https://doi.org/10.2352/ISSN.2470-1173.2019.13.COIMG-140

Through the Windshield Driver Recognition

***David Cornett III*[*], *Alec Yen*[**], *Grace Nayola*[***], *Diane Montez*[***], *Christi R. Johnson*[*], *Seth T. Baird*[*], *Hector Santos-Villalobos*[*], *David S. Bolme*[*]**

Oak Ridge National Laboratory; Oak Ridge, TN, USA; **University of Tennessee; Knoxville, TN, USA; ***Texas A&M University-Kingsville, Del Mar College; Kingsville, TX, USA

Abstract

Biometric recognition of vehicle occupants in unconstrained environments is rife with a host of challenges. In particular, the complications arising from imaging through vehicle windshields provide a significant hurdle. Distance to target, glare, poor lighting, head pose of occupants, and speed of vehicle are some of the challenges. We explore the construction of a multi-unit computational camera system to mitigate these challenges in order to obtain accurate and consistent face recognition results. This paper documents the hardware components and software design of the computational imaging system. Also, we document the use of Region-based Convolutional Neural Network (RCNN) for face detection and Generative Adversarial Network (GAN) for machine learning-inspired High Dynamic Range Imaging, artifact removal, and image fusion.

Introduction

Imaging through the windshield presents some unique challenges. Face images are captured at significant standoff distances, which reduces the amount of available light. This challenge is compounded by attenuation from windshield coatings, including tints that block Ultraviolet (UV) and Near Infrared (NIR) wavelengths and the required short exposure times for motion blur reduction at even low vehicle speeds of 15 MPH. The windshield also produces specular highlights that must be removed for unobstructed line of sight between the camera and the subject's face. In addition, environmental and behavioral factors such as structural shadows and face pose, respectively, tend to degrade the performance of face recognition systems.

There are current solutions for capturing images through windshields. The most common are automatic traffic cameras that include high power strobes, such as the Gatekeeper system [7], which operates using red or NIR illuminates. These systems all produce significant flashes that distract drivers, especially at night, and capture faces from a single camera point of view, which limits the system's ability to overcome occlusions, illumination fluctuations, and off-angle face poses.

We are developing a face recognition system designed to withstand harsh environments as well as high throughput for computer-aided site access control. The system is operated with passive illumination during daylight, and it is modular and extensible. The current prototype consists of two computational imaging units with three cameras each. The cameras include particular optics to address the challenges stated before. Figure 1 shows a picture of one of the units. This paper contributes with a deep learning-based reconstruction model that simultaneously performs High Dynamic Range (HDR) fusion and artifact correction. In particular, our solution is inspired by the latest work on GANs. The network models were effectively trained on data sets from a different domain. The proposed optical and machine learning-based solution increases the likelihood of consistent collection of high quality frontal face images.

The paper is organized as follows: First, we discuss the imaging challenges that drive the system design and specify the components of each computational unit. Second, details are provided about the software architecture and its components, which include a documentation of algorithms for image registration, HDR, artifact correction, and face detection and recognition. Finally, we conclude the paper with experimental through-the-windshield face recognition results and future directions.

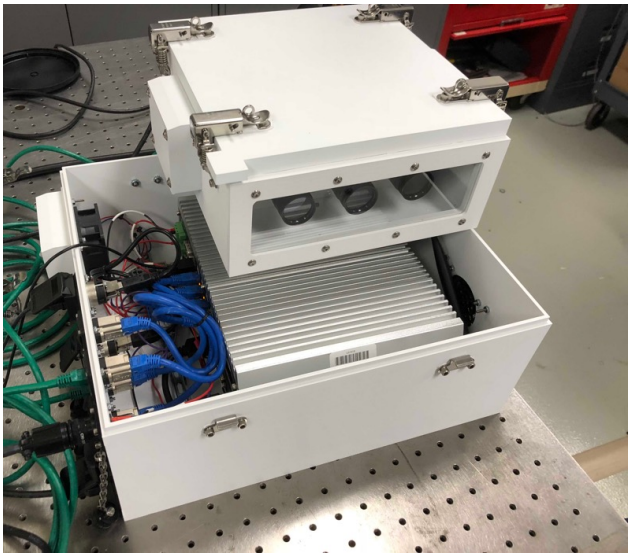


Figure 1: Computational imaging unit, which includes two weatherproof enclosures; one for a camera array and the other for a ruggedized computer.

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Figure 2: Images from all six cameras showing the left and right views of the vehicle.

Imaging Unit Design and Hardware Components

Imaging passengers through vehicle windshields presents a host of challenges for providing high quality templates for facial recognition. The system under development aims to address the following issues: glare and reflections at the windshield from uncontrolled illumination, low contrast/illumination at the vehicle cabin, motion blur, out of focus blur, and head pose variation.

To mitigate image degradation caused by windshield glare and reflections, we employed linear polarization filters. The electromagnetic polarization of reflections and glare is biased towards the incidence angle of the last reflection surface. For example, the road is approximately horizontal or zero degrees. Therefore, road glare is mostly composed of zero-degree-polarized light waves. Consequently, filtering such light artifacts with a 90-degree polarization filter reduces road glare. However, windshields are mounted at different angles. Thus, polarization filters were placed in all cameras at different polarization angles in the range of $90° \pm 10°$.

Contrast at the vehicle cabin is driven by the amount of ambient light (e.g., sun light) and the shadows generated by adjacent structures and vehicle components, such as trees and the vehicle roof, respectively. In addition, these light conditions can change instantaneously (e.g., sudden cloud coverage). Commercially available cameras can adapt exposure and gain to global light conditions. However, automated camera gain and exposure adaptation for moving targets is challenging. Consequently, HDR techniques were employed to guarantee adequate dynamic range for faces inside the field of view independently of changing lighting conditions. Unlike conventional HDR imaging, where a single camera takes a rapid succession of images at different exposure times (i.e., bracketing), the system under development emulates bracketing with an array of cameras with varied neutral density filters. Note that the traditional HDR approach introduces ghosting artifacts due to the movement of the vehicle through the frames. As shown in Figure 2, camera #3 does not have a Neutral Density (ND) filter for a 100% transmission of light into the sensor, while camera #2 and #1 use a 0.3 and 0.9 ND filter, respectively, for a 50% and 25% light transmission.

Motion and out-of-focus blur are controlled by different design parameters. Additionally, the imaging units need to capture a variety of vehicle sizes ranging from compact cars to service trucks and the face images require 100 pixels between the eyes which provides enough resolution for accurate recognition. We

Figure 3: Illustration of the computational unit camera array depth of field.

found that a 10 meter trigger distance on average meets these requirements. However, an extended depth of field is desired in order to reduce out-of-focus faces. Consequently, as shown in Figure 3, the middle camera is focused at a target plane 9 meters away (immediately following the trigger), while the other cameras are focused at distances of 8 and 7 meters to extend range where the driver is in focus. For the selected distance, depth field, and average vehicle speed of 10 MPH, an exposure time of 10 ms provided the optimal trade off between motion blur and contrast.

The two-unit system was key at increasing the likelihood of capturing a frontal face image. As shown in Figure 4, two computational units are placed at each side of the lane and as close as possible to the road to get a near frontal windshield image. Each imaging unit is triggered by the same signal and frame rate acquisition is also coordinated. Therefore, at time t two simultaneous views of the vehicle and passengers are obtained. As the vehicle moves across the field of view, the driver may change face pose to track other vehicles or due to other distractions. Our dual system takes advantage of these movements and near to frontal faces are expected in more than one frame. Note that an overhead capture is not ideal because it reduces the available capture window and vehicle components tend to occlude the faces.

System Hardware

The weatherized prototype system contains independent, modular imaging units that serve up their data to a control program that manages and performs image fusion, detection, and recognition processes. These units consists of an array of three Basler GigE cameras mounted to a custom 3d-printed bracket, which aligns them to a common focal point at 10 meters. Although individual camera exposure time variation is controlled with ND filters, global gain and exposure time is adjusted at the time of data acquisition based on that day's expected illumination conditions. To mitigate overheating of components, each computational imaging unit has a cooling fan for air cycling and each Basler camera is outfitted with a heat sink. Each imaging unit transmits raw images via Power over Ethernet (PoE) connections to a ruggedized and weatherized computing node. The computing node contains four power over Ethernet (PoE) ports, 32GB of memory, an Intel Core i7 CPU with 8 cores, and a NVIDIA GTX1050 Ti Graphics Processing Unit (GPU) with 4GB of memory. The GPU was selected to support our deep learning-heavy algorithms. Face detection was found to benefit the most from the GPU. The current face recognition algorithm process an entire 3MP image in 27 and 0.17 seconds for Central Processing Unit (CPU) and GPU,respectively. The GPU provides a 158x speedup to face detection.

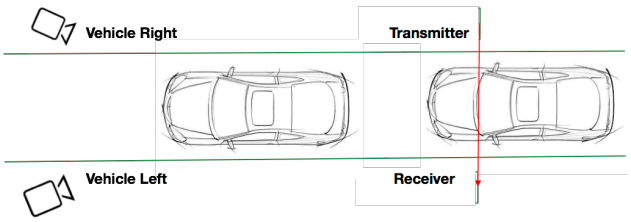


Figure 4: Two-unit camera setup.

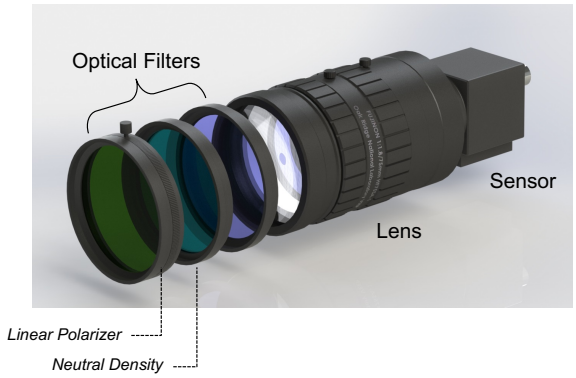


Figure 5: Cameras and filters used to penetrate windshield glare.

For frame synchronization the same trigger signal was distributed across units and their cameras. The triggering system consists of an Optex through beam photoelectric sensors placed a couple meters behind the middle camera focus plane (See Figure 4). After this initial triggering event, each camera would then acquire twenty frames equally spaced in sequence. The frames were then stored with syncing metadata (e.g., time stamp, event number, frame number, etc.). This metadata is critical for proper image registration and fusion.

Based on the specifications and design selections stated above, the following are our optical components: 1) a Basler camera with 2048 x 1536 pixel matrix and 3.45 μm pixel size, 2) a 50 mm focal length and 34 mm aperture lens, 3) a Midwest Optical Systems UVIR filter to sample in the visible spectrum, 4) a Midwest Optical Systems ND filter, and 5) a Meadowlark Optics linear polarizer with extinction ratio greater than 10,000,000:1. Figure 5 illustrates the optical configuration of the cameras.

System Software Architecture

The modular design concept used for system hardware components was continued at the software level. At its core, Google Remote Procedure Call (gRPC) was employed to separate and manage communications between software processes. The gRPC works as remote servers. For example, a camera server handled camera triggering and raw image acquisition, and processes such as face extraction, HDR fusion, and face template matching were also designed to execute independently. This independent server architecture allows for easy system upgrades and seamless multithreading. The gRPC architecture allows upgrade of low level algorithms without the need of significant code re-factoring and the risk ok breaking other application processes. The QT Software Development Toolkit version 5 [4] was selected to implement the Graphic User Interface (GUI) for the application.

The following subsections describe the core processes for image acquisition and facial recognition.

Image Acquisition

The image acquisition module coordinates the communication with the cameras and receives raw images every time the cameras are triggered. For each trigger event and imaging unit, a three-raw-image set is generated. After a trigger event, each camera module would take a sequence of twenty frames, which results in 120 total raw images or 40 three-raw-image sets. The

image frames are pushed to a memory queue data structure until computational resources are available for further processing. Each image frame is tagged with metadata, such as camera module number, event number, frame id, and timestamp. Each three-raw-image set is passed to the next processing step together.

Initial Alignment

During imaging unit setup, a calibration image is taken to estimate a rigid transformation that will coarsely align the images in a three-raw-image set–with the middle camera as the reference. The quality of the alignment is confirmed by visual inspection of the resulting alignment of several three-raw-image sets. A close spatial alignment of the sets confirms that the cameras are also temporally synchronized. Although initial calibration can be performed without specialized physical calibration tools, empirical experience dictates that use of high contrast patterns, such as a checkerboard patter, facilitates and increases the accuracy of the coarse spatial alignment. A software tool was also developed to refine some camera parameters, e.g., gain and exposure time, to the current environmental conditions. The subsequent image processing routines depend on a proper coarse alignment of each three-raw-image set.

Face Detection

Face detection occurs after the initial coarse alignment and before the HDR steps. Face detection is significantly less computational expensive than full-frame HDR. Also, the face detector is configured to detect even low contrast and noisy faces. Consequently, we can save computational resources by only performing HDR on Regions of Interest (ROIs) with faces.

The detection process starts with computing the Region of Interest (ROI) coordinates for the faces in the frame of the camera without a ND filter. Given that the three-raw-image set is coarsely aligned, we extract the same ROIs for the other frames in the set. This step extracts three different optical (e.g., exposure) versions of a face. Each face image is 256x256 pixels in size. This face set is known as the three-raw-face set. Note that each three-raw-image set may contain several three-raw-face sets. Each three-raw-face set is post-processed for alignment refinement and HDR fusion.

Adaptive Alignment

The coarse alignment registers the whole images so that the three-raw-face set fully contain the same face in all three images. However, due to lens distortions and inaccuracy of the coarse rigid transformation at depths that depart from the depth used during camera setup, the coarse alignment is not sufficient enough for HDR fusion. Consequently, the alignment of the three-raw-face image set needs refinement. The initial coarse alignment was refined with a median threshold bitmap technique [14] and a scale-invariant Fast Fourier Transform (FFT) registration technique [13]. From these two techniques, the FFT-based technique consistently produced accurate alignments.

High Dynamic Range Fusion

HDR fusion is an example of a computational imaging technique. In computational imaging, there are physics-based, data-based, or statistical models of the world, the object, and/or the instrument that "sees" the world (i.e., measurements) that are used to indirectly reproduce the object of interest (e.g., image) from the measurements. In HDR the radiance of the scene is sampled by acquiring images at different radiance dynamic ranges. These separate images are fused together following a radiance model for each frame camera configuration (e.g., exposure time, f#, density filter). In this paper, we compare physics-and data-based models for HDR. For the physics-based model we employed two algorithms for exposure fusion. The first approach is inspired on Debevec radiance mapping [5] and Durand[6] tone mapping algorithms. The second approach is an implementation of Mertens et al. [12] exposure fusion technique. For a data-based HDR approach, a GAN was trained to both fuse images with different exposures, and to correct for the expected artifacts in through-the-windshield face images.

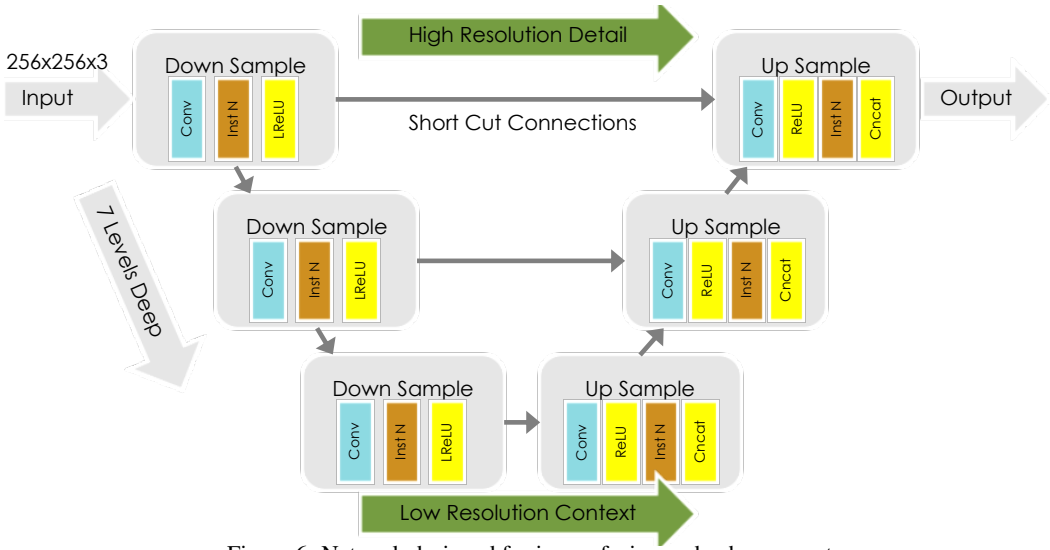


Figure 6: Network designed for image fusion and enhancement.

Design and Training of GAN for Through-the-Windshield Image Fusion and Artifact Removal

One of the contributions of our work is the use of GANs to perform HDR processing and image enhancement across the three-raw-face image set. Figure 6 shows the generator module for our GAN. The GAN is implemented in Keras and is based on tutorials provided from [10], which was configured for super resolution and based on the Wasserstein GAN implementation [1].

For our purposes we have made significant modifications to the network to make it more consistent with the implementation described in [9]. The original network transformed the image from 64x64x3 pixels to 256x256x3 pixels. As shown in Figure 6, this has been changed to produce a 256x256x3 to 256x256x3 transformation which is roughly the size of the face when vehicle images are captured at 10 meters from the camera.

The network consists of 7 down-sampling steps and then 7 up-sampling steps. Short cut connections bridge the layers at the same size. This has been found to provide enough depth to perform complex face analysis operations at the low resolution layers while providing shortcuts to transmit higher frequency information and facial details for accurate face reconstruction.

The original super-resolution version of the network applied a down-sampling to a high-resolution face and then trained the network to reconstruct the original image. Our network's intent is for the purpose of HDR processing, however, as stated previously, there are a variety of other image artifacts that degrade image quality, such as glare, noise, blur, alignment, lighting, shadows. For this reason, we pushed the capabilities of the network to fuse different exposures and correct for these artifacts.

The network is trained using the Celeb A [11] dataset, where faces larger than 256x256 have been extracted. The images are corrupted with synthetic artifacts that resemble those found in the through-the-windshield face recognition application. The network goal is to receive three low quality face images and reconstruct the original high-quality face image. The original Celeb A images are RGB color images. Therefore, the network attempts to predict skin color also. In addition, more than one synthetic artifact can be applied to each input image. The simulated artifacts are:

- A random rigid transformation (e.g., rotation, scale, and translation) is applied to the image.
- A grayscale version of the image is created by randomly weighting the original image.
- The image is multiplied by a randomized camouflage image to simulate random shadows or glare on the windshield and to encourage the network to mitigate non-face artifacts.
- The image pixel values are scaled by three randomly selected values to approximate the ND filters on the cameras. The order of these images into the input channels is randomized.
- Small translations of three pixels are applied to two of the images. The "center" image is kept aligned to the original.
- The images are blurred using a random Gaussian filter with a sigma value centered at 2.0 and a standard deviation of 2.0.
- Randomized noise is generated independently for each image channel. It is also blurred using random Gaussian filters and applied to the images.

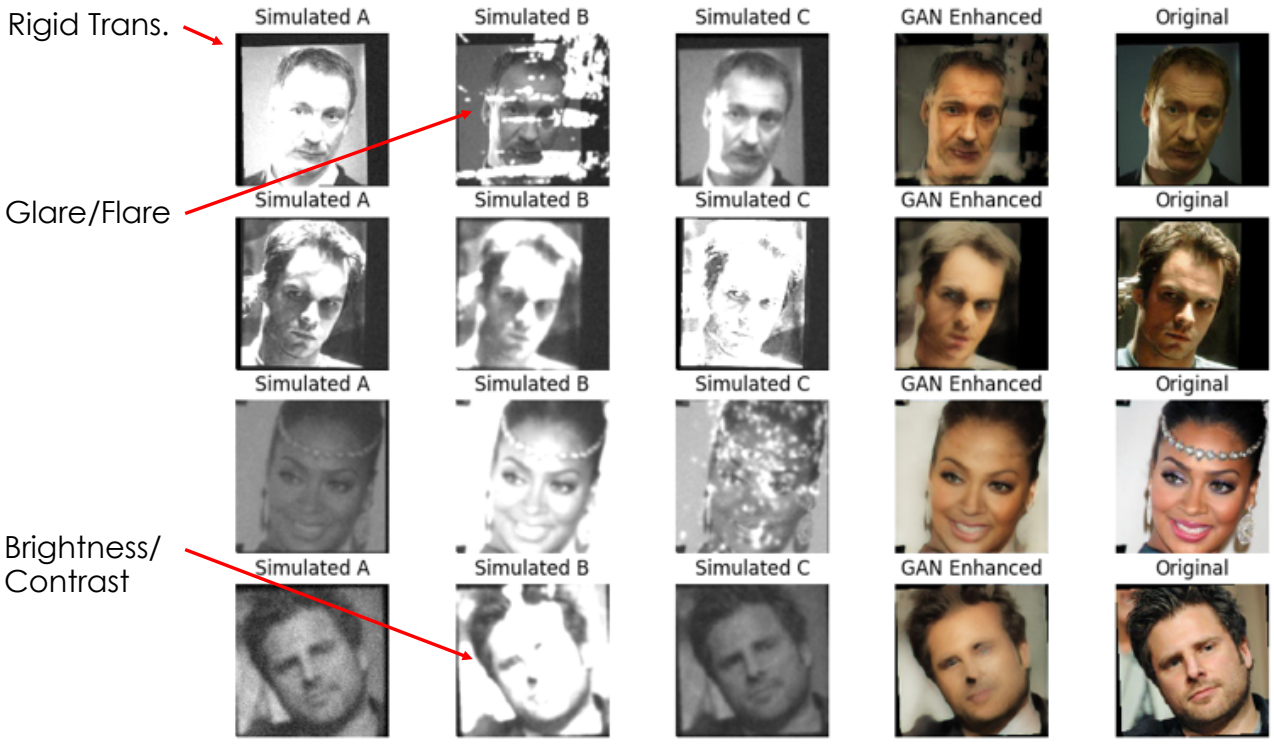


Figure 7: Celeb A examples of GAN (first three columns) inputs, (4th column) reconstruction/fusion, and (last column) original images. Note that the first three columns show examples of the GAN training data and synthetic artifacts.

Experimental Setup and Results

After the original image is corrupted, the pixel values are clipped to the range 0 to 255 to maintain the original image pixel value range. The original image and corrupted input are both re-scaled to the range -1 to 1 before being presented to the network. Examples of the generated dataset are shown in Figure 7. Note that the network attempts to perform a combination of corrections and processes, including alignment, HDR processing, artifact reduction, noise reduction, super resolution, and colorization (since it attempts to reconstruct the original 3-channel color image).

Two different network models were trained. The first model (GAN1) used approximately 2,500 iterations. This model includes a perceptual loss [9] function. The loss function was based on the first 9 layers of a VGG19 network using ImageNet weights from Keras. The second model (GAN2) used 2,500 iterations. It included the perceptual loss of the GAN1 model and also included a face recognition model loss based on the first 9 layers of the VGG Resnet50 face model. The intention for GAN2 is to produce reconstructions that are improved for face recognition.

An initial prototype experiment was staged at a closed traffic loop with vehicles that ranged both in size and windshield variety. Participants passed through the system numerous times to produce an initial test data-set consisting of 2,949 raw images containing faces. Additional variability was added to the data-set with natural lighting conditions ranging from overcast to very bright, presence of rain, various head poses and face obstructions (sun-

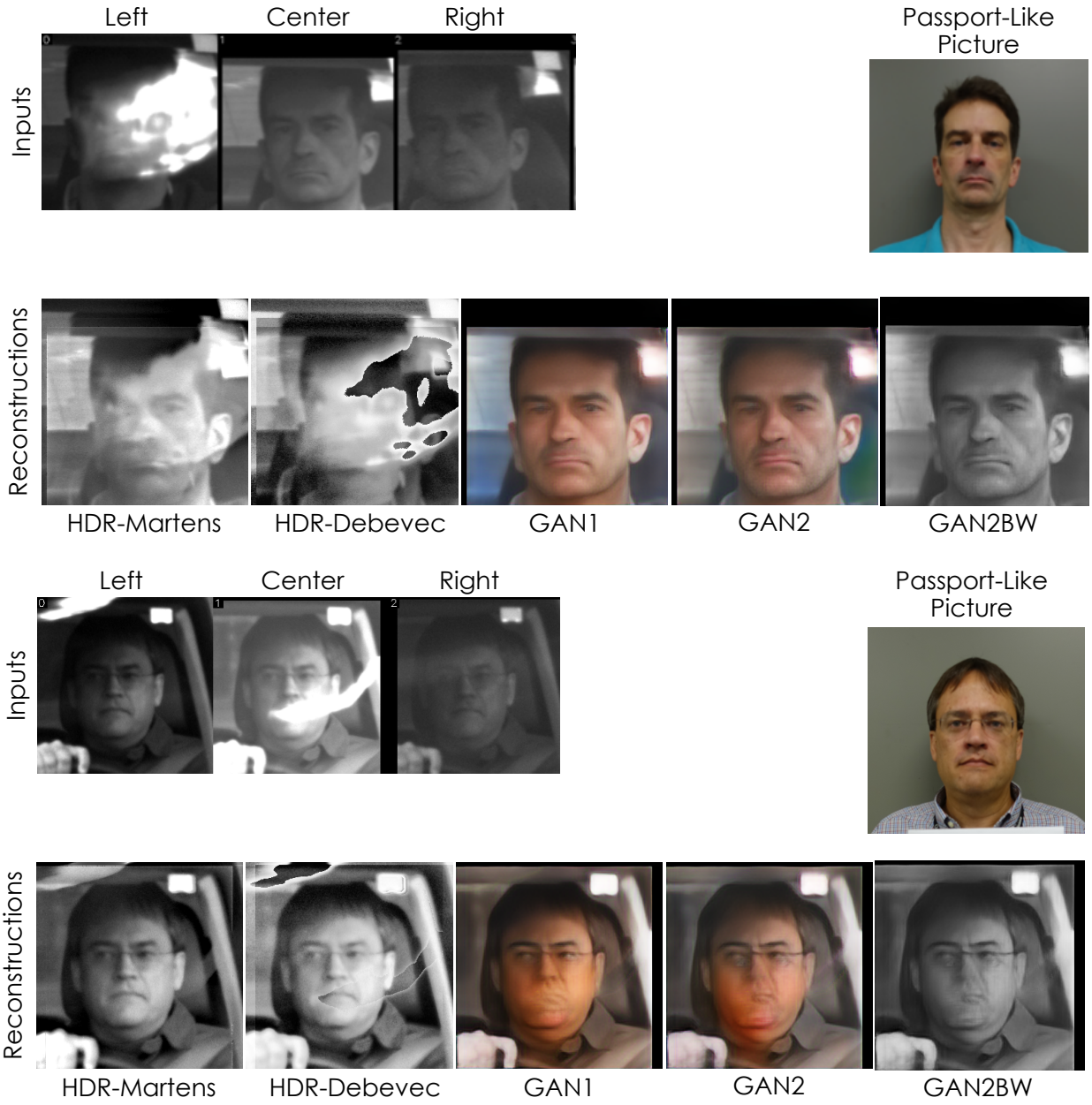


Figure 8: Example reconstructions from the vehicle driving test.

glasses, etc.), operating windshield wipers and varying speeds. Drivers alternated between starting from a dead stop before the trigger and accelerating to 15mph as well as driving through the system maintaining a speed of 15mph. These speed variations faithfully approximate the targeted driving conditions occurring at security portals.

Continuous daily use in direct sunlight with daytime temperatures above 80F coupled with intermittent rain events provided ample validation of system ruggedness. System setup in the field also predicated the need for additional convenience features (such as laser leveling) to quickly orient the system components and provide indication of optimal performance.

Given the unconstrained nature of through-the-windshield face recognition, the crux of facial recognition relies on the accurate detection of faces. We used an open source RCNN[15] that has been adapted for face recognition by Jiang and Learned-Miller [8]. In the initial prototype experiment the face detector was able to detect driver and passenger faces at a rate of 97.15%. There were 84 instances of missed faces - the vast majority of these due to face occlusion from cups, visors, windshield wipers, and/or a significantly dirty windshield. Additionally, there were only 3 events when a driver was able to pass through the system without a single face detection across the event frame set. Therefore, this high detection rate should greatly boost the accuracy of recognition efforts in later experiments. Face recognition results were not calculated for this initial experiment.

A second two-day experiment was conducted at a facility security portal to provide live field test performance metrics. The experiment included 22 enrolled participants who entered the facility under standard security procedures as the system actively captured images for identification. Images were gathered at this portal from before sunrise till early afternoon providing for various illumination conditions in which to capture driver images. Approximately 56,520 raw images were captured from the two systems from 471 vehicle triggered events. Of these, the captured participant images were then merged to produce the two-GAN variant images as well as with standard Debevec[5] and Mertens[12] algorithms. These processed images, along with each of the raw images from the three system cameras, were then used as probes in the recognition system. Figure 8 shows empirical examples of input through-the-windshield face images, a passport-like picture (i.e., gallery image), and the reconstructions from our physics-and data-based models. Note that GAN2BW is the same network as GAN2, but the output image is mapped to a monochrome space.

To generate features for recognition, we have used the VGG2 RESNET model [3]. The last layer of that classification network was removed, leaving a 4096 element vector for matching. Scores were produced using correlation, where higher correlations are assumed to be better matches. Figure 9 shows the scores for each method and illustrates that GAN2 which optimizes for recognition does have a small boost in the mean match score. However, the traditional Mertens HDR method [12] still produces best results. The face recognition configuration used for these experiments will be available so that others can reproduce results [2]

One interesting artifact of our GAN implementation is that the network output (i.e., reconstruction) is biased towards image input #2. This is illustrated in Figure 10, where the ordering of inputs are swapped to produce different reconstructions. It was found that if the best image was provided to input #2 reconstructions always seemed to be better. We attempted to mitigate input bias by randomly ordering the training images with the expectation that the network would learn to evaluate quality and base reconstructions on the best inputs available. The artifact is most likely a result of how the network converged during training.

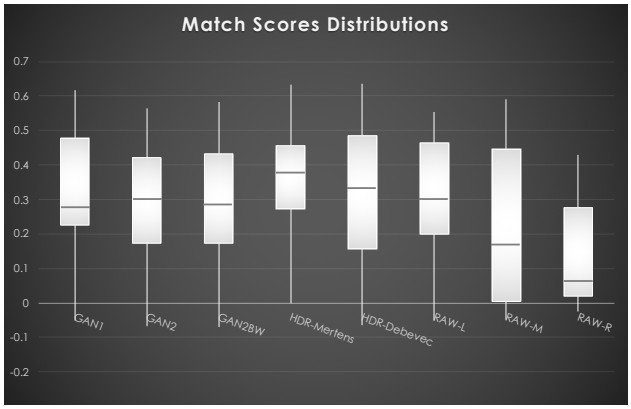


Figure 9: Match score distributions sorted by method used to produce the probe image for comparison to the ground truth gallery. GAN1 was designed to produce images for perceptual quality whereas GAN2 and GAN2BW were trained for perception and recognition accuracy.

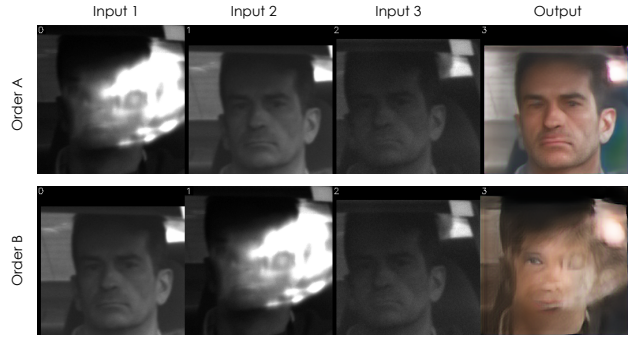


Figure 10: The ordering of the inputs presented matters and can have significant effects on reconstruction quality

Conclusion

Preliminary field tests of the prototype through-windshield face recognition system have demonstrated that the system has the potential to meet desired timing, accuracy, and durability constraints. The efforts of the HDR GAN design has shown to be particularly effective in reconstructing high quality HDR images from the raw exposures that can be used for facial recognition.

Additional plans to enhance the overall system performance and reliability include:

- Revisiting outdoor alignment procedure
- Setup and calibration improvements
- Modifications to user GUI design
- Outdoor recognition testing
- Separate HDR fusion from artifact correction

Acknowledgements

This research was supported in part by an appointment to the Oak Ridge National Laboratory Post-Bachelors Research Associate Program, sponsored by the U.S. Department of Energy and administered by the Oak Ridge Institute for Science and Education, the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI), and the Texas A&M University-Kingsville, Office of University Programs, Science and Technology Directorate, Department of Homeland Security Grant Award # 2012-ST -062-000054.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] David Bolme and David Cornett III. Faro: FAce Recognition from Oak Ridge. https://github.com/ORNL/faro, 2019.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[4] The Qt Company. Qt5 software development toolkit. https://www.qt.io/, 2019.

[5] P Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. *Proceedings OF ACM SIGGRAPH*, pages 369–378, 1997.

[6] F. Durand and J. Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics*, 21(3):257–266, 2002.

[7] Dave Harmon. Facial recognition and registration plate reading, 2017.

[8] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 650–657. IEEE, 2017.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2 (3), page 4, 2017.

[10] Erik Linder-Norn. Keras implementations of generative adversarial networks. https://github.com/eriklindernoren/Keras-GAN, 2018.

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[12] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Computer Graphics and Applications, 2007. PG'07. 15th Pacific Conference on*, pages 382–390. IEEE, 2007.

[13] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, Aug 1996.

[14] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski. *High Dynamic Range Imaging*, volume 2nd. Morgan Kaufmann Publishers Inc, 2010.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

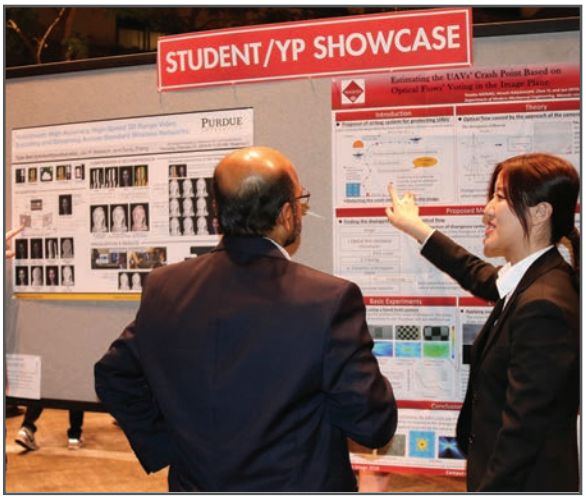
JOIN US AT THE NEXT EI!

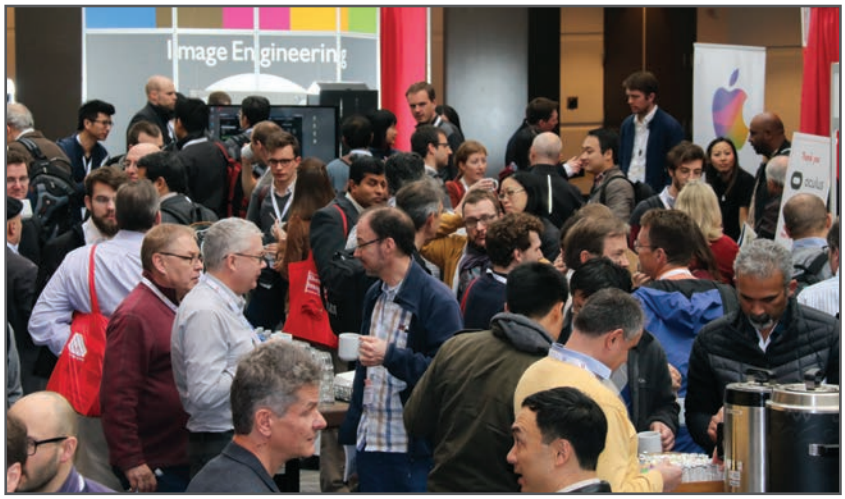
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org