HCMUS at MediaEval 2021: Facial Data De-identification with Adversarial Generation and Perturbation Methods

Minh-Khoi Pham*[1,3], Thang-Long Nguyen-Ho*[1,3], Trong-Thang Pham*[1,3], Hai-Tuan Ho-Nguyen[1,3], Hai-Dang Nguyen[1,2,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM, [2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{pmkhoi,nhtlong,ptthang,nhhtuan,nhdang}@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

The 2021 MediaEval Multimedia Evaluation introduces a new data de-identification task, which goal is to explore methods for obscuring driver identity in driver-facing video recordings while maintaining visible human behavioral information. Interested in the challenge, our HCMUS team participate in searching for different ideas to tackle the problem. We propose two novel approaches as our main contribution: one is based on generative adversarial networks and the other is based on adversarial attacks. Moreover, a specific combination of evaluation metrics is also included for the later method for a fair comparison. The source code is available at https://github.com/kaylode/mediaeval21-drsf

1 INTRODUCTION

Car accidents are an urgent problem for many countries today. Currently, we develop vehicles to become safer, more durable, but the number of accidents is still worrisome. In [3], it is shown that driver-related problems (e.g. distraction, emotionally agitated, fatigue) account for 90% of crashes. These findings will help governments, driver educators, vehicle companies, and the public better understand the situation so that appropriate measures can be taken.

With the desire to study driver behavior, it is required to collect more driving data. This leads to concerns about privacy and security. At the "Driving Road Safety Forward: Video Data Privacy" competition, we need to perform de-identification on SHRP2 dataset [5]. The dataset shows the drivers' faces, bodies, genders, and behaviors. We aim to de-identify such that we can keep as much information as possible for the behavioral experts.

In this work, we want to focus on hiding their full face, which is the most important identifier in the given dataset. Specifically, we propose two approaches:

- A simple process that swaps the driver's face with an anonymous face to keep the most facial information.
- An adversarial pipeline to perturbate the face identity while preserving main facial attributes in form of embedding features. This approach is followed by specific evaluation functions for appropriate assessment.

* Equal contribution

MediaEval'21, 13-15 December 2021, Online

2 METHOD

2.1 Run 01 - Face Swapping

In this run, we implement the idea of swapping face to hide the real face of the driver while keeping all other facial features like gaze, eyes, mouth, nose, and head pose. The proposed procedure is shown in Figure 1. We first extract face from the given RetinaFace [6] detection results. Then, we use the swapping method from [7] to swap between an anonymous face identity and the driver's face. Finally, we bring back the swapped face to the original video, by using an overlay module. The implemented overlay module in our work is the overlay feature of a third-party tool (e.g. `ffmpeg`).

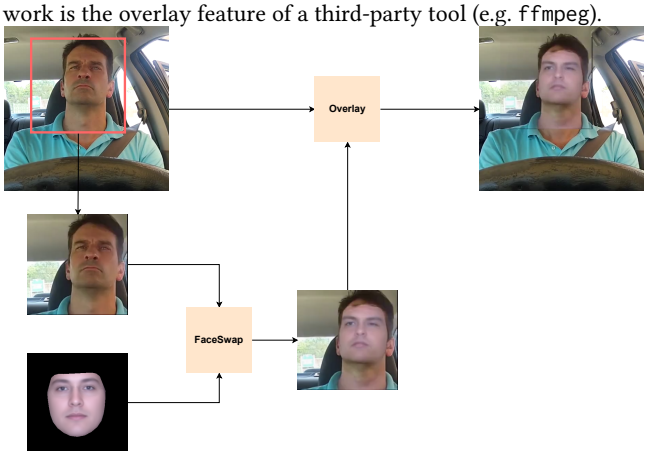


Figure 1: An illustration of overall pipeline for our Face Swapping method

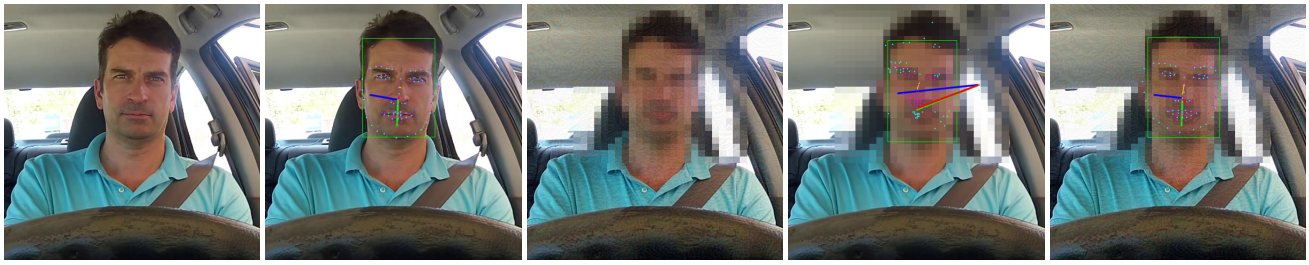
2.2 Run 02 - Adversarial Attack

This run focus on hiding a person's identity in the image from human view and preventing unauthorized deep vision algorithms from extracting useful information while ensuring correct prediction for authorized algorithms only. In our research, we only consider the position with rotation of the human face and its eye gaze vector as principal information for studying the person's action and behavior.

The specific approach is a process consisting of two main steps:

(1) Safeguard identified information from being inferred by unauthorized models.
(2) Guarantee that the model with a defined set of weights can extract information with low error.

In the first step, we apply a simple identity masking technique - pixelation and blurring to anonymize the driver's faces. This step



Original images | Prediction on original images | De-identified images | Unauthorized prediction on de-identified images | Authorized prediction on de-identified images

Figure 2: Adversarial de-identification results for run 2 (green box, dots, 3D RGB axis, yellow vector indicates head localization, face landmarks, head pose, and gaze vector respectively)

provides strong perturbation to the original image such that the faces may not be identified by humans nor any vision models.

Secondarily, to ensure that the hidden attributes, which are the bounding box, landmarks, and gaze vector of the face, can be revealed only to model with a defined set of weights, we utilize Iterative Fast Gradient Sign Method (I-FGSM) [4].

In general, I-FGSM works by exploiting the gradient of the cost function with respect to the input image to create a new image that maximizes the loss such that it drives the model's outputs towards the desired target. In our case, I-FGSM is used in the opposite trend to the common goal, which is to modify the input image and minimize the loss, resulting in changing the model prediction on the adversarial sample from false to true.

The targeted models (and their default model weights) which we choose to attack are listed below:

- For face detection, we experiment on the Retina Face [2] and MTCNN [8]
- For facial landmark detection, we explore the 2D-FAN [1]
- For gaze vector, we use a simple ResNet pretrained on ETH-XGaze [9] and MPII-Gaze [10] datasets.

We carry out the backpropagation process and find the corresponding quantity to change on the image. Our objective is to minimize function as described follows:

$$L = L_B + L_{LM} + L_G$$

Where:

- L_B is the box proposal loss of the detection model.
- L_{LM} is the L2 error between the predicted heatmap and the ground truth heatmap of the landmarks estimator.
- L_G is the L2 error between the gaze vector predicted by the model and the true gaze vector.

With the proposed loss function, we compute the network gradient and use it as a perturbation to update the current input image.

$$X_{N+1}^{adv} = Clip_{X,\epsilon}\{X_N^{adv} + \alpha sign(\nabla_x L)\}[4]$$

Given the $X_0^{adv} = X$ the raw input image, we iteratively add the perturbation to X until L becomes smaller than a predefined threshold or until N meets the maximum number of iterations.

In the concept described above, the goal of the problem is to ensure that the facial attributes extracted by authorized models between the original image and its de-identified version have a slight deviation. Therefore, we propose the following assessment method, which indicates whether the attributes are well hidden or not:

$$DiffScore_m(v_a, v_o) = \frac{1}{N}\sum_{i}^{N}(1 - IOU_B(f_{ai}, f_{oi})) + d_{LM}(f_{ai}, f_{oi}) + \Theta_G(f_{ai}, f_{oi})$$

Where v_a, v_o is the adversarial and raw video sequence consisting of N frames. f_{ai}, f_{oi} is the predictions of model m on two adversarial and original frames at the same timestamp i respectively. The IOU_B is the Intersect-over-Union between the two faces location. The d_{LM} is measured based on the euclidean distance between landmark points of the faces, Θ_G is the angle between two gaze vectors. The final $DiffScore$ is calculated by summing all differences between pairs of frames across the timestamp.

Consequently, we expect $DiffScore$ is low for authorized algorithms and higher $DiffScore$ for unauthorized ones.

3 EXPERIMENTS AND RESULTS

In the second run, we pass consecutive video sequences as a batch with a size equal to 64 into our de-identification pipeline to generate perturbated videos whose facial attributes have been hidden. We perform a full adversarial pipeline on 720 videos from the SHRP2 dataset and demonstrate some visual results as shown in Figure 2.

4 CONCLUSION AND FUTURE WORKS

Conclusively, we present different strategies to address the data privacy issues for MediaEval Challenge 2021. In the future, we aim to study the performance of our adversarial attack for the information preservation method on several deep vision models regarding facial attributes. We are intent to analyze our proposed metrics on these experiments as well.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.116

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. (2019). arXiv:cs.CV/1905.00641

[3] Thomas A. Dingus, Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences* 113, 10 (2016), 2636–2641. https://doi.org/10.1073/pnas.1513271113 arXiv:https://www.pnas.org/content/113/10/2636.full.pdf

[4] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. (2017). arXiv:cs.CV/1607.02533

[5] Transportation Research Board of the National Academy of Sciences. 2013. The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset. (2013).

[6] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE.

[7] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2020. Motion Supervised co-part Segmentation. *arXiv preprint* (2020).

[8] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503. https://doi.org/10.1109/lsp.2016.2603342

[9] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. (2020). arXiv:cs.CV/2007.15837

[10] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. (2017). arXiv:cs.CV/1711.09017