D heads from facial scans. The dataset also had the ability to change illumination and expressions of the models in addition to pose. Iyer et al. [12] define a taxonomy for classifying types of synthetic face data and their relation to experimental control. They define semi-synthetic as a re-rendering of real data such as a facial scan or a 2D image.

With this guided-synthetic method we have created two outdoor long range datasets of pose and blur as well as a screenshot dataset of pose. It is our intention to provide the toolkit and associated datasets presented in the following sections of this paper to the biometric community. The 3D models and the program used to display them are available from the authors . This paper validates this approach, and we are now building hundreds of 3D models to be used in a real maritime collection.

This paper is organized as follows. In Section II we discuss previous work using the photohead concept. In Section III we describe the new datasets we have collected with the photohead methodology. Experiments on the new datasets are discussed in IV. Finally we conclude and discuss how to obtain the dataset in Section V.

II. PREVIOUS SYNTHETIC DATA WORK

Pinto et al. [17] conducted pose variation tests using unverified 3D models created using the commercial software package FaceGen, produced by Singular Inversions (http://www.facegen.com/). As stated before their results on LFW [11] were comparable to more advanced algorithms. Running the same algorithm on the generated models, which were considered to be much simpler, their results quickly dropped to around 50 percent as pose variance was increased.

Our previous work in [12] expanded on the original concept of *photoheads* created by Boult et al. [5], [4]. The original concept re-imaged pictures from the FERET [16] dataset on a waterproof LCD panel mounted outside on a roof. Two cameras were mounted at 94ft and 182ft from the panel to capture images of the screen. Since it was a permanent setup it allowed for data capture at different times of day and weather over long periods of time.

In the new setup described in [12] we defined classifications for synthetic data and expanded on the overall concept. Our redesigned setup uses a high powered projector instead of an LCD. Instead of re-imaging pictures, we created guided-synthetic 3D models based on the PIE dataset. Since the display of the new apparatus was not weatherproof and we did not have a permanent display area we could not gather the same types of long-term data as the photoheads described in [5], [4]. However with better imaging equipment we were able to create a frontal dataset of the guided-synthetic models re-imaged from 214 meters.

While the photohead method potentially introduces new covariates, the results from using an LCD in [5] and our most recent work with the projector in [12] show that any new covariates introduced do not affect recognition scores for the algorithms tested. While removing some covariates is a goal of photoheads that is not the only motivation for using synthetic data. Collecting datasets with a large number of subjects, which is not a feasible task using real humans, becomes possible with the use of synthetic data. The main reason however that one would want to use photoheads is to be able to conduct the same experiment in a repeatable fashion. Even if the re-imaging process of photoheads is adding covariates, it will always add these covariates thus creating a controlled repeatable experiment.

Going one step further than similar synthetic datasets such as [3], [10] we also validated that our models were equivalent to the 2D images of the subjects used to create them. This is important as many researchers will attempt to invalidate results solely based on the fact the experiments were conducted using synthetic data. To do this we followed a procedure similar to [4], consisting of "self-matching at a distance" tests that matched the same image from the FERET dataset back to a re-imaged version taken 15ft away. However, our test for the 3D models was not exact self-matching, as the images were not the same. Instead we used three frontal gallery images of the same people not used to generate the models. In this way we ensured the recognition algorithms were not simply matching the texture back to the picture. Since the models are based on a well known dataset, validating the models allows us to compare the guided-synthetic data to our collected data.

III. TOOLKIT & DATASETS

For both blur and pose datasets, we used the 3D models generated and tested in our previous work [12]. One set of models was created using the Forensica Profiler software package from Animetrics, Inc. (see http://www.animetrics.com/products/Forensica.php). We also tested another software package, FaceGen which was used by Pinto et al. for their experiments in [17]. Both software packages were given a single frontal and profile image to generate the model with. FaceGen has the user manually adjust key feature points on the images provided. The software from Animetrics takes tries to map a set of major and minor feature points to the image. These points can then be manually adjusted by the user. Using these points, both software packages generate 3D points and a texture that can be saved in a WaveFront Object file format. As stated in Section II the models were generated from images out of the PIE dataset. Recognition results on screenshots from both packages allowed us to determine that the software package from Animetrics generated a more accurate model from a statistical standpoint. Also stated in Section II we used a gallery of three photos from the real PIE dataset and a single screenshot as the probe. Using this testing protocol Animetrics models were able to achieve 100% recognition rate. On the other hand FaceGen was only able to achieve a recognition rate of 47.76%. The experiments conducted in this paper use the Animetrics model set.

Data Set	Distance	Poses	Subjects	Total Images.
PIE Pose Screenshot	NA	13	68	884
PIE Pose Distance	214	13	68	883
Blur Set	214	1	67	67

TABLE I. The Pose Screenshot and Distance databases contain 68 subjects in 13 different poses at a distance of 214 Meters. The Pose Distance dataset is missing one image from view C22. The Blur set is missing one of the 68 models as well.

We created both re-imaged long distance datasets and a screenshot dataset. The screenshot dataset was taken while running our custom display software which leverages OpenGL to render the models. These were taken at a resolution of 1900x1080. For imaging the long distance sets we used a Canon EOS 7D. A Sigma 800mm F5.6 EX APO DG HSM lens and 2X adapter is attached to the camera. Images from the camera were saved in the proprietary Canon CR2 format at a resolution of 5194 x 3457. The models were displayed in a specially designed display box running our display software at 214 meters away. The system used a 4000 lumen BENQ SP820 4000 projector, displayed at a resolution of 1024x768 with a refresh rate of 85Hz, approximately 18in from the screen.

A. POSE DATASET

In a synthetic environment we have nearly infinite pose configurations. However, this is not very useful unless we validate our models to be equivalent to different poses of their human counterparts. In order to validate the guided-synthetic pose set, comparison to similar human poses was necessary. The logical choice was to re-create the PIE set they were modeled after in synthetic form since the set itself has multiple poses in which the models can be validated against. Using all angles documented by Gross et al. [9], except for pose C07 (refer to Figure 1 for angles used), we created our own guided-synthetic version of the PIE dataset. We did not use C07's documented angles because when used in the rendering program it did not change pose variance when compared to C27. Instead we estimated the angle by slowly varying the pose until it looked close to the original PIE picture angle.

Like the PIE dataset each set created has 68 subjects imaged at 13 poses. Two sets of pose were generated. The first consisted of screenshots of the models in each of the 13 poses. This set was used mainly to validate the ability of the models to accurately reproduce the pose of the subjects. The second was of all the poses re-imaged at 214 meters. Note for the distance pose set we are missing one image from pose C22(see Figure 1). Set statistics can be seen in Table I and examples of each of the 13 poses from the set captured at 214 meters can be seen in Figure 1

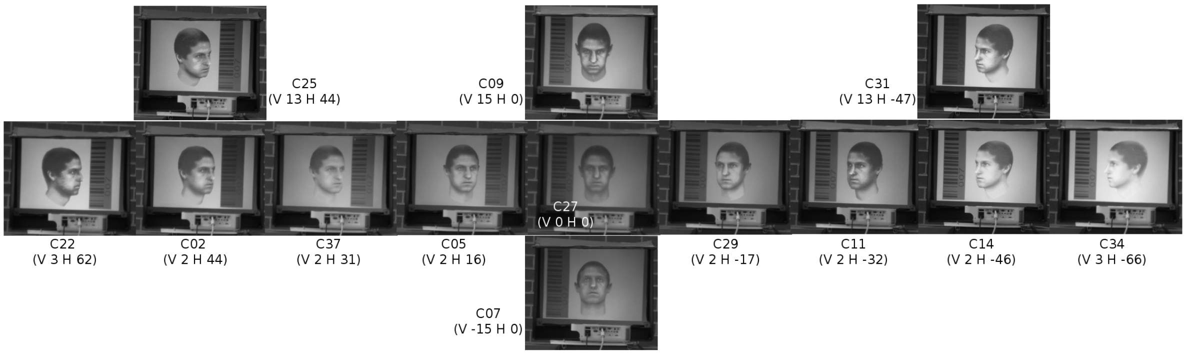

Fig. 1. This figure shows all of the 13 poses captured in the long distance pose set. The vertical and horizontal orientation of the faces are labeled in each image. The camera names from the original PIE database are used to allow easier comparison between datasets

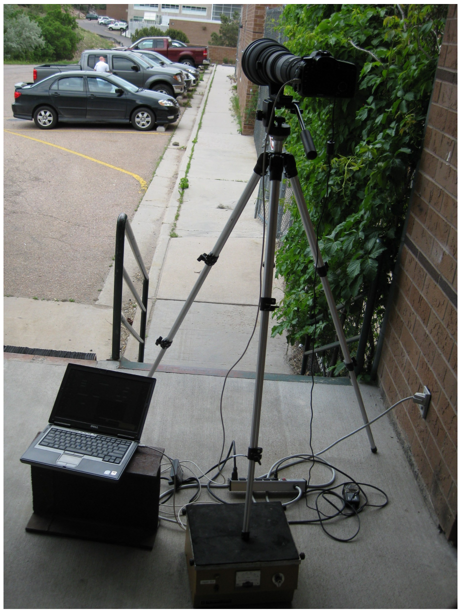

Fig. 2. The shaker table moves the leg of the tripod to cause varying degrees of blur depending on its speed. This type of motion blur is relevant to operational settings as it could be caused by someone walking close to the capture setup causing vibrations. This produces motion blur in the captured images.

B. BLUR DATASET

In the maritime environment we expect significant motion blur. Before testing in that environment we wanted a challenge set with blur, some controls, and the ability to validate performance. Creating a blur set proved to be more of a challenge than initially expected. Our first idea was to use the photohead program's ability to animate the head as a way to create motion blur. However, because the screen's refresh rate was too high no motion blur was picked up by the camera or human eye. The second idea was to create a fake blur within the program using OpenGL texture tricks that create the illusion of blur on the screen. We did not pursue this idea as it became apparent that it would only create pure synthetic blur for us to re-capture. Our goal was to capture real blur.

Finally we decided to actually cause motion in the capture setup itself. Using a "shaker table", used for shaking elements in a chemistry lab, we tried making one of the endpoints of the apparatus have motion. The display end could not be moved fast enough as it was too heavy for the table. Instead we looked to the capturing side of the setup. By propping one of the legs of the tripod setup onto the shaker table we effectively created motion blur in the images. A picture of the shaker table setup can be seen in Figure 2. Since the shaker table's speed could be manually adjusted, the amount of blur could easily be controlled and repeated as necessary. This type of motion blur actually has a high operational relevance as it is possible for someone to shake the tripod holding a camera by simply walking in the vicinity. Using this method we captured a blur dataset at 214 meters. Assuming a linear blur model described in [6], images in this set contained an average of approximately 18 pixels of motion blur. Due to an error during data capture, images were taken of 67 of the possible 68 subjects in the set. Also only the blur set contains no pose variation. A breakdown of the set can be seen in Table I.

IV. BASELINE EXPERIMENTS AND RESULTS

To run experiments we used the same two recognition algorithms used in [12]. One is a "V1-like" algorithm described in [17], the other is a leading commercial algorithm. Both were implemented into a pipeline setup consisting of a watchlist enrollment phase followed by a recognition phase. Both portions have the option to use the Viola Jones face detector [21] and the eye detector used in [18]. The "V1-like" algorithm is the same as in [12]. The geo-normalization from the CSU Face Identification Evaluation System [2], and Self Quotient Image (SQI) lighting normalization described in [22] are integrated into the recognition process.

For the commercial algorithm we re-implemented the code using a provided SDK to speed up the testing process. In our previous work [12] we did a 1-to-1 verification comparison of each probe and gallery as opposed to generating a watchlist. Also in the previous work we cropped the image around the display apparatus and used the commercial algorithm's face detector. This worked fine for the frontal tests as the face detector performed well and was essentially equivalent

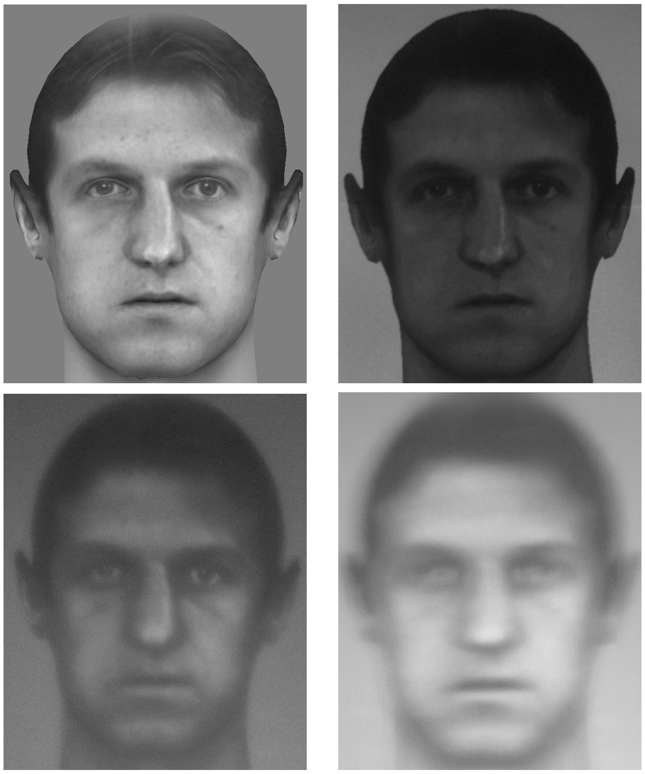

Fig. 3. Top row: Left image of screenshot. Right image re-imaged 81 meters indoors. Bottom Row: Left re-imaged 214 Meters. Right re-imaged 214 meters with motion blur. We are able to add difficulty to the exact same model simply by changing the setting in which it was imaged.

to our “V1-like” tests; that work also used ground-truth to geo-normalize the images before applying the recognition algorithms. Instead of cropping the images we now use ground-truth for both algorithms.

A. *POSE EXPERIMENTS*

To be able to compare to previous work we replicated the experiment design in Gross et al. [9]. For each test they selected a single pose for the gallery and proceeded to match on each pose as the probe. This resulted in 169 different test combinations. We conducted this test with 6 different probe/gallery variations. Table II shows the probe gallery combinations of the dataset. We used two subsets from the expressions set of PIE, screenshots of our guided-synthetic models and re-imaged models at a distance of 214 meters.

For our tests we ran two different variations. The first test ran without ground-truth allowing both cores to find the face and features automatically. Eye coordinates for each image were given to the recognition pipelines for the second test. A total of 2028 tests combinations were evaluated on each pipeline (2 test types * 13 probe poses * 13 gallery poses * 6 probe/gallery combinations). We wanted to avoid the possibility of potentially having an incomplete watchlist generated. So for the enrollment phase ground-truth eyes were used exclusively for each test. Due to large number of tests we cannot display all the collected data. The toolkit available from the authors provides additional data not presented in this paper so that researchers may have an accurate baseline when attempting to replicate experiments.

Probe	Gallery
Real Pie	Real Pie
Real Pie	Synthetic Pie Screenshots
Real Pie	PIE-Pose-Distance
Smile Pie	Real Pie
Smile Pie	PIE-Pose-Distance
Smile Pie	Synthetic Pie Screenshots

TABLE II. A list of probe gallery combination. Real PIE refers to the neutral pose subset of the PIE expression set. Smile Pie is the Smile subset of the expression set. Synthetic Pie Screenshots are screenshots of the 3D models rendered. PIE-pose-distance is the 3D models re-imaged from 214 meters away.

Validating our guided-synthetic models in our most recent work with photoheads was done by achieving 100% recognition on screenshots of frontal oriented pose. The gallery used consisted of 3 different frontal images that were not used for model generation. Since pose is not solved to the same degree as frontal imagery we could not use the same metric to validate. We re-ran the self matching tests using the neutral expression in PIE as both the probe and gallery. This was to give ourselves a baseline to compare to since we could not directly compare to Gross et al.’s results. Using the screenshots as the probes we conducted the same tests. As seen in Figure 4 the rank 1 recognition results for self matching tests generally performed better except in a few cases when the pose was varied much farther from the gallery image orientation. Recognition performance of the screenshots decreased in the same manner as the self matching set.

Since the performance is of a similar nature one could argue this is enough to show equivalency of the models to their real life counterparts. However using frontal pose for both gallery and probe the screen shots actually missed one image when we used the “V1-like” pipeline. This is not the 100 percent we had seen before with a similar validation test. As we stated above our original test contained 3 gallery images instead of one. Even with this explanation the test results being lower across the poses for the screenshots prompted us to conduct another experiment for further validation. Our hypothesis was that a different image of the same person used as the gallery would yield similar results to that of the screenshots. To do this we re-ran both of these tests but used the smiling subset of PIE expression as the gallery. Although this adds the variable of expression into the mix, Beveridge et al. [1] concluded that if only one image is enrolled in a gallery, having the person smiling is better than a neutral pose. On the frontal test it actually missed more images than the 3D models, not being able to recognize three subjects. Referring back to Figure 4 it can be seen that when smiling PIE is used as the gallery real PIE has recognition rates closer, and in some cases at the exact same level, to that of the screen shots.

Figure 5 shows a comparison of rank 1 recognition percentages. The graph displays results both commercial and “V1-like” pipelines while using ground-truth eye points. The gallery image has a horizontal orientation of 17 degrees to the right. Since the gallery is still facing in a relatively frontal view, it is no surprise that only extreme variations

V1			
DataSet	GT	No GT	Cropping no GT
Blur Set	%47.76	%0	%26.87
PIE-Pose-Distance Camera_27	%54.41	%0	%30.88
Commercial algorithm			
DataSet	GT	No GT	Cropping no GT
Blur Set	%97.06	%0	%97.05
PIE-Pose-Distance Camera_27	%100	%5.88	%98.53

TABLE III. This table shows rank 1 recognition results for both pipelines on 2 sets of data containing frontal poses. Both sets of data are taken from 214 meters distance. Motion blur is added to the blur set. This has more of an effect on the "V1-like" algorithm than the commercial. With GT or cropping, the commercial algorithm significantly outperforms the "V1-like" algorithm on both sets of data. When given the whole image, without any ground truth data, both pipelines fail miserably.

in pose perform poorly across all tests. It is obvious that the commercial algorithm steadily outperforms the "V1-like" core. On the distance set the commercial algorithm shows the largest difference in performance when compared to the "V1-like" core. Figure 6 shows the same results when the gallery camera is varied an additional 15 degrees for 32 degrees of horizontal rotation. Even with a small increase to rotation both algorithms see a significant decrease in performance.

The next graph in Figure 7 shows some of the more interesting rank 1 recognition data. It shows the results from both the recognition pipelines using a frontal image of real PIE as the gallery and the PIE pose distance set as the probe. As with the graphs in Figures 5 and 6 the commercial pipeline far exceeds the results of the "V1-like" pipeline when given ground-truth. However when the ground-truth is taken away suddenly both behave extremely poor. Not a single face is recognized with the "V1-like" pipeline. The commercial pipeline cannot achieve above 5.88%.

B. BLUR EXPERIMENTS

To test blur we used the frontal neutral expression pose from PIE as the gallery and compared it to the blur dataset captured at 214 meters. Since we had already validated the frontal pose of our models in [12] there was no need to do so for this set of tests. In table III we show the recognition results of the blur set.

We compared the recognition results on the blur set to that of frontal pose from the pose set as seen in Figure 1 as C27. A cropped version of this and the blur image can be see in the bottom row of Figure 3. As expected adding blur makes recognition on the same dataset more difficult. As with the pose set, the commercial algorithm outperforms the "V1-like" core when using ground-truth or a cropped image. When given the entire image without any ground-truth both perform dismally. Only on the frontal pose set was the commercial algorithm able to recognize any faces at all. Every other test resulted in no faces being recognized.

V. CONCLUSIONS

After conducting multiple tests we conclude that while Pinto et al.'s claim that datasets are too easy and not relevant enough to practical recognition scenarios has some validity, their concerns are not the only problem. We agree with their conclusion that algorithms may be exploiting attributes of certain datasets, yielding unrealistically optimistic results, which is the first half of the problem. Thus, improving dataset design to limit these variables is part of the solution. The second half of the problem we concluded is researchers actively or implicitly applying significant constraints to the problem by the way they conduct testing. Most experiments on recognition algorithms are given clean data with a cropped image and/or coordinates of feature points. As seen with our tests on the blur set when either algorithm is given nothing but an unprocessed image, which is truer to a real life implementation, they perform poorly. When given ground-truth or cropped images, both algorithms see a drastic improvement in performance. Even if no cropping is done, most datasets are at very close range with the face dominating the image, so there is little difference if the image had been cropped around the face.

Pose and blur are still unsolved, but important problems. Furthermore, outdoor distance adds complexity to recognition problem. Close range frontal recognition is widely viewed as essentially solved. By simply adding distance we turned what seemed to be an easy problem based on a well known dataset into an extremely difficult challenge. The photohead system allows us to evaluate algorithms, and more importantly entire face recognition systems, with more relevance. The key is for researchers to use the data collected appropriately and not over constrain it to the point of making experimental results look as though the problem has been solved.

The true contribution of this paper is the toolset for and validation of our 3D photohead methodology. Our previous work using guided-synthetic 3D models in [12] was evaluated only using nearly frontal approach images. This paper shows that even under different poses our guide-synthetic models are equivalent to their real life counterparts. While PIE might be viewed as a smaller set, now that we we have validated the process it opens the door to a much larger variety of data that can be generated from our models. Using this method has enabled us to create multiple datasets for public release. It has also created a framework for evaluating entire face recognition systems. By releasing the 3D models and our display program as a complete toolkit, we are enabling researchers to use this set of tools to conduct their own experiments. Researchers have the potential to improve recognition rates not only by using better algorithms but also by using higher quality imaging techniques. [17] asked for a more difficult dataset to solve. Not only are we providing a more difficult dataset, we are giving people the tools needed to design and create datasets. If researchers feel our dataset did not provide them with enough challenge, they now have the ability to design a dataset as difficult as they desire.

The complete toolkit, including datasets and the 3D models presented in this paper is available from the authors. Additionally, the program used to render the models is also available. Anyone producing proof of license for the PIE [19] dataset will be allowed to obtain the PIE models. This is due to the licensing restrictions of PIE and our agreement with CMU. Similarly we can also provide our (2D) FERET photheads and MBGC-based dataset with hundreds of 3D

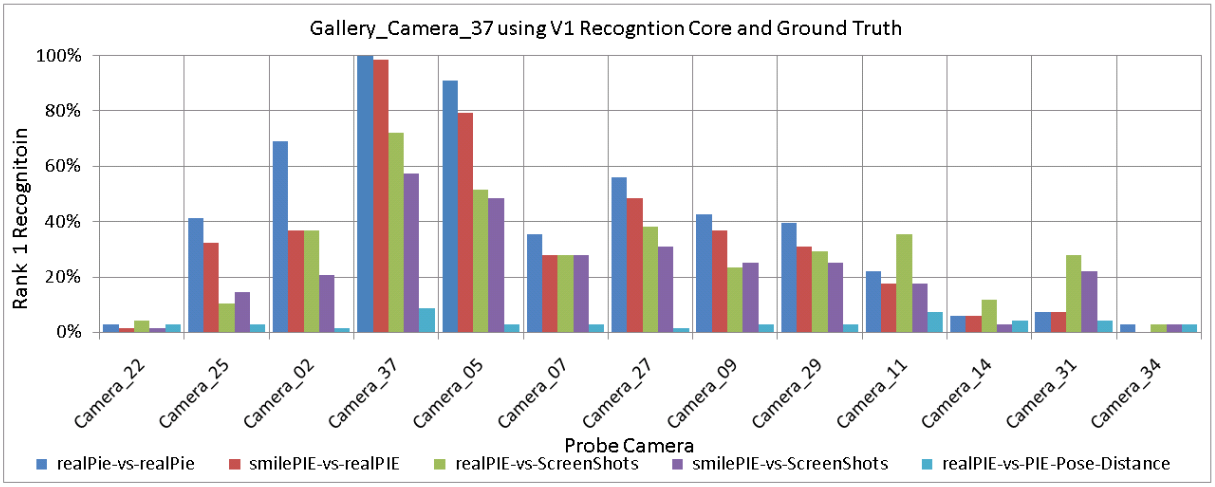

Fig. 4. This graph shows the results from the "V1-like" recognition core using Ground-Truth points. Each Bar is a Gallery-vs-Probe combination. When using the real PIE set as the gallery the screenshots perform worse when compared to the real PIE set, except for a few cases. However, when the gallery is changed to the Smile PIE subset, the real PIE set results go down. In many cases it even gets the same recognition results as the screenshots, showing that the difference in performance is mainly due to the similarity of the pictures and not because the data is synthetic.

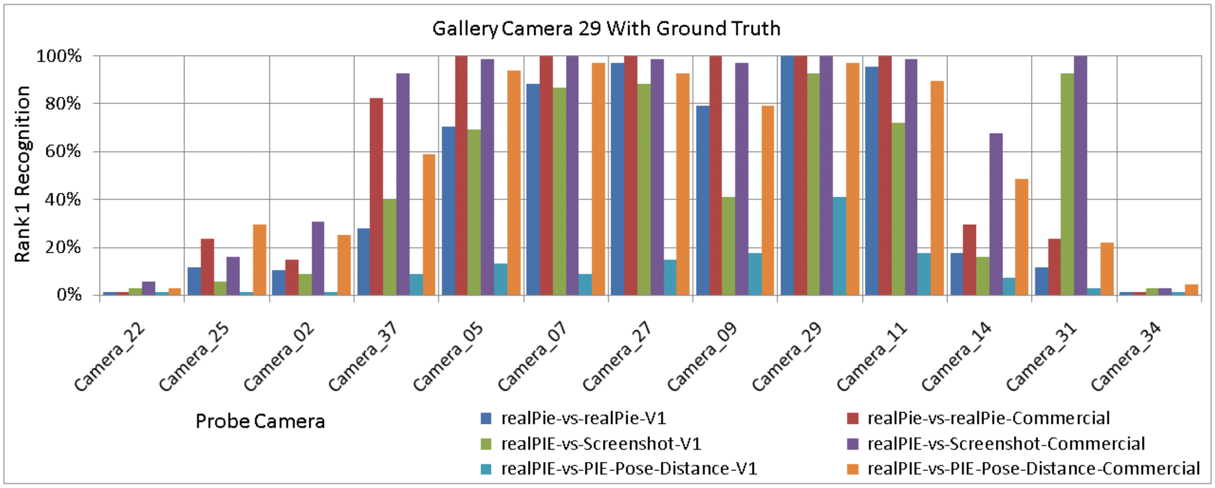

Fig. 5. Using a small degree of pose variation (17 degrees) for the gallery image results in relatively good results when the real PIE and screenshot sets are used as probes, if the pose variance was within 32 degrees. Both algorithms were given eye coordinates. The commercial algorithm clearly out performs the "V1-like" core. For the distance pose set, the commercial algorithm still gets useable results to the rest of the tests. The "V1-like" core is no better than chance even with little to no pose variation in the probes.

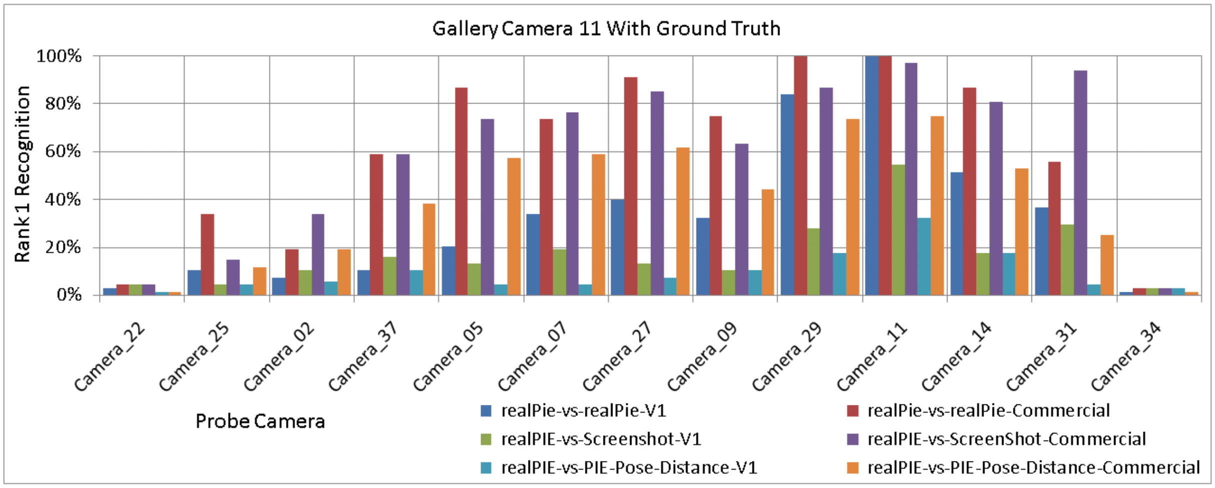

Fig. 6. Same type of graph as Figure 5 except the gallery camera has horizontal pose variance of 32 degrees. Even with this small increase the results go down drastically for both algorithms even on images with little or no variation in pose.

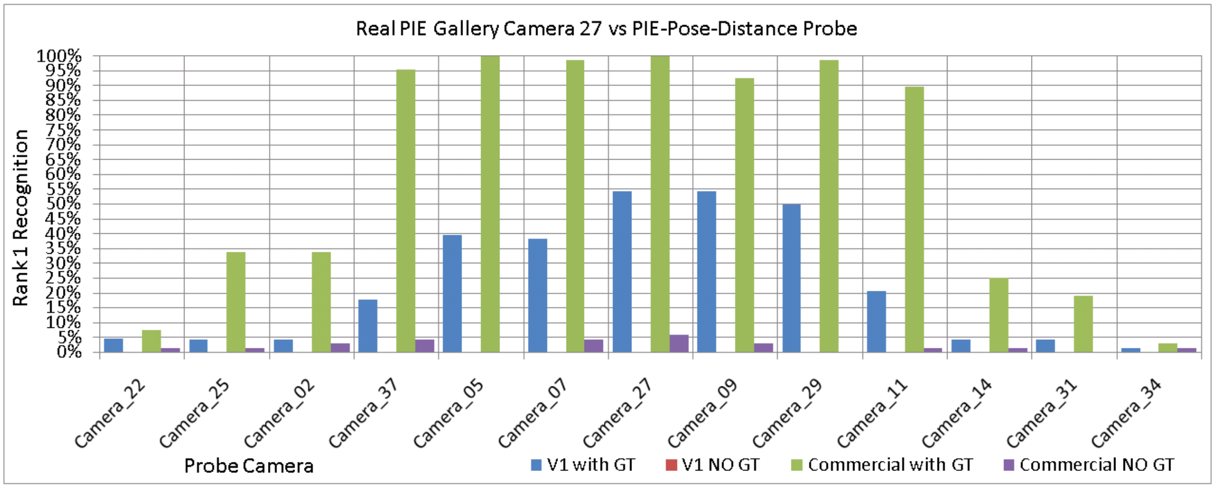

Fig. 7. This graph shows the results of both the "V1-like" and commercial algorithms when using an image with no pose variation as the gallery. Again, the commercial algorithm outperforms the "V1-like" core when given ground-truth eye coordinates. However, when given no ground-truth at all both algorithms fail. The "V1-like" core cannot recognize a single image and the commercial algorithm fails to get above 5 percent except on the frontal pose image where even then it achieves only 5.88% recognition.

models, again to individuals that have licenses to the underlying data. For further details please contact one of the authors.

VI. ACKNOWLEDGMENTS

This work was supported by ONR STTR N00014-07-M-0421 and ONR MURI N00014-08-1-0638 and NSF PFI 650251. We would also like to thank: Justin Schiff for his work on the photohead 3D display software. Dana Scheirer for model generation. Chris Olsowka and Jon Parris for model generation, ground truth and data captures. Anthony Magee and Brian Parks for their work on the blur setup and data captures.

REFERENCES

[1] 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.

[2] 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] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] 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.

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

[6] M. Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE T. on Acoustics, Speech and Signal Processing*, 24(1):58–63, 1976.

[7] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 277, Washington, DC, USA, 2000. IEEE Computer Society.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Proceedings of the IEEE F&G*, Sept. 2008.

[9] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, December 2001.

[10] Y. Hu, Z. Zhang, X. Xu, Y. Fu, and T. S. Huang. Building large scale 3d face database for face analysis. In *MCAM'07: Proceedings of the 2007 international conference on Multimedia content analysis and mining*, pages 343–350, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] V. N. Iyer, S. R. Kirkbride, B. C. Parks, W. J. Scheirer, and T. E. Boult. A taxonomy of face-models for system evaluation. In *To be published in AMFG 2010: Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*. IEEE Computer Society, 2010.

[13] B. Kroon, A. Hanjalic, and S. Boughorbel. Comparison of face matching techniques under pose variation. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 272–279, New York, NY, USA, 2007. ACM.

[14] M. D. Levine and Y. Yu. Face recognition subject to variations in facial expression, illumination and pose using correlation filters. *Computer Vision and Image Understanding*, 104(1):1 – 15, 2006.

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

[16] 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.

[17] 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.

[18] W. Scheirer, A. Rocha, B. Heflin, and T. Boult. Difficult detection: A comparison of two different approaches to eye detection for unconstrained environments. pages 1–8, 2009.

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

[20] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recogn.*, 39(9):1725–1745, 2006.

[21] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[22] 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.

[23] 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.

[24] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recogn.*, 42(11):2876–2896, 2009.

[25] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.