Report #23-UT-129

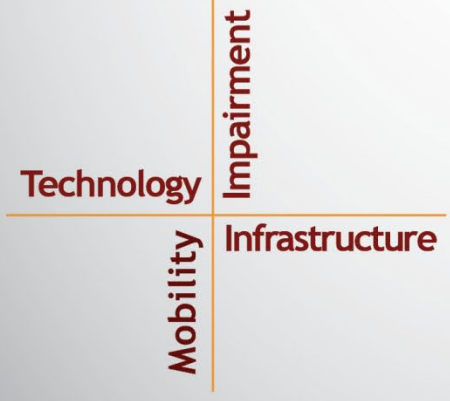
NSTSCE

National Surface Transportation Safety Center for Excellence

Face De-identification of Drivers from NDS Data and Its Effectiveness in Human Factors

Surendrabikram Thapa, Julie Cook, Abhijit Sarkar

Submitted: August 8, 2023



Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Zac Doerzaph from the Virginia Tech Transportation Institute; John Capp from General Motors Corporation; Terri Hallquist from the Federal Motor Carrier Safety Administration; Mike Fontaine from the Virginia Department of Transportation and the Virginia Transportation Research Council; and Melissa Miles from State Farm.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

Advancements in artificial intelligence (AI) and the Internet of Things (IoT) have made data the foundation for most technological innovations. As we embark on the era of big data analysis, broader access to quality data is essential for consistent advancement in research. Therefore, data sharing has become increasingly important in all fields, including transportation safety research. Data sharing can accelerate research by providing access to more data and the ability to replicate experiments and validate results. However, data sharing also poses challenges, such as the need to protect the privacy of research participants and address ethical and safety concerns when data contains personally identifiable information (PII).

This report mainly focuses on the problem of sharing drivers' face videos for transportation research. Driver video collected either through naturalistic driving studies (NDS) or simulator-based experiments contains useful information for driver behavior and human factors research. The report first gives an overview of the multitude of problems that are associated with sharing driver videos. Then, it demonstrates possible directions for data sharing by de-identifying drivers' faces using AI-based techniques. We have demonstrated that recent developments in generative adversarial networks (GANs) can effectively help in de-identifying a person by swapping their face with that of another person. The results achieved through the proposed techniques were then evaluated qualitatively and quantitatively to prove the validity of such a system. Specifically, the report demonstrates how face-swapping algorithms can effectively de-identify faces while still preserving important attributes related to human factors research, including eye movements, head movements, and mouth movements.

The experiments were done to assess the validity of GAN-based face de-identification on faces with varied anthropometric measurements. The participants used in the data had varied physical features as well. The dataset used was under lighting conditions that varied from normal to extreme conditions. This helped to check the robustness of the GAN-based techniques. The experiment was carried out for over 115,000 frames to account for most naturalistic driving conditions. Error metrics for head movements like differences in roll angle, pitch angle, and yaw angle were calculated. Similarly, the errors in eye aspect ratio, lip aspect ratio, and pupil circularity were also calculated as they are important in the assessment of various secondary behaviors of drivers while driving. We also calculated errors to assess the de-identified and original pairs more quantitatively.

Next, we showed that a face can be swapped with faces that are artificially generated. We used GAN-based techniques to generate faces that were not present in the dataset used for training the model and were not known to exist before the generation process. These faces were then used for swapping with the original faces in our experiments. This gives researchers additional flexibility in choosing the type of face they want to swap. The report concludes by discussing possible measures to share such de-identified videos with the greater research community. Data sharing across disciplines helps to build collaboration and advance research, but it is important to ensure that ethical and safety concerns are addressed when data contains PII. The proposed techniques in this report provide a way to share driver face videos while still protecting the privacy of research participants; however, we recommend that such sharing should still be done under proper guidance from institutional review boards and should have a proper data use license.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

ADAS	advanced driver assistance systems
AI	artificial intelligence
CJWR	cheek-to-jaw width ratio
CV	computer vision
EAR	eye aspect ratio
EHWR	eye height-width ratio
ERGAS	Relative Dimensionless Global Error Synthesis
FWHR	face width height ratio
GAN	generative adversarial network
LAR	lip aspect ratio
MAE	mean absolute error
MSE	mean squared error
NAS	National Academy of Science
NDS	naturalistic driving study
ORNL	Oak Ridge National Laboratory
PERCLOS	percentage of eye closure
PII	personally identifiable information
PSNR	peak signal-to-noise ratio
PUC	pupil circularity
RMSE	root mean squared error
SAM	spectral angle mapper
SHRP 2	Second Strategic Highway Research Program
UIQI	universal image quality index
VAE	variational auto encoders

CHAPTER 1. INTRODUCTION

Over the last decade, naturalistic driving studies (NDS) have been pivotal in answering key questions in transportation safety, operation, and mobility. Analysis of drivers' behavior from their face videos has shown how secondary behavior (e.g., cell phone use), distraction, and drowsiness play roles in the occurrence of safety-critical events. While face videos of drivers contain valuable information for human factors and safety research (e.g., driver gaze, percent eye closure [PERCLOS], and active head movements), their wide-scale distribution and scope of research are restricted due to the personally identifiable information (PII) that they carry. To eliminate such restrictions, we need to create de-identified videos of naturalistic drivers with two key considerations: (1) all PII information will be eliminated, and (2) all human factor cues will be preserved. This project explores how to achieve such a goal by leveraging recent developments in computer vision (CV).

This project consisted of three major tasks:

Task 1: Understanding PII information and human factors attributes. In this task, we listed a set of PII attributes present in the face videos and selected a set to be addressed. For example, if a tattoo is considered PII, it will not be considered part of this project. Next, we summarized the human factors attributes or behavioral attributes that are traditionally used in transportation research. Examples could be PERCLOS, head pose, and movement of the mouth.

Task 2: Test off-the-shelf algorithms to create de-identified videos under different operating conditions. Recent advances in variational autoencoder (VAE) and generative adversarial networks (GANs), variants of deep neural networks, have enabled algorithms like DeepFake (Westerlund, 2019) that can take videos of a certain person and replace the identity of that person with a second person while keeping their facial attributes and actions. Figure 1 shows an example where the identity of the person in the leftmost column is replaced frame-by-frame in their original video by each of the personalities from the top row. In the process, only a single image of the imposter person is provided, and the algorithm preserves the facial attributes of the original videos, including *blinking of the eyes*, *head pose*, and *mouth movement*. Recently, we have worked with NDS videos from the Virginia Tech Transportation Institute's NextGen, MiniDAS, and FlexDAS data acquisition systems and developed and tested a series of algorithms that can (a) detect driver faces, (b) identify facial key points (e.g., corner of eyes, mouth), and (c) create a 3D shape model of the driver's face (see Figure 2a). The results demonstrate the future prospects of face swapping in addressing privacy concerns.

This task also included the following subtasks: (a) detailed literature survey of multiple DeepFake/face-swap algorithms; (b) application and testing of selected DeepFake algorithms to generate de-identified face videos; and (c) selection of an appropriate face for the face-swap. We also explored using artificially generated faces (Karras et al., 2020) that are close to the appearance of the driver (see Figure 2b).

Task 3: Use manual reduction and computer-based techniques to test the effectiveness of human-factors-related attributes, including PERCLOS analysis, in de-identified face videos. Once the de-identified videos were created, we needed to confirm that the output video did not contain any PII. Manual reduction was performed on a set of de-identified videos. The reduction team

manually annotated PERCLOS for a set of videos before and after the de-identification was performed. We compared the results to check effectiveness. This task also included discussing the challenges and limitations of such algorithms and their applicability under diverse conditions.

Figure 1. Illustration. Example of DeepFake where the personality of the left column is replaced by personalities of the top row (Siarohin, Lathuilière, Tulyakov, Ricci, & Sebe, 2019).

(a) (b)

Figure 2. Illustration. (a) Example of 3D face mask on NDS data; (b) example images of faces generated using GAN-based techniques and were not present in the dataset used for training the model.[1]

REPORT LAYOUT

In this report, we first explore PII and standard methods and norms to eliminate PII for extended sharing. Chapter 2 provides an overview of data sharing practices. Chapter 3 provides a brief background on CV approaches for face de-identification and general data protection practices. Chapter 4 gives an outline of the methods used in this project. Similarly, Chapter 5 highlights the importance of various behavioral attributes. Chapter 6 provides a detailed analysis of the results. Finally, Chapter 7 concludes with suggestions on how data sharing can be made effective with face de-identification. It also includes considerations about sharing data and discusses the limitations of our work.

[1] https://www.wired.com/story/artificial-intelligence-fake-fakes/

CHAPTER 2. PROTECTING PII

In transportation safety research that involves machine learning or data-driven research, reproducibility is a critical aspect that helps in facilitating the reliability and transparency of scientific findings. For research to be considered as "good research," it must have basic traits of correctness, novelty, sound methodology, reproducibility, and transparency (Burrell & Toyama, 2009). Data sharing provides the ability to validate the reproducibility, transparency, and correctness of research. Apart from these merits, data sharing also helps researchers to optimize time and resources. When data originates from human subjects, sharing the data also provides the potential for larger gains and benefits resulting from the participants' time and effort (Myer, 2018).

However, when research involves data from human subjects, great care must be taken to respect the research participants involved and protect their privacy and dignity. Respecting participants involves obtaining informed consent to participate in the original research effort, as well as obtaining consent regarding how their data may be used in the future. U.S. Department of Health and Human Services regulations detail a comprehensive set of protections for research subjects. These regulations (Federal Policy for the Protection of Human Subjects 2018, 45 CFR part 46 subpart A) are also known as the Common Rule. The Common Rule states that one criterion for institutional review board approval is "when appropriate there are adequate provisions to protect the privacy of subjects and to maintain confidentiality of the data." The benefits of sharing the data do not outweigh the importance of the privacy provisions. Researchers who wish to share data must consider the benefits of doing so along with a thorough examination of promises that were made to the participants in the original study. Then, a careful process must be put in place to ensure that data sharing efforts continue to protect the privacy of the participants and maintain the confidentiality of the data.

STEPS TO DE-IDENTIFY DATA

The National Academy of Sciences (NAS) developed a process aimed at sharing data from clinical trials (National Academies Press, 2015). This 12-step process is designed to ensure that the benefits of sharing data are maximized and the risks to participants are minimized. In the current effort, we apply this process (modifying as necessary) toward the sharing of face video data collected in NDS. However, many of the proposed steps, while important, are outside the scope of this effort. Therefore, the five key steps addressed within this effort are:

1. Determine the direct identifiers in the existing dataset.
2. Mask the identifiers in the dataset.
3. Perform threat modeling.
4. Determine minimal acceptable data utility.
5. Determine the reidentification risk threshold.

Determine the Direct Identifiers in the Existing Dataset

The first step is to determine the direct identifiers in the existing dataset. Direct identifiers in data are pieces of information that can be used to identify a specific individual. These may include name, address, phone number, email address, Social Security Number (SSN), passport number,

or a recognizable image of a person's face. While this may be a simple and clear process for some datasets, it becomes it bit more complicated when looking at multimedia (e.g., video, images). Ribaric et al. (2016) have generated a taxonomy of identifiers found in multimedia content that provides a useful reference for this work. First, there are the physiological and behavioral biometric identifiers that are generally unique and permanent. Examples include images of the face, ear, iris, and vein patterns, as well as things like gestures and lip motion. Second, there are soft biometrics that are not necessarily permanent or distinctive. Examples of these in a naturalistic driving face video would be eye color, silhouette, moles, tattoos, birthmarks, and scars. Lastly, there are non-biometric identifiers such as hairstyles, dressing styles, and ID badges.

Mask the Identifiers in the Dataset

The second step requires considering and developing different strategies to mask the identifiers. This can be simply deleting an identifier from the dataset that is intended to be shared or masking the identifier in such a way that it cannot be recovered. For example, the face of a person in an image can be blurred or can be covered with a masking box. For more simplistic cases, names of participants can be removed and replaced with a unique participant ID. For GPS data, the start and/or end locations of a trip can be eliminated to preserve PII.

Perform Threat Modeling

An additional step is to identify potential threats to data should it be distributed. This would involve identifying any adversaries. In the case of the Virginia Tech Transportation Institute's (VTTI's) datasets, this may include institution competitors attempting to embarrass or harm VTTI, insurance companies attempting to individualize the risk profiles of current or prospective clients, someone seeking to take legal action against a participant or their employer, or someone with intent to cause harm or embarrassment to an individual participant. For example, if de-identified driver face videos are being used to study the effects of distracted driving on accident rates, an insurance company may be interested in re-identifying the data in order to identify specific individuals who may be more likely to be involved in accidents due to distracted driving. One must also consider additional information that may be available to potential adversaries and explore whether or not other datasets or information exist that could be linked to the de-identified dataset with the end result leading to reidentification of an individual. Examples of these may be employee records such as assigned routes, accident reports, and vehicle assignments. public and quasi-public databases and information, such as registered voters, the Fatality Analysis Reporting System (FARS), property owners, the Motor Carrier Management Information System (MCMIS), and the National Registry. The availability of newspaper articles about accidents and police accident reports should also be considered.

This exploration of adversaries should also take into account the motivations of the adversaries and the cost incurred to them. Myer (2018) points out that the well-known cases of data reidentification have been performed by researchers simply to prove the point that reidentification of the data was possible. In addition to this possibility, one should weigh the motivation and availability of funds for a fleet, a competitor, an individual, etc., to attempt to reidentify data.

Determine Minimal Acceptable Data Utility

The next step considered for this research effort was Step 4 of the NAS-proposed steps, which involves determining the minimal acceptable data utility (which data must be retained in order for the dataset to be considered useful). The answer to this question will depend on the specific research question. For example, research investigating driver distraction or driver fatigue would require a dataset that retained images of drivers' eye-glance and eye-closure behavior, while a study examining seat belt use would not require the retention of any facial features. The research team can consider creating custom de-identified datasets to share, each containing the minimum number of potential identifiers to answer specific research questions. This approach may require extensive resources that are not readily available. If this is the case, and only one de-identified dataset will be created, researchers will need to determine what it must contain.

Determine the Reidentification Risk Threshold

When considering how the de-identified dataset will be distributed and who will have access, there are a number of steps that should be taken to determine the reidentification risk threshold. The steps provided by NAS are complex and outside of the scope of this work, but the initial part of this process may include determining the acceptable reidentification risk for a publicly available dataset compared to that of a nonpublic dataset. Providing data under the terms of a data use license to a known recipient with an established reputation for research in the scientific community would carry a different reidentification risk than publicly releasing the same dataset.

CHAPTER 3. BACKGROUND AND RELATED WORKS

PII is critical, and privacy should always be maintained while dealing with such sensitive data. There have been instances of serious consequences, like high penalties or even cessation of operations, when PII has been compromised (Isaak & Hanna, 2018). When sharing data involving patients' clinical data, it has become a standard practice to remove PII. As mentioned in Chapter 2, as well as in the literature, various de-identification techniques are practiced. Some of these are discussed below:

- ***Pseudonymization:*** Pseudonymization is a popular de-identification technique in which PII within the record or data is replaced by one or more dummy identifiers or pseudonyms.
- ***Aggregation:*** Aggregation removes the personal details in the dataset by providing a summary or aggregate information about all participants instead of revealing the individual identity of the participants or individuals. For example, releasing the average statistics of customers instead of releasing individual data of each customer is an example of aggregation.
- ***Data Reduction:*** Data reduction deals with completely removing the direct identifiers from the data to make it de-identified. The quasi-identifiers which can be used to potentially identify the individual are also removed.
- ***Data Suppression:*** Data suppression deals with suppressing the direct information. Techniques include giving a data range instead of exact values and clustering the data. Other practices like random rounding are also a part of data suppression.
- ***Data Masking:*** Data masking helps to de-identify data by masking the direct identifiers. Approaches include adding random noise or numbers. Data masking can involve pseudonymization as part of its process.

Table 1 supplements the de identification techniques listed above with examples and details.

Table 1. De-identification techniques and examples.

Method	Original Data	De-identified Data	Details
Pseudonymization	Jacob Williams Age 19 Resident of Blacksburg Student of Virginia Tech	Preslav Ibrahimovic Resident of Blacksburg Student of James Madison University	• The information is partially hidden using dummy names.
Aggregation	John Smith (Age 45 yrs) Tom Brown (Age 20 yrs) Andy Johnson (Age 25 yrs)	Average age is 30 years.	• Total processing of data is done to obtain statistics.

Method	Original Data	De-identified Data	Details
Data Reduction	Student Registration 1998Jan20-Fall2022	The student joined in 2022.	• Deletion of direct identifiers from registration ID data.
Data Suppression	Anthony Robinson Born in 1997	Mr. A Born between 1985 and 1999	• Range is given instead of exact numbers.
Data Masking	Jacob Williams Age 19 Resident of Blacksburg Student of Virginia Tech	J**** W******* 19 years old Resident of beautiful city Student of **** University	• Random noise is added. • Substitution for places, organization.

In addition, when data is shared publicly, there are chances that the data might be used in cases that are not intended by the researchers. Uses of data beyond the consent provided by the participants is unethical and violates the promises made to the participants who consented in the collection of data. The researchers who are involved in collection of the data should ensure that the use of the data aligns with the promises made to participants. In terms of driver face video, data sharing usually involves limited usage terms that allow researchers to use the data only for specific purposes.

NDS DATA

NDS data involves real-world driving from drivers under unconstrained scenarios. The aim is to capture the typical behavior of drivers during their regular driving. The Second Strategic Highway Research Program (SHRP 2) NDS is one of the largest naturalistic light vehicle studies. It has recorded data from more than 3,400 drivers over more than 5 million trips. The dataset includes vehicle kinematics (speed, accelerations), radar data, and GPS data. It also includes videos that capture in-cabin driver behavior and the outside scene. The driver-facing, in-cabin video captures key information about drivers, including their gaze patterns, secondary behavior, and interaction with the vehicular environment.

CONSIDERATIONS FOR DE-IDENTIFICATION OF DRIVER FACE VIDEOS

Data sharing can become easier with fewer restrictions. For sharing driver face videos, de-identification methods generally rely on removal of PII features. For driver face videos, these features can include eyes, skin features, and facial features like nose and lip structure. When PII removal is attempted, steps should be taken to prevent the human cues necessary to build intelligent systems from being lost. For example, there is often a need to preserve the ambience (illumination) and the behavior that the driver exhibits while driving. Typical driver behaviors include head, eye, and lip movements. Preserving the features that contain critical information related to driving behavior allows researchers to use the data in different applications of ADAS. For example, preserving lip movements allows researchers to monitor and analyze behaviors like yawning, laughing, and talking during driving. Similarly, preserving information like eye movements enables researchers to monitor driver attentiveness. The data involving drivers' faces is a challenging task in CV, particularly because of reasons like the large variation in ambient

illumination, driver appearance, and posture. Also, drivers can be involved in a number of different behaviors, such as smoking, drinking water, changing FM stations, or gesturing to other drivers on the road. Sometimes there are even occlusions due to hand movements covering the face or eating behavior. Wearables are another common challenge in face de-identification.

FACE SWAPPING USING CV

CV in the field of AI deals with making computers understand and interpret the visuals around us. With unprecedented developments in computational power and research collaborations, there has been significant progress in CV tasks like facial recognition, face detection, scene perception, emotion analysis, and pose detection. In recent years, advancements in CV have increased by leaps and bounds. Generative models have been the center of everyone's interest since research started with GANs to create photorealistic images. The generative models have been successful in manipulating the appearance of human faces. For example, MichiGAN proposed by Tan et al. (Tan et al., 2020) dealt with changing the hair colors of people with different shades of hair colors. Similarly, Zhu et al. (Zhu, Urtasun, Fidler, Lin, & Change Loy, 2017) proposed a GAN-based method to change peoples' outfits. Advancing a step forward, researchers started to change the face of a person using GAN-based techniques. The generation of new faces using generative models dates back to 2014 when Goodfellow et al. (Goodfellow, 2016) proposed GANs. With more research in the direction of generative models, face swapping started to become increasingly popular among researchers. Face swapping is a task in CV in which the face of a given subject (target) is replaced using a given imposter image (source). It is sometimes referred to as "deepfakes" in the literature (Westerlund, 2019). Face swapping has been applied more recently to meaningful uses. These uses include resurrecting dead artists, creating deepfake reporters who can perform news readings, and even manipulating the movements of famous paintings. Forbes, in 2020, interpreted deepfakes as a friend of humanity (Forbes, 2020) as it has multiple useful applications. Research is ongoing in the direction of using deepfakes to get synthetic data that can be shared easily with fewer restrictions. In the case of the privacy of drivers' face videos, 3D masks have been used to help remove PII. Ideally, a 3D mask should be able to de-identify the video while protecting certain features that are important in transportation research. Although 3D masks do help to remove PII to a great extent, they generally fail to preserve various important information like eye movements and gaze direction (Shao, Lan, & Yuen, 2018).

The advantages of data sharing are immense in a world of technologies that rely heavily on data. One of the main advantages is the flexibility that data sharing provides in experimental setup. By sharing data with fewer restrictions, researchers have more freedom to design and execute their experiments in a way that best suits their needs. This can lead to more accurate and reliable results, as researchers have access to a wider range of data to work with. Additionally, sharing de-identified data can help to overcome limitations in the original data. The data may have limitations on the infrastructure it can be run on, be limited to specific use cases, or have restrictions on sharing even the results. By sharing de-identified data, researchers can process the data in more flexible ways, such as running it on different platforms or using it for different applications.

CHAPTER 4. METHODS AND EXPERIMENTAL SETUP

DATASET

In this study, we used multiple sources of data for validation and testing. The primary source of data that we used for the validation of the algorithms was collected by Oak Ridge National Laboratory (ORNL) and had nine participants. For all the videos, the participants drove a short trip between 6 minutes and 10 minutes long. Additionally, the dataset contains headshot images of the participants. The images of drivers were taken from various angles. For this study, we mainly used the front-facing images. These images were also used for anthropometric measurements of the drivers. The dataset has a unique picture ID and a unique video ID for each image and driving video. The picture ID in the first column of Table 2 and unique video ID in the second column correspond to the image and video of the same person, respectively. The dataset was rich in terms of the demographics of the subjects provided and included both male and female participants and diverse age groups, which helped to understand and qualitatively evaluate the validity of our approach in a detailed way.

Table 2. Participants and physical features.

Unique data ID		Physical Demographics							No. of frames in video
Picture ID (PID)	Video ID (VID)	Male	Female	Old	Young	Glasses		Facial Hair	
						Images	Videos		
863	T002	Y	N	N	Y	N	N	N	11,217
873	T003	Y	N	Y	N	Y (T0)	Y (T0)	Y	14,457
876	T004	Y	N	Y	N	Y (T0)	Y (T0)	N	13,677
880	T005	N	Y	Y	N	N	Y (T2)	N	11,433
883	T006	N	Y	Y	N	Y (T0)	Y (T2)	N	11,517
886	T007	Y	N	N	Y	N	Y (T1)	N	11,769
893	T008	N	Y	N	N	Y (T0)	Y (T1)	N	14,169
897	T009	Y	N	N	N	N	Y (T1)	Y	14,871
906	T011	N	Y	N	N	N	N	N	12,576

Participants and Anthropometric Measures

The physical features of the participants are shown in Table 2. Apart from the qualitative physical features, anthropometric face measurements such as the face width height ratio (FWHR) and cheek-to-jaw width ratio (CJWR) were also calculated. Anthropometric facial analysis helps to quantitatively evaluate the morphology of the human face. It has been used widely in multiple disciplines like pediatrics, facial surgery, and orthodontics. Researchers have used such metrics

widely in the assessment of different transformations that involve facial structures. Since this study involves the transformation of faces to a great degree, these metrics can be of huge importance. FWHR is the ratio of the width of the face to the height of the upper face. Similarly, CJWR is the cheek-to-jaw width ratio as shown in Figure 3. Equations 1–3 and 4 show the mathematical implementation of FWHR and CJWR. For FWHR, there are two settings used in this analysis, FWHR-Brow and FWHR-Lid. FWHR-Brow takes the FHWR where height is taken till the eyebrow, as shown in Figure 4. Similarly, FWHR-Lid takes the FHWR where the height of the face is taken to the eyelid, as shown in Figure 5. For calculation of FWHR, the rectangle whose height and width are taken passes through landmark points 0 and 16 on the sides and through landmark points 50 and 52 on bottom. The upper line of the rectangle slightly differs in the case of FWHR-Lid and FWHR-Brow. The rectangle's top passes through landmark points 18 and 25 in the case of FWHR-Brow, as shown in Figure 4. Similarly, in case of FWHR-Lid, the rectangle's top passes through landmark points 37 and 43. Table 3 shows the anthropometric measures of the participants. From Table 2, the physical demographics like sex and age for various participants can be found. Similarly, the use of glasses by the subjects can also be found in Table 2. For the use of glasses by participants, the glasses have been categorized into three groups. T0 is spectacles with transparent glasses. T1 is glasses with photochromic lenses, which are lightly tinted in normal lighting and turn into dark sunglasses-like shades when exposed to ultraviolet light. Similarly, T2 is sunglasses through which the eyes cannot be seen properly.

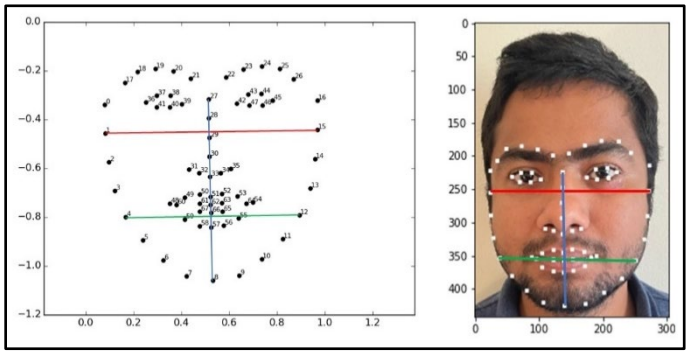


Figure 3. Screen capture. Landmarks for calculation of CJWR (left); example of a face with landmarks (left).

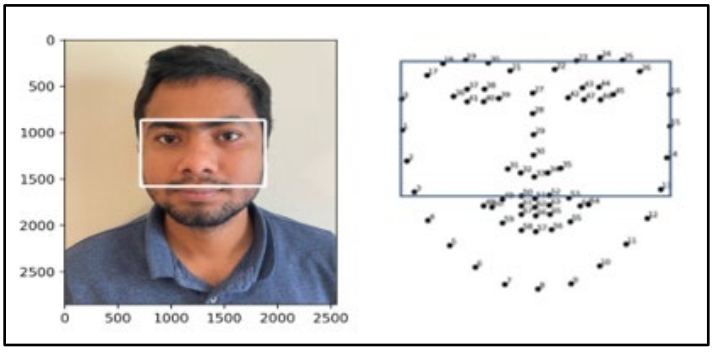


Figure 4. Screen capture. Use of face for calculation of FWHR till eyebrow (left); FHWR till eyebrow of face with landmarks (right).

$$Facial\ Width\ to\ Height\ Ratio\ (FWHR) = \frac{Width\ of\ the\ Face}{Height\ of\ the\ Face} \quad (1)$$

$$FWHR - Lid = \frac{Width\ of\ the\ Face}{Height\ of\ the\ face\ till\ eyelid} \quad (2)$$

$$FWHR - Brow = \frac{Width\ of\ the\ Face}{Height\ of\ the\ face\ till\ eyebrow} \quad (3)$$

$$Cheek - to - Jaw - Width\ Ratio\ (CJWR) = \frac{Cheek\ Width}{Jaw\ Width} \quad (4)$$

From Figure 3, the landmarks for cheek length are (1,15). Similarly, the landmarks for jaw are (4, 12). The 68 landmarks are calculated using Dlib (King, 2009). Dlib's 68-point landmarks help to locate the face features like eyes, nose, lips, and eyebrows. Thus, from Figure 3 and Equation 4, the CJWR is defined as follows:

$$CJWR = \frac{d_h^c}{d_h^j} \quad (5)$$

where,
d_h^c = distance between point 1 and point 15
d_h^j = distance between point 4 and point 12

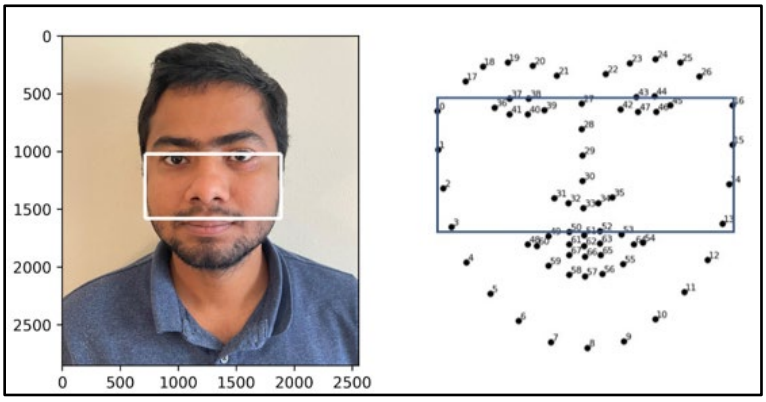


Figure 5. Screen capture. Use of face for the calculation of FWHR till eyelid (left); FHWR till eyelid of face with landmarks (right).

Table 3. Anthropometric measures of participants.

PID	FWHR-Brow	FWHR-Lid	CJWR
863	1.709	1.974	1.1609
873	1.882	2.401	1.1402
876	1.769	2.186	1.1472
880	1.711	2.264	1.2360
883	1.812	2.467	1.1168
886	1.805	2.187	1.1363
893	1.649	2.028	1.1850
897	2.023	2.551	1.2085
906	1.732	2.25	1.1357

Statistics of the Driving Videos

On average, the ORNL driving videos were 6.59 minutes long. The videos were high-quality and were taken at a frame rate of 30 fps. On average, there were 11,862 frames in each video. The number of frames in individual videos are given in Table 2. The videos were taken in changing lighting conditions. The change in lighting in the videos can be attributed to the naturalistic lighting conditions, such as changes in light due to shade from trees or houses, for example.

ALGORITHMS FOR FACE SWAP

The wide accessibility of public data has made it possible to train face-targeted models with large-scale data, which has led to the evolution of deep learning techniques like autoencoders and GANs (Kingma & Welling, 2013). Face swapping or face manipulation usually involves four different levels of manipulation. Entire face synthesis deals with the creation of faces that are not based on existing images using GANs. StyleGAN has been able to create such photorealistic, high-quality images of people (Karras et al., 2020). Similarly, identity swap deals with replacing the face of one person with the face of another person (i.e., an imposter). Identity swap is used for the de-identification of drivers' face videos in which a face in a driver video is replaced by an imposter face. Another face manipulation technique is attribute manipulation, which consists of face-editing techniques that allow changes to facial attributes like beards, hair, gender, or age. Similarly, expression swap, the fourth manipulation, primarily deals with the modification of facial expression (Tolosana, Vera-Rodriguez, Fierrez, Morales, & Ortega-Garcia, 2020). In our implementation, we carried out initial experiments with various algorithms. Ultimately, the SimSwap algorithm was robustly assessed (R. Chen, Chen, Ni, & Ge, 2020). Compared to other algorithms, SimSwap allows the identity to be swapped while preserving attributes like facial expressions and gaze direction, which are important in transportation safety research. The candidate algorithms and their brief features are shown in Table 4.

Table 4. Algorithms for face swapping.

S.N.	Algorithm	Source	Target	Loss Functions	Remarks
1.	FSGAN: Subject Agnostic Face Swapping and Reenactment (Nirkin, Keller, & Hassner, 2019)	Image or Video	Video	Domain-Specific Perpetual Loss, Reconstruction Loss, Adversarial Loss	Uses face reenactment and segmentation, inpainting and blending
2.	SimSwap: An Efficient Framework For High Fidelity Face Swapping (R. Chen et al., 2020)	Image	Video or Image	Identity Loss, Reconstruction Loss, Adversarial Loss and Gradient Penalty, Weak Feature Matching Loss	Uses Encoder, Decoder, and Information Injection Module
3.	Fast FaceSwap (Korshunova, Shi, Dambre, & Theis, 2017)	Image or Videos	Image or Video	Content Loss, Style Loss, Light Loss	Uses realignment and stitching
4.	First Order Motion Model for Image Animation (Siarohin, Lathuilière, Tulyakov, Ricci, & Sebe, 2019)	Video or Image	Image	Different Losses (20 loss terms)	Combines Local Motion for occlusion-aware image generation
5.	FaceSwap GUI ("Face Swap," 2022)	Image	Image	SSIM, MSE, L2 Reg Term, Eye Multiplier, Mouth Multiplier	Has command line interface and graphical user interface

The basic working architecture of a GAN is given below in Figure 6.

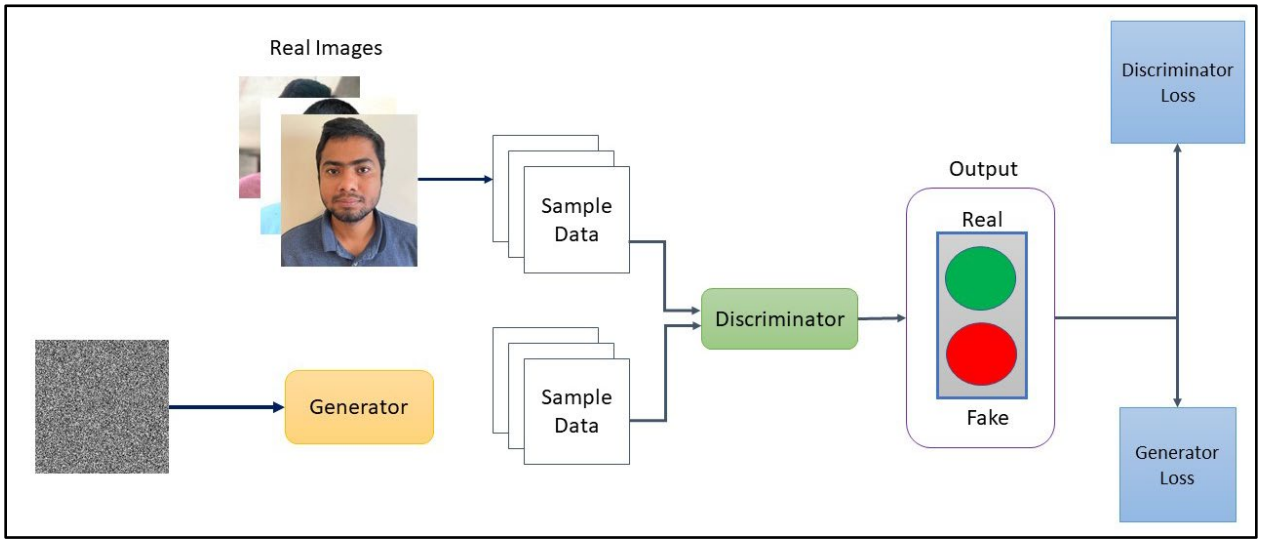


Figure 6. Diagram. Architecture of a GAN. The real images are taken by the GAN model to give output as the de-identified images.

The GAN, as shown in Figure 6, learns to generate new data that is close in statistics to the real image. There are two major networks, a generator and a discriminator. The generator typically

learns how to create realistic images, while the discriminator tries to be better in identifying whether the image is real or fake. The generator's aim is to increase the miss rate of the discriminator, whereas the discriminator aims to be better in flagging the synthetic images. At one point, the generator becomes so smart that it can create photorealistic synthetic images that can fool the discriminator.

The SimSwap algorithm (R. Chen, Chen, Ni, & Ge, 2020) is a high-fidelity face-swapping method. SimSwap aims to achieve high-fidelity face swapping while preserving facial attributes like expression and gaze direction. The authors propose a method called ID Injection Module (IIM), which transfers the identity information of the source face into the target face at the feature level. This allows the algorithm to handle arbitrary face swapping, rather than being limited to specific identities. Additionally, they propose Weak Feature Matching Loss, which helps to preserve facial attributes in an implicit way. The authors demonstrate, through experiments on faces that are captured in unconstrained environments (referred to as "faces in the wild" in CV), that SimSwap achieves better attribute preservation and competitive identity performance compared to previous state-of-the-art methods. The architecture is shown in Figure 7.

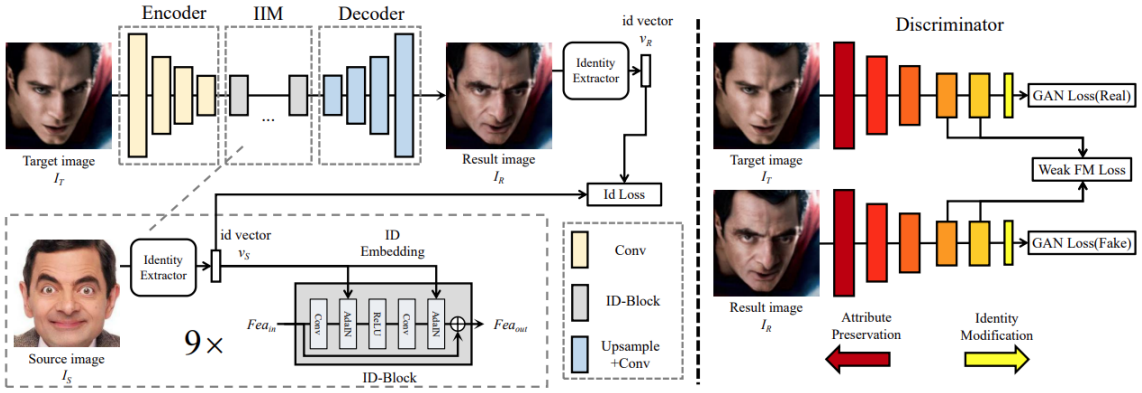


Figure 7. Diagram. Architecture of SimSwap (taken from R. Chen, Chen, Ni, & Ge, 2020).

EXPERIMENTAL SETUP

Our experimental setup can be divided into two parts, as shown in Figure 8. The first one is the use of face-swapping algorithms to swap the faces. The second part is analysis of the output from face-swapping algorithms.

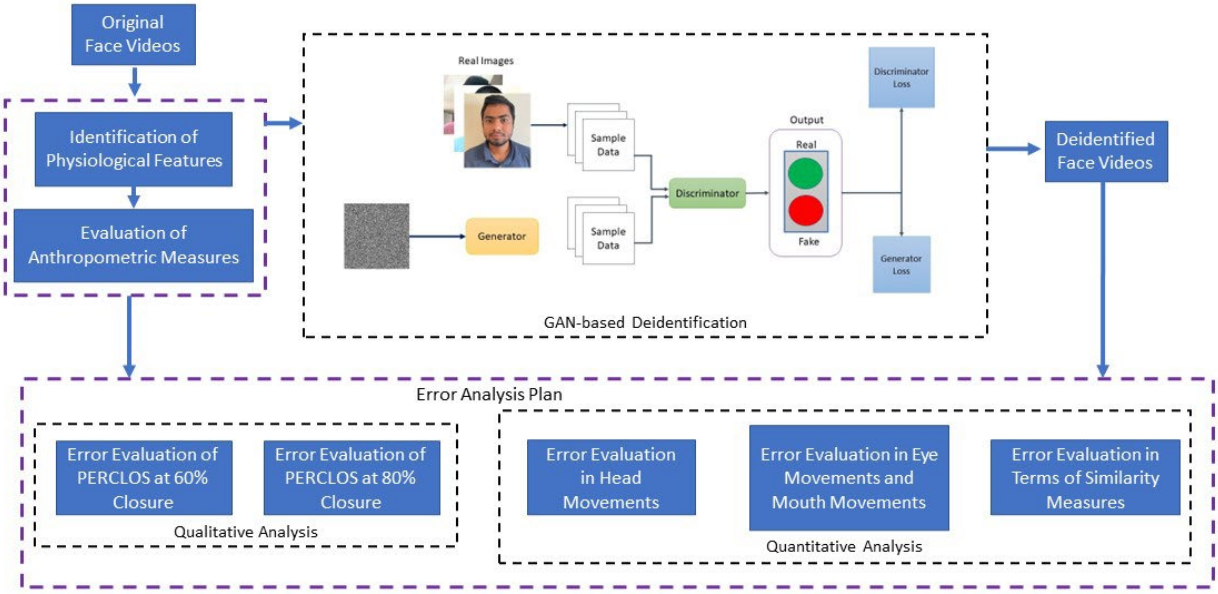


Figure 8. Diagram. Overall process for the de-identification of videos along with error analysis plan.

Face swapping: The first part of the experimental setup involved the use of face-swapping algorithms to replace the faces of the drivers in the videos. The videos were first split into frames, and then each frame was processed individually to swap the face with an imposter face. This process is illustrated in Figure 9, which shows how the original face is replaced with an imposter face.

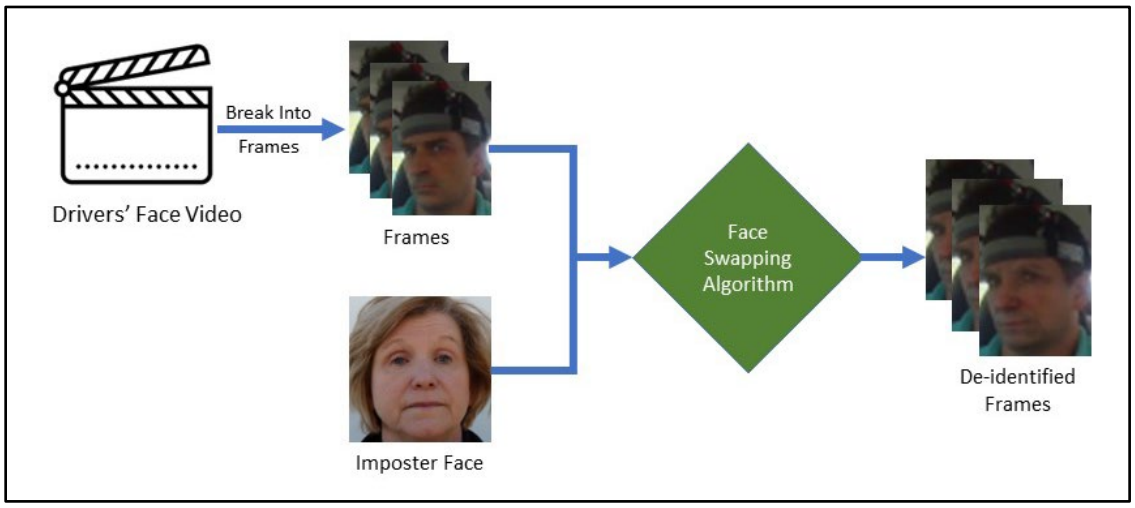


Figure 9. Diagram. Face swapping using imposter face.

The second part of the experimentation involved analyzing the output from the face-swapping algorithms. The head movements, eye movements, and lip movements were taken as the topic of interest for analysis. The analysis of the frames had two different objectives: (1) ensuring that all the frames were de-identified, and (2) ensuring that the human cues important in transportation

safety research were preserved. To ensure that all the frames were de-identified, six error metrics were calculated for all the frames. These metrics were used to measure the similarity between the original face and the imposter face, and to ensure that the imposter face was sufficiently different from the original face. To preserve human cues important in transportation safety research, the error in human cues was calculated and analyzed. This analysis determined if the head movement, eye movements, and lip movements were similar between the original face and the imposter face, so that the human cues important for transportation safety research were not lost in the de-identification process.

CHAPTER 5. DRIVERS' BEHAVIORAL ATTRIBUTES FROM FACE VIDEO

The in-cabin videos from NDS play a key role in studying drivers' behavior. Human factors researchers often use attributes from driver face videos for crash analysis or driver attention monitoring. Different facial cues are important to understand these attributes. For example, the movements of the eyes can tell a lot about how attentive the driver is while driving the vehicle. Similarly, mouth movements can tell whether the driver is yawning or not. Apart from that, mouth movements can also tell if the driver is involved in other secondary activities like eating, drinking, or speaking. Recent research in transportation has proven the importance of algorithms for facial detection, gaze estimation, and eye and mouth movement analysis. Thus, the de-identification task should account for the preservation of such important features. In this research, priority has been given to preserving the head movements, eye and lip movements, and fiducial points. In this section, we discuss a number of key attributes that are widely used in human factors research and deduced from driver face video. We also discuss the error metrics that were used to assess the quality of images.

ROLL, PITCH, AND YAW ANGLES (HEAD POSE)

Head pose is one of the most important factors used for driver behavior monitoring. Gaze direction and patterns can be studied from the head pose information (Murphy-Chutorian, Doshi, & Trivedi, 2007). Head pose also provides important information about the visual attention of the drivers. This can also help in modeling ADAS (Ba & Odobez, 2008). The face is generally taken as a rigid body with three degrees of freedom in the pose (Saeed, Al-Hamadi, & Ghoneim, 2015). These degrees of freedom in the pose are characterized by three rotation angles: roll, pitch, and yaw, as shown in Figure 10. Pitch is up and down like a box lid. Yaw is left and right like a door on hinges, and roll is tilting. More precisely, in the case of the human face, with the human head facing the camera, the pitch is the movement of the head up and down (rotation along the X-axis). Similarly, yaw is the angle of movement of the head left and right (rotation around the Y-axis). Finally, the roll is the tilt angle (rotation around Z-axis).

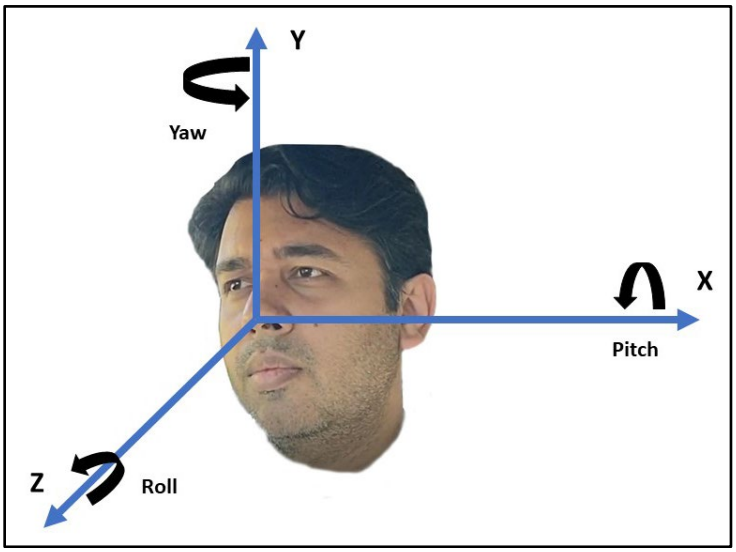


Figure 10. Illustration. The head poses rotation angles (roll, pitch, and yaw). Pitch is along the X-axis. Yaw and roll are the rotation around Y-axis and Z-axis, respectively.

EYE ASPECT RATIO (EAR)

Eye aspect ratio (EAR) is a parameter that is based on the eye landmarks of the face. The eye landmarks were obtained through the Dlib toolkit (King, 2009). The Dlib toolkit gives the landmarks of the frontal face of the driver as shown in Figure 11. Previous works in the literature used methods like ellipse fitting (Tolba, 2019) to evaluate the state of the eyes. The ellipse fitting method uses segmentation techniques for the segmentation of pupils. After segmentation, the ellipse is fit with white pixels that represent the size of the eyes. From the ellipse, the values of the major and minor axes are taken. The ratio of the major and minor axes is then calculated to evaluate the state of the eye. With techniques like ellipse fitting, there are certain problems like inaccurate segmentation. For example, in naturalistic driving situations, the participants might be wearing glasses. Also, NDS occur in dynamic environments that cause the lighting to change. These changes also might pose challenges in the segmentation. Another widely used parameter is eye height-width ratio (EHWR) (Shen et al., 2012). This is basically the ratio of height and width of the eye as shown in Equation 6. This method again relies on just the four pixels of the eyes, which might be inaccurate in naturalistic driving scenarios. Thus, in our experimentation, a more stable parameter, EAR, was used. It is stable as it is based on the landmarks and hence avoids the traditional image segmentation upon which most algorithms are based. As shown in Figure 11, there are 68 points as the facial landmarks. For the eyes, there are six major landmark points (shown in Figure 12).

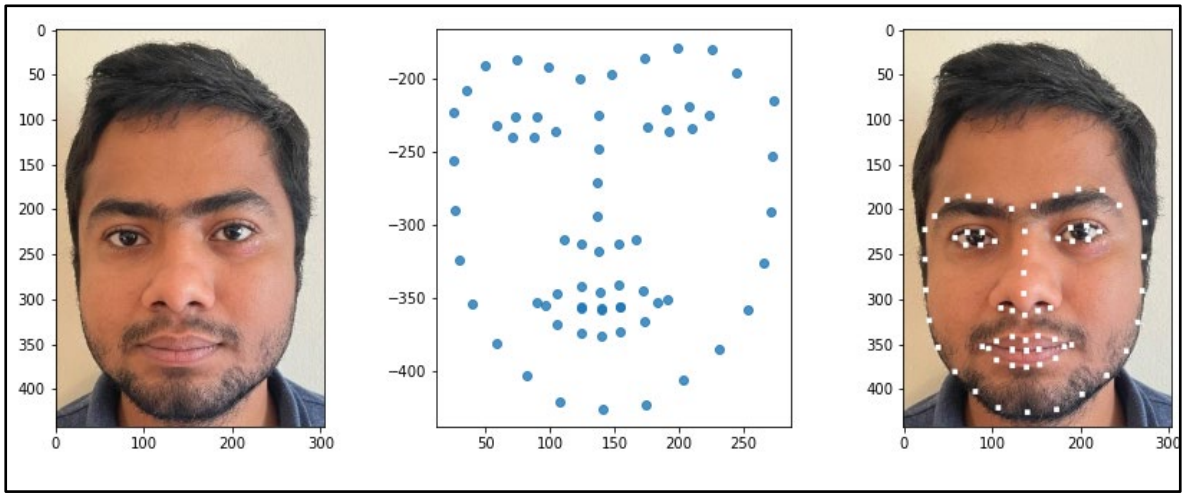


Figure 11. Screen capture. The 68 landmark points for a face.

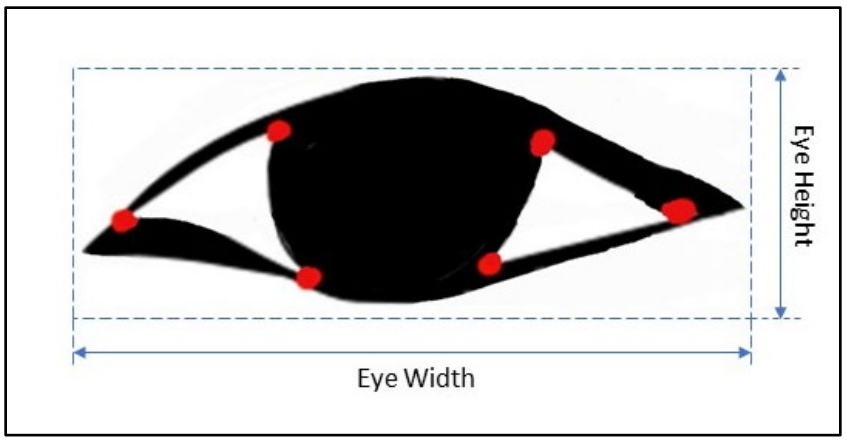


Figure 12. Diagram/ Variables for EHWR.

Mathematically, EyeHeightWidthRatio is given as:

$$EyeHeightWidthRatio\ (\mu) = \frac{EyeHeight}{EyeWidth} * 100\% \quad (6)$$

If $\mu < 27\%$, the eye is identified as closed. If $27\% \leq \mu \leq 45\%$, the eye state should be identified by the state of the eyelid. The conditions are given in Table 5.

Table 5. Conditions for Eye State w.r.t. EHWR.

Conditions		Result
$\mu < 27\%$		Eye is identified as closed.
$27\% \leq \mu \leq 45\%$	$\mu < 40\%$	The eye is closed if eyelid is concave.
	$\mu \geq 40\%$	Eye is open if eyelid is convex.

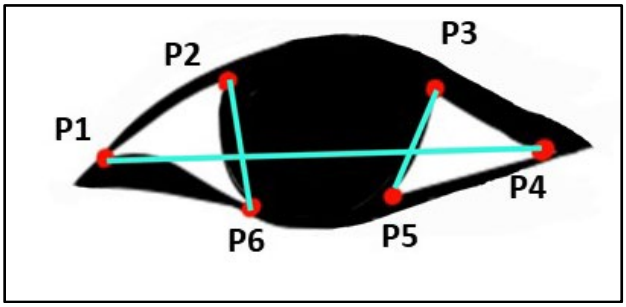


Figure 13. Diagram. Landmarks for calculation of EAR.

From Figure 13, the landmarks used to calculate EAR can be seen. The expression for the calculation of EAR is shown in Equation 7.

$$Eye\ Aspect\ Ratio\ (EAR) = \frac{(d_{v1}^{e} + d_{v2}^{e})}{2 * d_{h}^{e}} \quad (7)$$

where,

d_{v1}^{e} = distance between P2 and P6

d_{v2}^{e} = distance between P3 and P5
d_{h}^{e} = Horizontal length of eyes (between P1 and P4)

.

From empirical observations of the anthropometric measures of the faces of the drivers available, the maximum bound and the minimum bound for EAR were calculated. For doing so, we have hypothesized that at a maximum the width of the eye would be double the mean of the distances d_{v1}^{e} and d_{v2}^{e} . Thus, for the maximum bounds, Equation 7 can be deduced as:

$$Eye\ Aspect\ Ratio\ (EAR) = \frac{(d_{v1}^{e} + d_{v2}^{e})}{2 * d_{h}^{e}}$$

$$EAR_{max} = \frac{(d_{v1}^{e} + d_{v2}^{e})}{2 * (d_{v1}^{e} + d_{v2}^{e})}\ [Since\ d_{h}^{e} = 2 * \frac{d_{v1}^{e} + d_{v2}^{e}}{2}, for\ maximum\ EAR]$$

$$EAR_{max} = \frac{1}{2}$$

Thus, the upper bound for EAR is 0.5. From Table 6, when the maximum EAR among all the frames taken into consideration is calculated, it is 0.47. This validates our hypothesis of the upper bound as 0.50. EAR is a simple yet robust metric, as it is just a ratio and is useful for eyes of all kinds. Since it is just the ratio of two distances, it is dimensionless.

LIP ASPECT RATIO

Similar to EAR, lip aspect ratio (LAR) is an important human cue that can be leveraged for transportation safety research, particularly in ADAS. The landmarks of the lips are taken using the Dlib library, as shown in Figure 11. The landmarks of the lips are shown in Figure 14. Using the landmarks, the LAR is calculated as shown in Equation 8. LAR is also a fractional term, where the numerator is the vertical length of the mouth, and the denominator is the horizontal length of the mouth. Thus, being a ratio of distances, LAR is dimensionless.

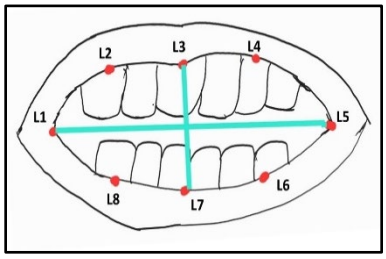


Figure 14. Diagram. Landmarks for calculation of LAR.

$$Lip\ Aspect\ Ratio\ (LAR) = \frac{d_{v}^{L}}{d_{h}^{L}} \quad (8)$$

where,

d_{v}^{L} = vertical distance between L3 and L7
d_{h}^{L} = horizontal distance between L1 and L5

PUPIL CIRCULARITY

Pupil circularity (PUC) is a measure complementary to EAR, but it focuses more on the pupil instead of the entire eye (Varghese et al., 2021). From Equation 9, it can be observed that a

person with eyes half-closed or almost closed will have a much lower PUC value compared to the case where the eyes are fully open. This makes PUC a more sensitive metric due to the squared term in the denominator. Similar to EAR, the expectation is that when an individual is drowsy, their PUC is likely to decline.

$$Circularity = \frac{4 * \pi * Area}{Perimeter^2} \quad (9)$$

where,

$$Area = (\frac{d_r^p}{2})^2 * \pi$$

$$Perimeter = d_{p1}^{p2} + d_{p2}^{p3} + d_{p3}^{p4} + d_{p4}^{p5} + d_{p5}^{p6} + d_{p6}^{p1}$$

where,

d_r^p = distance between P2 and P5
d_{p1}^{p2} = distance between P1 and P2
d_{p2}^{p3} = distance between P2 and P3
d_{p3}^{p4} = distance between P3 and P4
d_{p4}^{p5} = distance between P4 and P5
d_{p5}^{p6} = distance between P5 and P6
d_{p6}^{p1} = distance between P6 and P1

Table 6 shows the maximum, minimum, and the mean of various human cues in the ORNL dataset used in our study.

Table 6. Statistics of human cues for ORNL dataset.

Human Cues	Maximum	Minimum	Mean
EAR*	0.47	0.06	0.26
PUC*	0.70	0.21	0.43
LAR	0.63	0.0	0.06
Pitch	45.53	-54.49	1.71
Roll	45.01	-38.40	0.17
Yaw	87.77	-89.19	-7.84

*For EAR and PUC, only the subjects without glasses and the subjects wearing T0 and T1 glasses are taken.

ERROR METRICS FOR ASSESSING IMAGE QUALITIES

There are a number of metrics that are widely used in the calculation of the similarity of two images in terms of quality. While working on problems typical to that of the de-identification of facial videos, the robustness of the metric is of high importance. If the value of a metric for certain original and de-identified image pairs is different from what is expected, the de-identified frame should be the subject of scrutiny. Since the privacy of human subjects is something that cannot be compromised, there are six image quality metrics that were used in this

experimentation. The purpose of using multiple metrics is that it adds robustness to the identification of frames that should be scrutinized. Apart from this reason, there is not a **single** metric that aligns truly with the human perception of quality of the images. Usually, the image quality is assessed using a full-reference metric. A full reference metric provides a direct comparison between the test image and the reference image without any distortion.

Mean Squared Error

Mean squared error (MSE) is a full-reference metric (Jagalingam & Hegde, 2015). It is a simple yet robust metric that measures the average of the squared difference between the original and de-identified pixel values. However, there is no absolute indication of what MSE value is considered better. It depends widely upon the use cases of MSE. The general rule of thumb is that the lower the value is, the better is the image quality of the de-identified image. However, a zero value of MSE represents that the images are completely identical. Thus, expected ranges are useful. Mathematically, MSE can be formulated as shown in Equation 10.

$$MSE = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - D(i,j)]^2 \quad (10)$$

where,
m and *n* are the height and width of the image in pixels;
i and *j* are the row and column pixels of the given image;
I(*i*, *j*) is the original image;
D(*i*, *j*) is the de-identified image.

Root Mean Squared Error

Root mean squared error (RMSE) is another widely used metric to measure the differences between the given image (original) and de-identified image. It is basically the square root of the MSE (Asamoah, Ofori, Opoku, & Danso, 2018). Mathematically, RMSE can be mathematically formulated as Equation 11.

$$RMSE = \sqrt{MSE} \quad (11)$$

Peak Signal-to-Noise Ratio

The peak signal-to-noise ratio (PSNR) is the ratio between the maximum possible signal power and the power of the distorting noise which affects the quality of its representations (Sara, Akter, & Uddin, 2019). PSNR is widely used in order to calculate the reconstruction losses. Mathematically, PSNR is given by:

$$PSNR = 10log_{10}(peakval^2)/MSE \quad (12)$$

where,

Peakval = maximum value of pixels; for the ORNL dataset used, it is 255.

Universal Image Quality Index

This image quality metric goes beyond the traditional error metrics, which are mostly based on error summation methods. The universal image quality index (UIQI) goes beyond that by

assessing image qualities via a combination of three factors: loss of correlation, luminance distortion, and contrast distortion (Wang & Bovik, 2002).

Mathematically, the implementation for UIQI for pair of images I and D, $Q_{I/D}$ is given by:

$$Q_{I/D} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{ij} \tag{13}$$

where,

$$Q = \frac{\sigma_{xy}}{\sigma_x\sigma_y}.\frac{2\overline{x}\overline{y}}{\overline{x}^2\overline{y}^2}.\frac{2\sigma_x\sigma_y}{{\sigma_x}^2+{\sigma_y}^2} \tag{14}$$

$x = \{x_1, \dots\dots\dots x_n\}$ and $y = \{y_1, \dots\dots\dots y_n\}$ are the original and test image signals, $\bar{x}$ is the mean of x, σ_x is the variance of x, σ_y is the variance of y, and σ_{xy} is the covariance of x, y.

As mentioned earlier, the image quality is assessed using three factors. The first term in Equation 14 represents the correlation coefficient between x and y. The second term measures the degree of closeness of the mean luminance between x and y. Finally, the third term measures how close the contrasts of the two given images are.

Spectral Angle Mapper

The spectral angle mapper (SAM) metric assesses the similarities between the two images in terms of the spectral features. It is basically the cosine of the angle formed between the reference spectrum and the image spectrum. In this work, the reference is the de-identified image, and the image is the original image. The mathematical formulation for SAM is given as:

$$\cos(\alpha) = \frac{\sum XY}{\sum(X)^2\sum(Y)^2} \tag{15}$$

Relative Dimensionless Global Error Synthesis

Erreur relative globale adimensionnelle de synthèse (ERGAS), which translates to "relative dimensionless global error synthesis," is used to determine the quality of the images in terms of the normalized average error of each band of processed images (Renza, Martinez, & Arquero, 2012). This image quality metric is highly sensitive. A higher value of the metric shows that there is some distortion in the de-identified image, whereas a lower value of the metric shows that there is less distortion.

Mathematically,

$$ERGAS = 100\frac{b}{l}\sum_{i=1}^{N}\left(\frac{RMSE\ (i)}{\mu(i)}\right) \tag{16}$$

where,

b and l represent high-spatial resolution and low-spatial resolution images,
$\mu(i)$ represents the mean radiance of the spectral band,
and N represents the number of bands.

CHAPTER 6. RESULTS

De-identification of the drivers' face videos was assessed in terms of quality using two different techniques. First of all, we checked if behavioral features like head movements, lip movements, and eye movements were preserved. Second, we made sure that the de-identification task was performed properly. Figure 15 shows some of the results from the ORNL dataset.

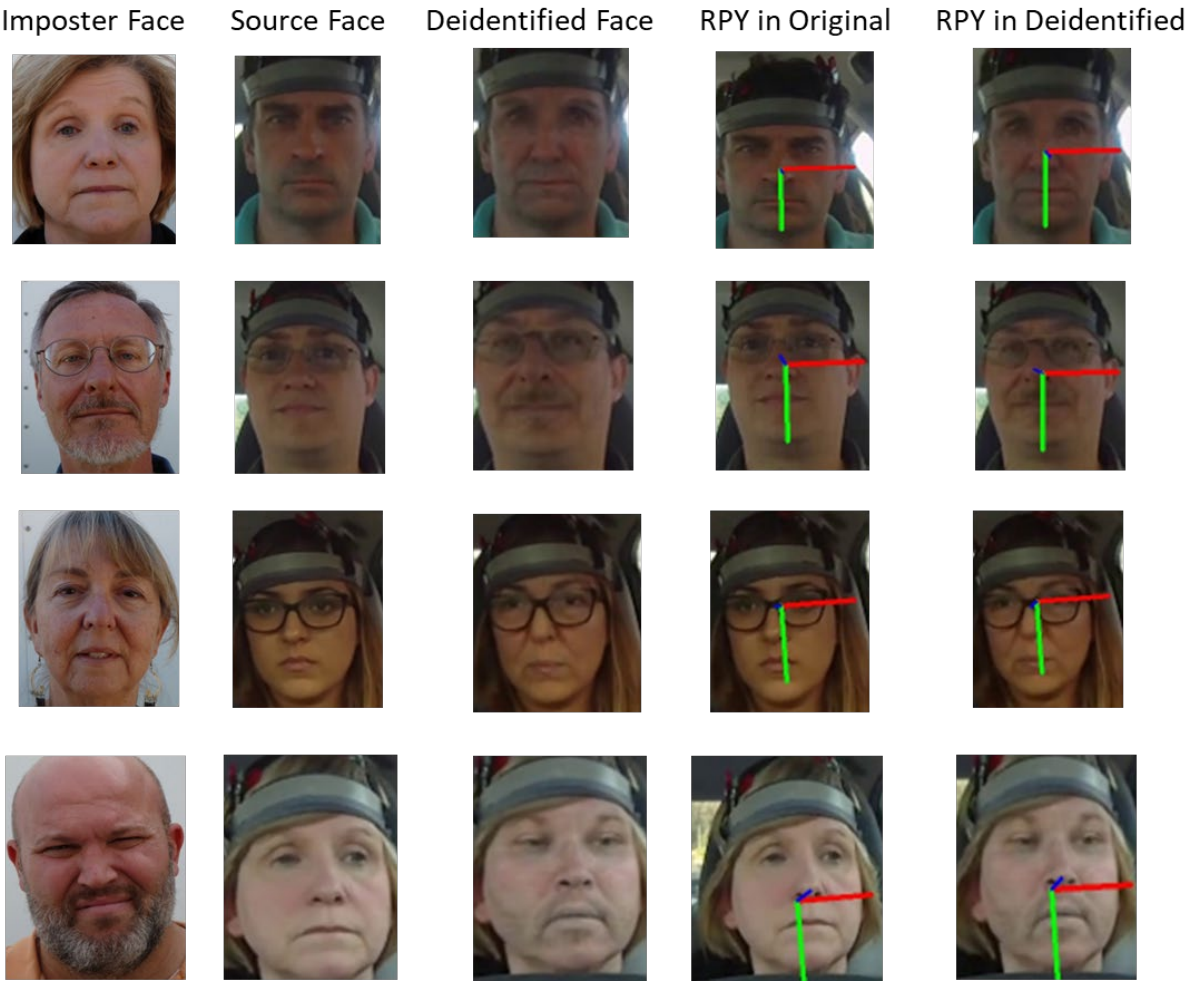


Figure 15. Screen capture. Results from de-identification technique. The figures show that the roll, pitch, and yaw angles are well preserved along with the gaze directions.

ERROR ANALYSIS IN HEAD MOVEMENTS

For the error analysis in head movements, three angles, roll, pitch, and yaw, were considered. In order to give an overview across all the head movements, mean absolute error (MAE) for roll, pitch, and yaw was also calculated. Figure 16 shows roll error and pitch error. The mean roll error and pitch error were less than 5 degrees. Figure 16 also shows yaw error and MAE. The angular yaw error, however, went beyond 8 degrees in one of the cases. The yaw error across VID-T007 for any imposter face was higher than for the other videos (see Figure 29 in the appendix). Inspection of the video revealed relatively higher head movements in the original

video. The higher yaw errors can be attributed to high head movements towards the left and right in the original video. Similarly, the high MAE was because of high yaw error, which predominates over smaller roll and pitch angular errors (see Figure 28 and Figure 29 in the appendix).

Figure 16 also shows a side-by-side comparison of the average errors in roll, pitch, and yaw across the whole ORNL dataset. The yaw error tended to be highly variable, whereas the roll error was the least volatile. The error pattern where Roll Error < Pitch Error < Yaw Error was in fact the expected error scenario. In a naturalistic driving scenario, there is less head-tilt variability and hence the error in roll is less. Also, the head movements up and down are lesser than the head movements sideways. The head movement is more sideways because in naturalistic driving conditions, drivers are required to check their mirrors and blind spots often. This makes sideways movements of the head the most common movement, and hence more error is seen in terms of yaw angles. Figure 17 shows the error statistics in terms of the face swapping. The term "FF" indicates that a female face in the video was replaced by a female imposter face. Similarly, "MF" means that a male face in the video was replaced by a female imposter face. FM indicates that a female face in video was replaced by a male imposter face, whereas MM represents that the male face in video is replaced by a male imposter face.

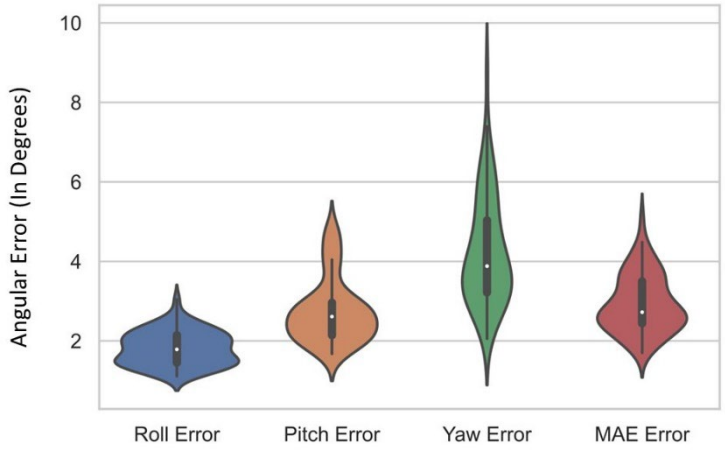


Figure 16. Chart. Violin plot for roll, pitch, and yaw angular errors along with MAE.

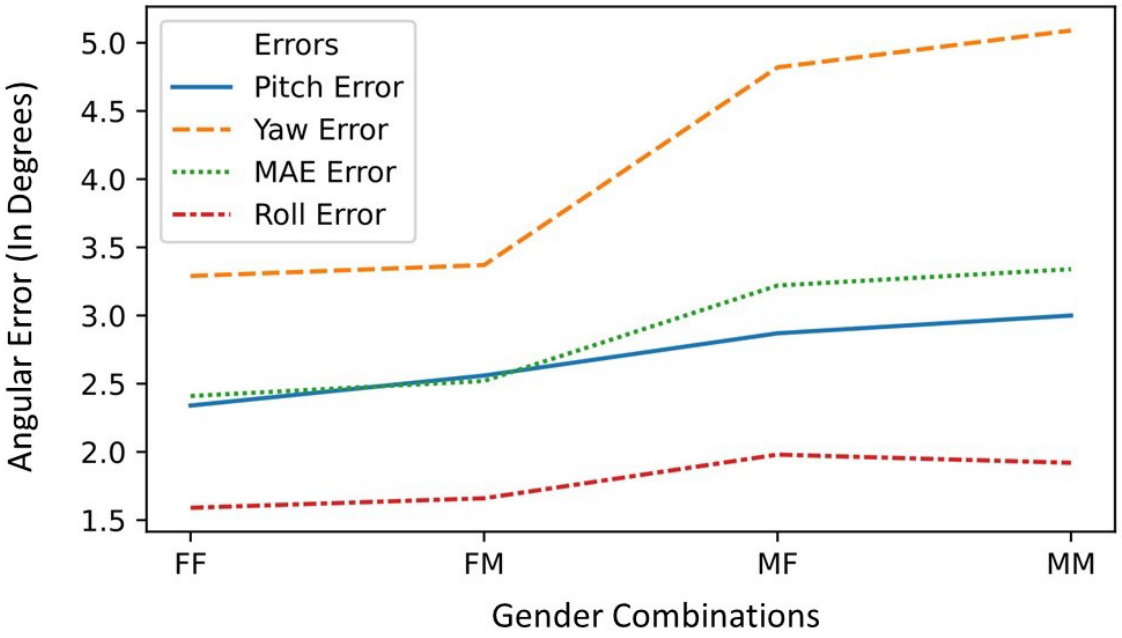


Figure 17. Chart. Error (in angles) with respect to gender pairs (target-imposter pair).

From Figure 17, it can be seen that the error was lowest when a female face in a video was replaced by a female imposter face. Similarly, from Figure 18, it can be seen that average yaw error for the ORNL dataset was 4.2 degrees. The error is in an acceptable range and, hence, it can be concluded that with respect to preserving head movements, face swapping algorithms do a great job.

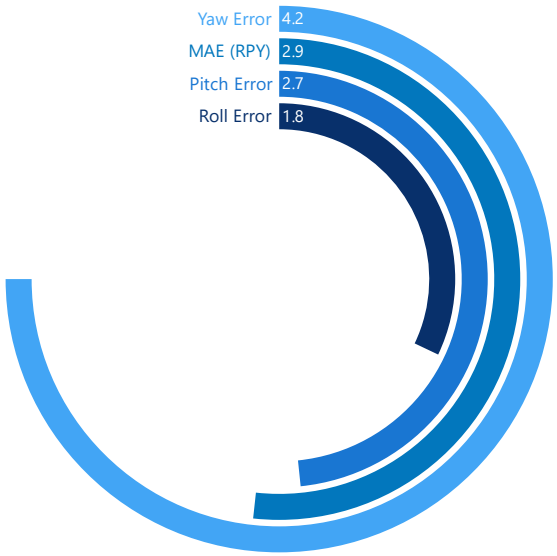


Figure 18. Chart. Overview of roll, pitch, and yaw angular errors in degrees.

ERROR ANALYSIS IN DRIVERS' EYE AND MOUTH MOVEMENTS

From Table 6, it can be noted that the average EAR across all the frames is 0.26. Similarly, the average circularity is 0.43. The highest EAR and circularity are 0.47 and 0.70 respectively. From Figure 19, the EAR error is less than 0.06 for most of the cases. The circularity is also nearly less than 0.075 for all the videos. This shows that even in the de-identified videos, the EAR and circularity are very well preserved. This low error shows that the de-identified videos can readily be used to build safe driving models, such as distracted driving detection models. The human cues are well preserved. The heatmaps for circularity and EAR (see Figure 30 in the appendix) look nearly identical. This shows that the de-identified face videos have promising future prospects to preserve the aspects of the eye well. Eye movements, being an important element in transportation safety research, help to understand secondary behavior in de-identified videos well.

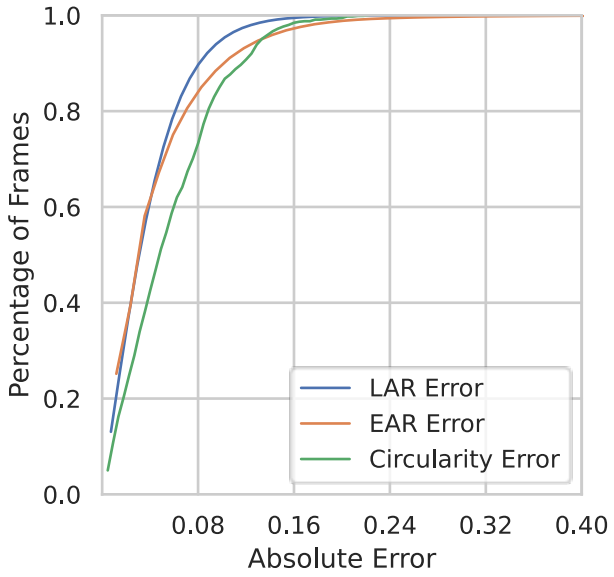


Figure 19. Chart. Percentage of frames vs. absolute error for EAR, LAR, and circularity.

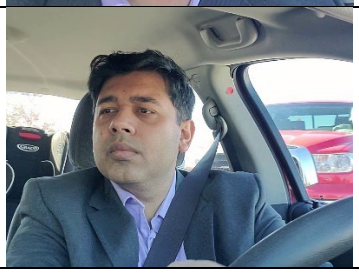
From Figure 19, it can be seen that for most of the frames, the error is well below 0.10 for all the parameters. This shows that the error is not very high, and prospects can be explored to define acceptable ranges. It also shows that there is less LAR error (more detail in Figure 31 in the appendix). Lesser LAR error means that the de-identified videos can be used to build models to predict yawns and other behavior in which lip movements are involved.

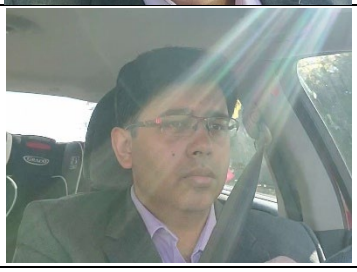
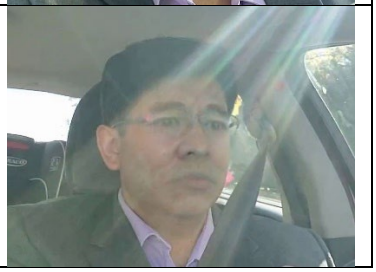
ANALYSIS OF SECONDARY ACTIONS

In order to analyze how well the de-identified videos preserve secondary actions, we performed a qualitative analysis of different cases. Some of the cases for secondary behavior are given in Table 7. For the driver's face in Table 7, the face was swapped with a face that had a different racial profile, as shown in Figure 20.

Figure 20. Photos. Replacement of face video of driver for analysis of secondary behavior.

Table 7. Analysis of secondary behavior for drivers' face videos.

Secondary Action	Original Face	Face after Face Swapping
Looking ahead with glasses		
Closing eyes		
Speaking		
Parking		

Secondary Action	Original Face	Face after Face Swapping
Using features on the dash		
Harsh lighting conditions		

QUALITATIVE ANALYSIS OF DE-IDENTIFIED VIDEOS

PERCLOS Agreement Analysis

Various measures are used in the assessment of drowsiness. There are two major approaches for drowsiness detection. The first one is vehicle-based, where driving information, such as the movement of the steering wheel and patterns in acceleration and braking, is used to assess drowsiness. The second approach for drowsiness detection is based on the direct observation of the driver for changes in vital signs and eye movements. Assessment of drowsiness based on driver behavior can further be divided into the use of visual and non-visual features. The assessments based on visual features typically take eye movements and mouth movements into account. On the other hand, assessment based on non-visual features takes vital signs like electroencephalogram, electrocardiogram, electrooculogram, and photoplethysmography into account. The vital-signs-based approaches can often be intrusive. Thus, with advancements in deep learning, visual-features-based techniques are widely used since they are non-intrusive. PERCLOS is one of the widely used measures. Similarly, there are other non-intrusive measures like eye closure duration and frequency of eye closure that are also used. In this report, we have taken EAR as the measure for evaluation of drowsiness. We have also tested our results with PERCLOS to assess the agreement of EAR with PERCLOS.

PERCLOS at 80%

PERCLOS at 80% means that if the openness of the eye is less than 80%, the data reductionist labels the image as eye-closed. For our PERCLOS assessment, we took six pairs of original and de-identified video, each of length 3 minutes. The confusion matrix is shown in Figure 21.

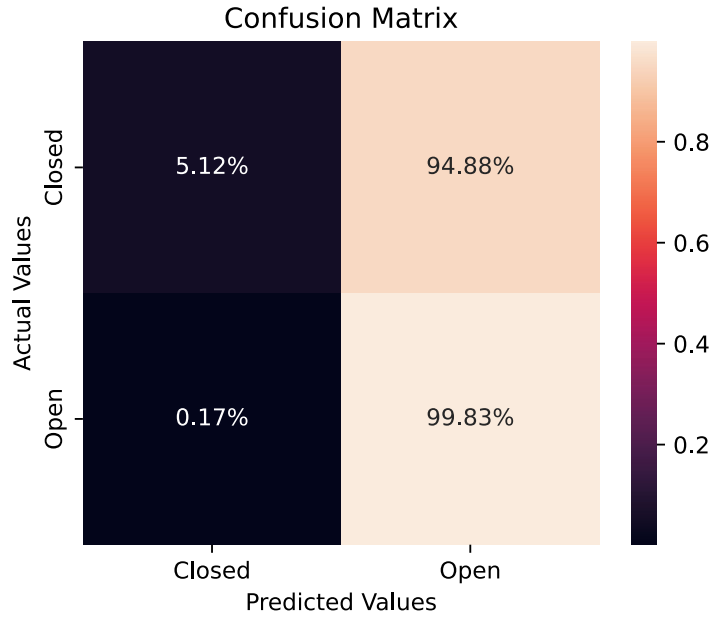


Figure 21. Chart. PERCLOS results for PERCLOS at 80%.

PERCLOS at 50%

PERCLOS at 50% means that if the openness of the eye is less than 50%, the data reductionist labels the image as eye-closed. For our PERCLOS assessment, we took six pairs of original and de-identified video. The confusion matrix is shown in Figure 22.

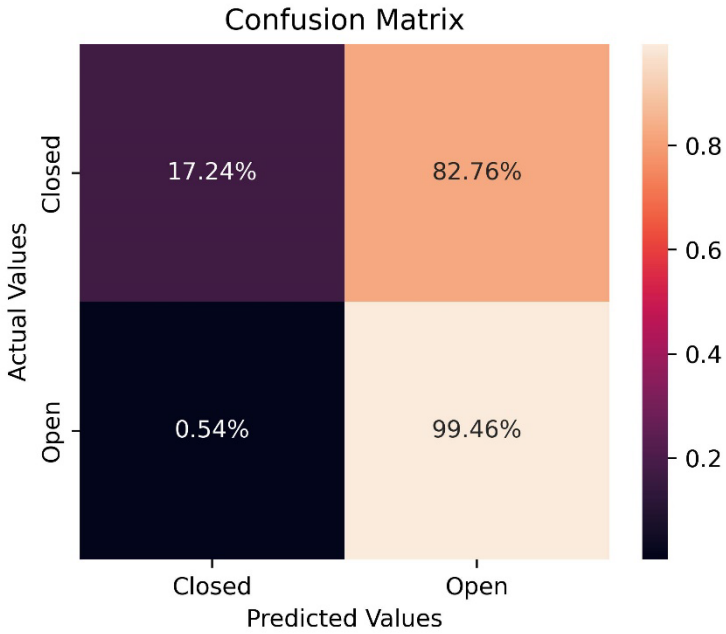


Figure 22. Chart. PERCLOS results for PERCLOS at 50%.

Analysis of PERCLOS

As the result in Figure 22 does not show a good agreement with the human annotator, we further investigated the reason for the errors. We plotted the EAR value for the original video and de-identified video for an eye blinking case. The result is shown in Figure 23. We can see a clear notch for eye closure (around 23 seconds) in both videos. This shows that EAR can identify the eye closure. However, the threshold may be different. PERCLOS has been a standard method for drowsiness detection in the literature. However, it needs manual annotation and may introduce human bias. EAR on the other hand provides a measure for automated calculation of eye closure using a CV algorithm. More research is required to understand how effectively EAR can substitute for PERCLOS, which is out of the scope of this work. Additionally, in the context of the de-identified video, we can see that the de-identified video still preserved the eye closure event, but its magnitude changed between the original and de-identified versions. A possible reason for that could be the subjective difference of the original face and de-identified face. The de-identification changes the anthropometric measurements of the face. This can result in a different EAR. Hence, more research is needed to study the variation of EAR for different anthropometric measurements. A trend analysis and a notch detection method can still detect an eye closure epoch.

From Figure 23, blinking is represented as a notch in the EAR value. This shows that the blinking pattern of the eyes is preserved. Despite having lower agreement with the manually reduced data, the blinks can be detected using EAR. De-identified videos have some limitations when it comes to manual validation using PERCLOS. In the de-identified videos, the EAR slightly changes from the original but the track of the blinks can be kept. PERCLOS is not highly robust in assessing eye closure in cases of varied eye size in de-identified videos. Thus, there is a poor performance, as shown by the confusion matrices in Figure 21 and Figure 22. PERCLOS validation in de-identified videos also has certain limitations in terms of working well in low illumination conditions, but EAR is a robust measure that can deal with even poor lighting conditions.

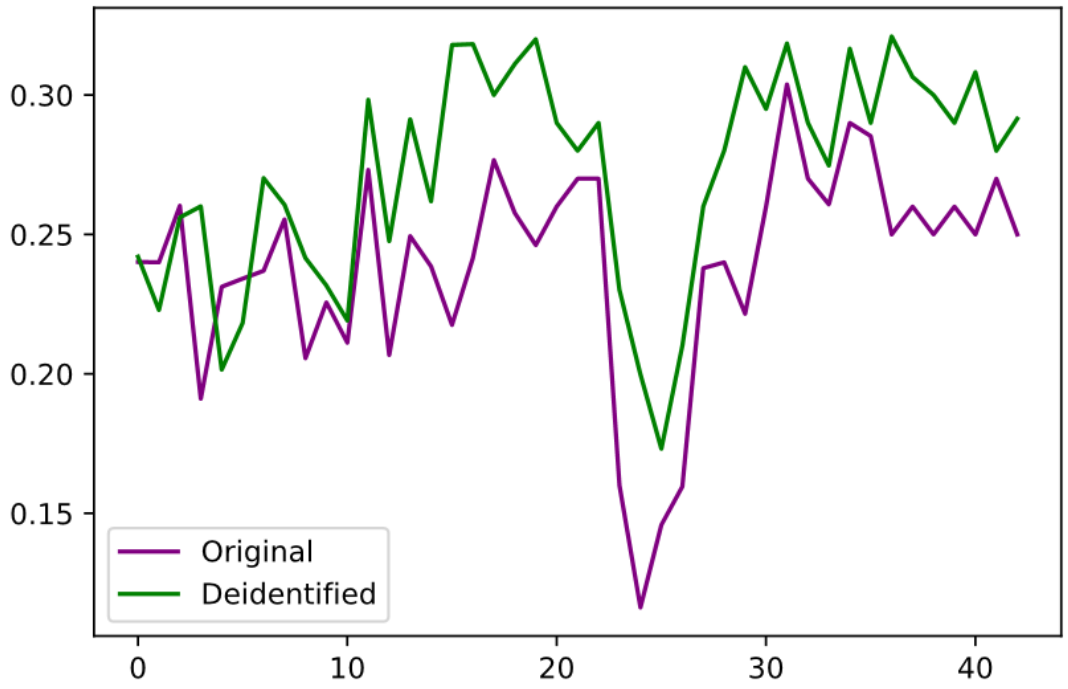


Figure 23. Chart. Comparison of EAR for short clip of de-identified and original videos.

QUANTITATIVE ANALYSIS OF DE-IDENTIFIED VIDEOS

Table 8. Statistics of error metrics for all frames of ORNL dataset.

Metric/Image	Original	De-identified			
		Minimum	Maximum	Mean	Remarks
MSE	0	0.873	99.39	15.88	Lower value means highly similar.
RMSE	0	0.935	7.969	3.909	Lower value means highly similar.
PSNR	inf	28.16	48.72	36.45	Higher value means highly similar.

Metric/Image	Original	De-identified			
		Minimum	Maximum	Mean	Remarks
UIQI	1	0.439	1	0.996	Higher value means highly similar.
ERGAS	0	247.21	16423.93	2200.12	Lower value means highly similar.
SAM	0	0	0.805	0.04	Lower value means highly similar.

For the quantitative analysis of the de-identified videos, the heatmaps the for six error metrics are given in the appendix (Figure 32 to Figure 34). Table 8 shows that the RMSE and MSE metrics are sensitive and can range from the smallest to the largest errors (see Figure 32). The metric UIQI has almost similar values for all the videos and hence is considered to be very insensitive. Similarly, from Table 8, the ERGAS metric seems to be very sensitive and hence can predict if the frames are de-identified (Figure 34). For all the de-identified frames, the errors with the original image should never be zero. If the errors are zero, trivially, the de-identified frames and original frames are the same and hence human interception is needed to check if de-identification has happened.

EXPERIMENTS WITH OTHER NDS DATA

Collision Avoidance System Field Operational Test

The de-identification framework was tested with data from the Collision Avoidance System (CAS) Field Operational Test (FOT). Due to privacy concerns, the examples used from the data are not shared in this report. The de-identification framework proposed here performed well in the tests conducted with the CAS FOT data. The face-swapping algorithms used effectively de-identified faces while still preserving important attributes related to human factors research, including eye movements, head movements, and mouth movements. The results were evaluated both qualitatively and quantitatively, and the proposed methods were found to be valid. Even with videos taken at night, the de-identification framework worked well.

VTTI L2 NDS

The de-identification framework was tested with a dataset from the VTTI L2 NDS. The results were promising, despite some limitations. The black-and-white video footage did not perform as well as the color video footage, as color information is crucial for facial recognition algorithms to accurately identify and de-identify individuals. Additionally, the framework struggled to effectively de-identify individuals who were wearing glasses, as the eyes appear more bulged in black-and-white footage, and glasses can obstruct important facial features used for identification. Despite these limitations, the framework performed well overall and was able to effectively de-identify a majority of the individuals subjected to testing.

Use of Synthetic Faces in De-Identification

With rising concerns over PII, using the faces of real participants in transportation may be problematic. To avoid this, we used fake faces that are not known to exist in real life, which were generated using StyleGAN (Karras et al., 2020). The faces are imagined by GANs and hence such faces can be used to improve diversity in data as long as the training data is unbiased and includes diverse examples. Figure 24 shows examples of synthetic faces used to replace the faces of real drivers. It is noteworthy that the face-swapping techniques hold well even when the participants are wearing glasses.

Figure 24. Illustration. Use of synthetic faces to replace the faces of the drivers.

Even with the fake faces, the roll, pitch, and yaw angles were preserved, along with other human factors like EAR and LAR. Figure 25 shows that there is less error in terms of roll, pitch, and yaw angles. Similarly, the EAR, LAR, and circularity error are also similar to the case when faces were replaced by real human faces. The major benefit of having drivers' faces replaced by synthetic faces is that it improves the diversity in the data as long as the training data is unbiased and includes diverse examples.

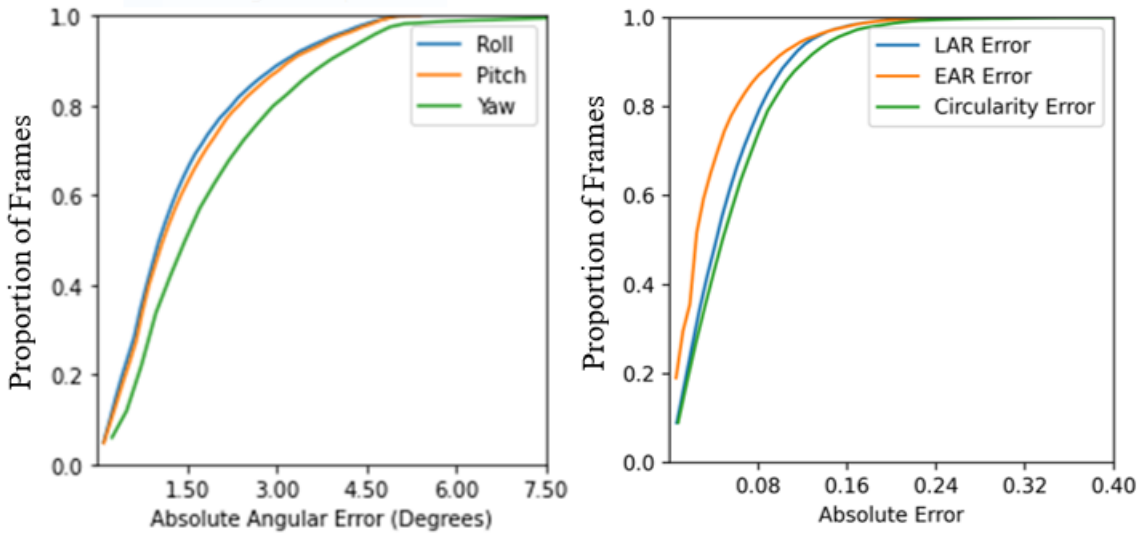


Figure 25. Graph. Error statistics when the drivers' faces are replaced with synthetic faces.

CHAPTER 7. CONCLUSION

In this work, we evaluated the effectiveness of face-swapping algorithms in protecting the privacy of drivers by de-identifying face videos. We evaluated the efficacy of face-swapping algorithms at preserving multiple human factors attributes. Our results show that face-swapping algorithms have potential to preserve the privacy of drivers. We have qualitatively and quantitatively shown that this work holds good future prospects for the de-identification of drivers' face videos, and the experimental setup can be used as a framework to evaluate other face-swapping algorithms in the future. Additionally, we provide a brief guide to the error analysis plan, possible future work, limitations, and additional advantages.

ERROR ANALYSIS PLAN

In order to benefit from face-swapping for de-identification of drivers' face videos, large-scale processing is important. The scalability of the proposed framework in this work can reap greater benefits in the curation of a large dataset. The algorithmic implementation of face-swapping can be made more efficient with distributed computing and recent advances in high-performance computing. On the other hand, checkpoints are needed to make sure that the de-identification is done with proper guidelines. For this, a proper error analysis plan that involves human-in-the-loop validation is also important. Since the purpose of this study was the de-identification of driver face videos while preserving the human factors cues, the evaluation of the results was done in two steps. First, the image quality of the frames in the de-identified videos was evaluated with respect to the original videos. Second, human cues that are useful in transportation safety research, such as head movements, lip movements and eye movements, were used. To automate the de-identification of large datasets using the framework provided, we suggest two major steps.

Creating Error Threshold and Spot Checking

From the experiments, we found that the error bounds for various error metrics needed to be defined. By training on a larger dataset that accounts for larger nuance, a more robust error threshold could be defined. For an example, in this experimentation with the ORNL dataset, it was found that for properly de-identified images (visual inspection and not recognized by recognition algorithm), the error was prevalent for all the metrics. For example, for a given original and de-identified pair, the error should never be zero. For a frame to be de-identified, it should have acceptable error across all the metrics. Even if there is a single instance of error being out of acceptable range, it should be subjected to spot checking.

Spot Checking for Frames with Abrupt Changes in Metrics

Most error metrics can detect unusual behavior. As shown in Figure 26 and Figure 27, unusual dips and peaks can be observed. Human-in-the-loop verification is suggested for such unusual dips and peaks. It is better to take a more sensitive metric like ERGAS (which can range from 0 to tens of thousands). In our data, the maximum ERGAS is in the range of 16,000, whereas the minimum is in the range of 200.

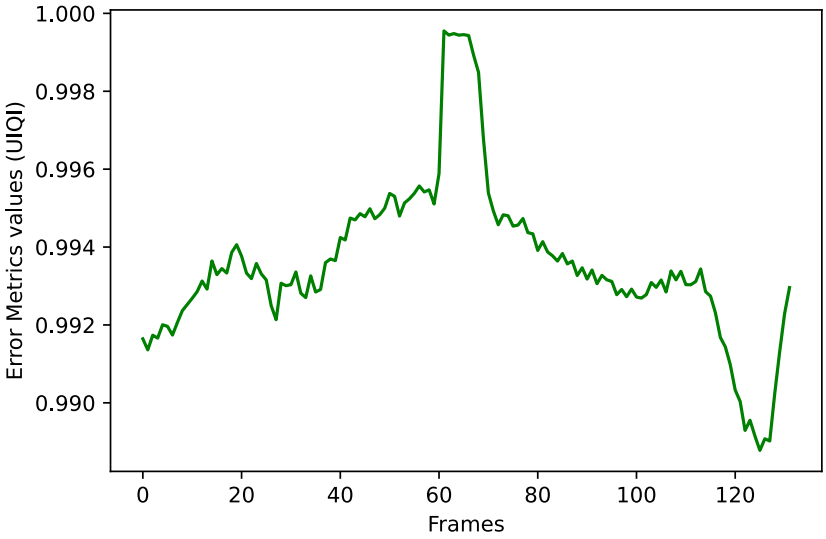


Figure 26. Graph. UIQI error over the frames of de-identified videos.

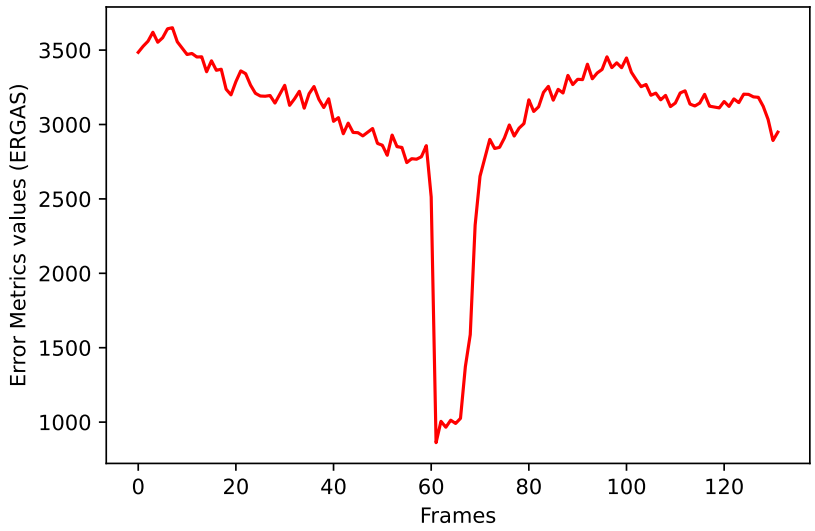


Figure 27. Graph. ERGAS error over the frames of de-identified videos.

Considerations

There are several fields of research where reidentification of a person is a major topic of interest. The IEEE Biometrics Council and the Computer Vision Foundation are some of the established research threads that actively work on reidentification. Reidentification is necessary on its own terms, but in the context of this project, we should be aware that reidentification is a risk for de-identified videos and we should develop and deploy enough countermeasures to minimize the chance. For this, threat modeling should be performed. Threat modeling generally includes steps

such as identifying adversaries, their access to additional information, and their benefits from reidentification. We demonstrated through multiple image-based metrics that can be used to guarantee this. However, more focused analysis and research may be needed to study further vulnerability to different kinds of advanced reidentification methods.

LIMITATIONS

The proposed framework for automating the de-identification of large datasets relies on the use of AI-based techniques, specifically face-swapping algorithms. While these algorithms have been demonstrated as a possible future direction at de-identifying faces while preserving important attributes related to human factors research, they are not without limitations. One limitation is that the accuracy and robustness of the face-detection algorithms used in the framework can be affected by a number of factors, such as lighting conditions, camera angle, and facial expressions. Most of the failure cases we observed originated from such extreme lighting and pose. In particular, the face-detection algorithms used in the framework may not be able to detect faces in certain lighting conditions or camera angles, which can lead to errors in LAR and EAR. Additionally, some other failure cases include rapid movements of the eyelids or mouth. Although we have tested on multiple data sources from VTTI's standard data collection systems that have different resolution and compression, we find that lower resolution and high compression can degrade the performance.

This work only deals with the de-identification of drivers' face videos. More work is needed in de-identification of drivers' appearances with non-biometric identifiers like clothing, hairstyle, etc. These features can be used for the reidentification of drivers' videos, especially if distributed to people who already know these people (fleet managers for example). Recent advances in CV have led to the development of various algorithms that can replace hairstyle, clothing styles, etc. A possible future direction is to integrate face-swapping algorithms with algorithms that help to de-identify the non-biometric identifiers as well.

FINAL THOUGHTS

In this report, we have presented a way to address restrictions due to PII in NDS. We have also proposed ways to quantitatively assess whether the human factors cues necessary for safety research (e.g., head movements, mouth and eye movements) are preserved. Leveraging the recent advancements in CV, this work primarily deals with the removal of biometric identifiers from the face area. Through extensive experimentation, we have demonstrated the feasibility of a face de-identification method. These methods show promise to develop realistic de-identified driver face video. As this research mainly focus on the face area of a driver, we should note that an adversary can use soft biometric identifiers or non-biometric identifiers to identify the participants, especially when reidentification attempts are made by people who are already familiar with the participants' soft biometrics and environmental cues. Therefore, we should have additional data user licenses and measures to protect privacy. In this light, this research is a first attempt toward driver de-identification, and we believe that this work will stimulate additional discussion and research initiatives that will create more robust face de-identification.

APPENDIX A. SUPPLEMENTARY FIGURES

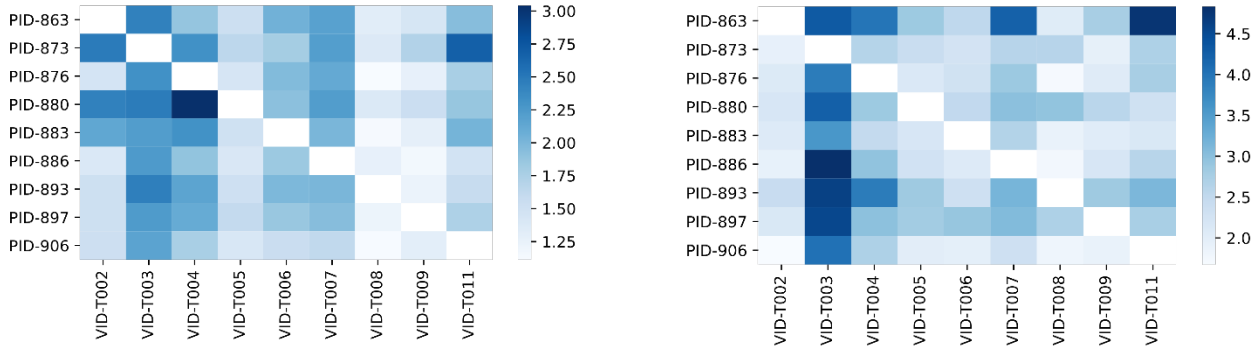


Figure 28. Graph. Roll error (left) and pitch error (right).

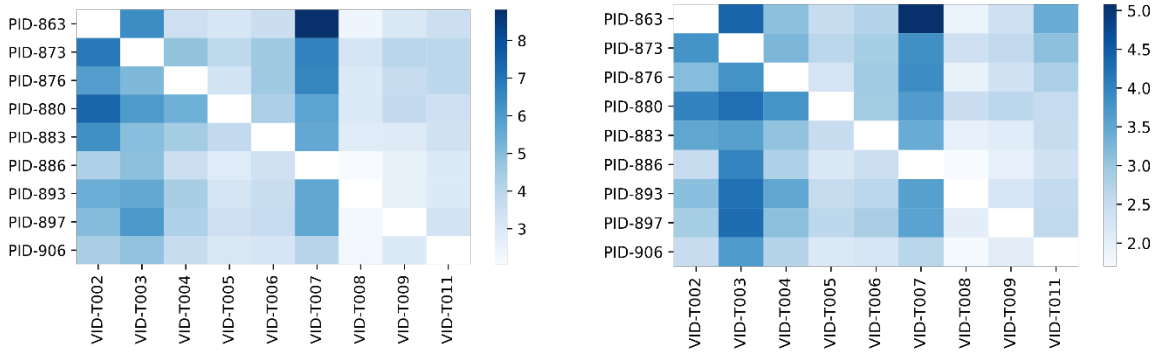


Figure 29. Graph. Yaw error (left) and MAE error (right).

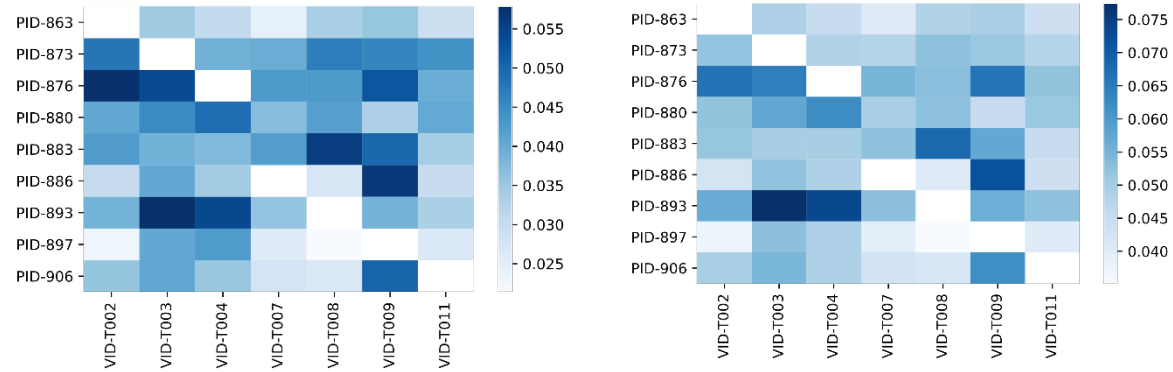


Figure 30. Graph. EAR error (left) and circularity error (right).

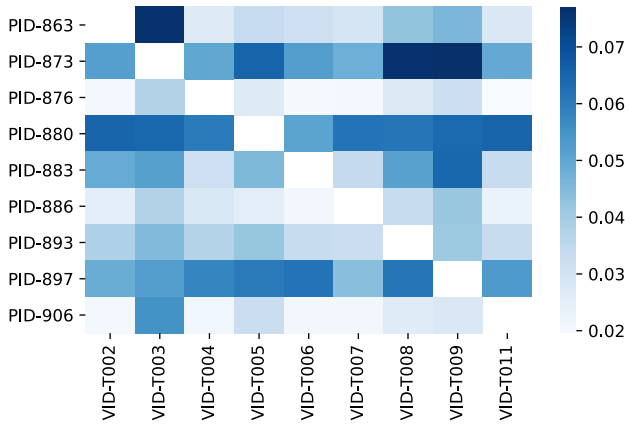


Figure 31. Graph. LAR error for ORNL dataset.

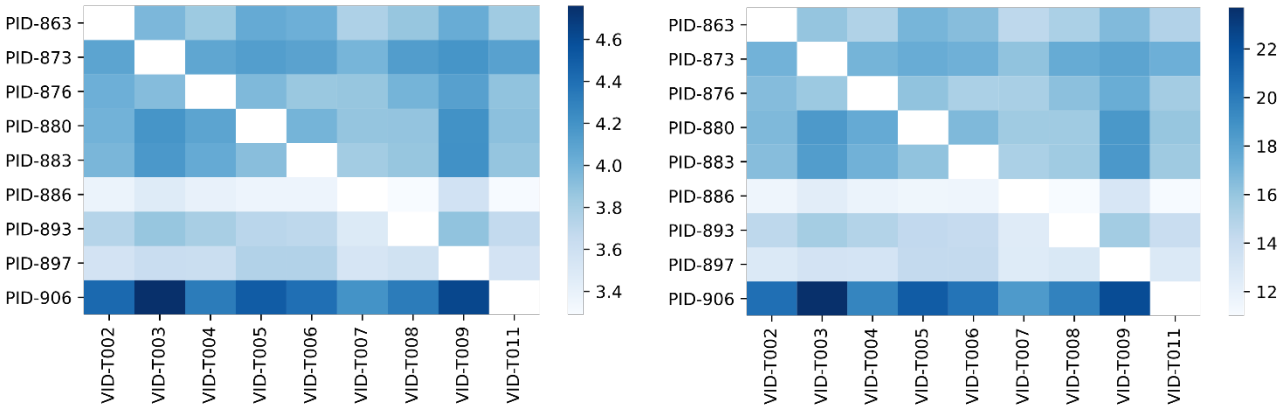


Figure 32. Graph. (a) Error metric MSE (right), (b) error metric RMSE (left).

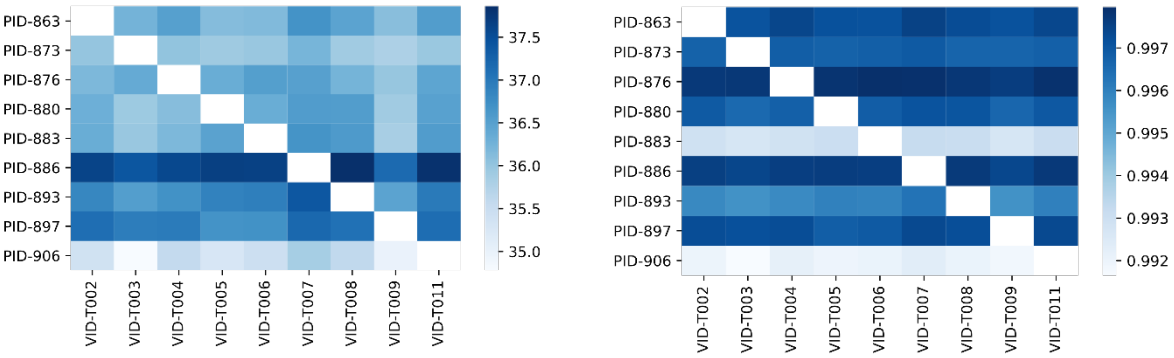


Figure 33. Graph. (a) PSNR error (left), (b) UIQI error (right).

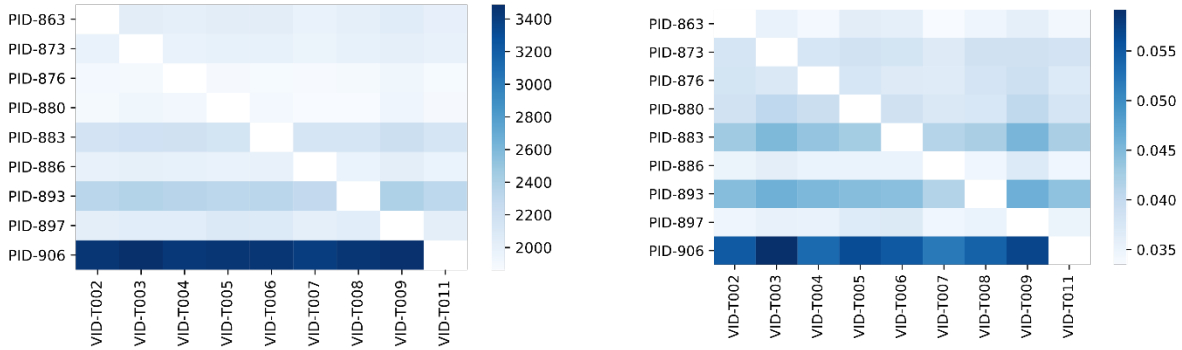


Figure 34. Graph. (a) ERGAS error (left), (b) SAM error (right).

REFERENCES

Asamoah, D., Ofori, E., Opoku, S., & Danso, J. (2018). Measuring the performance of image contrast enhancement technique. *International Journal of Computer Applications, 181*(22), 6-13.

Ba, S. O., & Odobez, J.-M. (2008). Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 16-33.

Burrell, J., & Toyama, K. (2009). What constitutes good ICTD research? *Information Technologies & International Development, 5*(3), pp. 82-94.

Chen, R., Chen, X., Ni, B., & Ge, Y. (2020). *Simswap: An efficient framework for high fidelity face swapping.* Paper presented at the Proceedings of the 2^{8t}h ACM International Conference on Multimedia.

Face Swap. (2022). Retrieved from https://faceswap.dev/

Forbes. (2020). Why deepfakes are a net positive for humanity. Retrieved from https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., . . . Badri, O. (2021). *Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51*(8), 56-59.

Jagalingam, P., & Hegde, A. V. (2015). A review of quality metrics for fused image. *Aquatic Procedia, 4*, 133-142.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of StyleGAN.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research, 10*, 1755-1758.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). *Fast face-swap using convolutional neural networks.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Menikoff, J., Kaneshiro, J., & Pritchard, I. (2017). The common rule, updated. *N Engl J Med, 376*(7), 613-615.

Murphy-Chutorian, E., Doshi, A., & Trivedi, M. M. (2007). *Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation.* Paper presented at the 2007 IEEE intelligent transportation systems conference.

Myer, M. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1) 131-144.

Nirkin, Y., Keller, Y., & Hassner, T. (2019). *FSGAN: Subject agnostic face swapping and reenactment.* Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision.

Renza, D., Martinez, E., & Arquero, A. (2012). A new approach to change detection in multispectral images by means of ERGAS index. *IEEE Geoscience and Remote Sensing Letters, 10*(1), 76-80.

Ribaric, S., Ariyaeeinia, A., & Pavesic, N. (2016) De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, *47*, 131-151.

Saeed, A., Al-Hamadi, A., & Ghoneim, A. (2015). Head pose estimation on top of Haar-like face detection: A study using the Kinect sensor. *Sensors, 15*(9), 20945-20966.

Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications, 7*(3), 8-18.

Shao, R., Lan, X., & Yuen, P. C. (2018). Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security, 14*(4), 923-938.

Shen, W., Sun, H., Cheng, E., Zhu, Q., Li, Q., & Shen, W. (2012). Effective driver fatigue monitoring through pupil detection and yawing analysis in low light level environments. *International Journal of Digital Content Technology and its Applications, 6*(17).

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. *Advances in Neural Information Processing Systems, 32*.

Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Yuan, L., . . . Yu, N. (2020). MichiGAN: Multi-input-conditioned hair image generation for portrait editing. *arXiv preprint arXiv:2010.16417*.

Tolba, A. (2019). Content accessibility preference approach for improving service optimality in internet of vehicles. *Computer Networks, 152*, 78-86.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131-148.

Varghese, R. R., Jacob, P. M., Jacob, J., Babu, M. N., Ravikanth, R., & George, S. M. (2021, October). An integrated framework for driver drowsiness detection and alcohol intoxication using machine learning. In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 531-536). IEEE.

Wang, Z., & Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters, 9*(3), 81-84.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11).

Zhu, S., Urtasun, R., Fidler, S., Lin, D., & Change Loy, C. (2017). *Be your own Prada: Fashion synthesis with structural coherence.* Paper presented at the Proceedings of the IEEE international conference on computer vision.