Building an Adversarial Attack Hat to Fool Facial Recognition

Morgan Frearson

Kien Nguyen

Science & Engineering Department, Queensland University of Technology

2 George St, Brisbane City QLD, Australia

morgan.frearson@connect.qut.edu.au

k.nguyenthanh@qut.edu.au

Abstract— The use of deep learning for human identification and object detection is becoming ever more prevalent in the surveillance industry. These systems have been trained to identify human body's or faces with a high degree of accuracy. However, there have been successful attempts to fool these systems with different techniques called adversarial attacks. This paper presents an adversarial attack using infrared light on facial recognition systems. The relevance of this research is to exploit the physical downfalls of deep neural networks. This demonstration of weakness within these systems are in hopes that this research will be used in the future to improve the training models for object recognition. A research outline on infrared light and facial recognition are presented within this paper. A detailed analyzation of the current design phase and future steps of the of the project are presented including initial testing of the device. Any challenges are explored and evaluated such that the deliverables of the project remain consistent to its timeline. The project specifications may be subject to change overtime based on the outcomes of testing stages.

Keywords— Adversarial attack, infrared light, physical perturbations, facial recognition

I. INTRODUCTION

Adversarial attacks aim to move an object's class across the decision boundaries of a DNN causing that object to be misclassified. This form of attack exposes fundamental blind spots in the training algorithms of DNN's [1]. A major adoption of these object detection systems by the surveillance industry (CCTV) has put pressure on the accuracy. Facial identification and human detection are two of the most prevalent and sort after features for a CCTV network.

Specifically, facial recognition systems have demonstrated an accuracy of 99.63% when identifying individuals [2]. To yield this result clean input data with no malicious intention from the individual is processed by the FR system. Several researchers have presented techniques to attack these systems via physical adversaries. Work done by Zhu

et al. [3] exploited flaws in FR training polices when they managed to decrease the confidence score of an individual using targeted makeup on the face. However, the most prevalent research towards fooling FR systems has been done using IR light. Zhou et al. [3] and Yamada et al. [4] both published successful papers utilizing IR light to misclassify FR. Their methods exhibited the strongest form of attack for the smallest resources required.

It was established that an IR adversarial attack on FR would be ideal as the project scope. The following paper explores the design, build and initial testing of an IRH. A wearable hat that uses LED's to project IR light onto the attacker's face with the intention to misclassify the individual on the FaceNet system. The hope is that exposing the downfalls of FR in this project will lead to a better understanding that can be made to prevent such attacks in the future.

The current progress of the IRH and a breakdown of each phase is discussed within this paper. Key challenges encountered, limitations and procedures of the system are also discussed. Finally, a comment on the timeline of the project such that future development and papers are also addressed.

II. IR AND FR LITERATURE REVIEW

A. IR Systems

IR light has a wavelength that causes it to be invisible to human eyes. However, IR light is detected by cameras as their sensors have a wider frequency range [3]. This results in photos and videos being susceptible to IR sabotage when it is present in the field of vision. An attack from IR can distorted images, making it an adversary to FR.

Yamanda et al. [4] invented a wearable device which makes the human face undetectable in capture images. Their type of adversarial attack relied on attaching IR LED's to a set of glasses (Fig. 1). The attack generated an IR noise signal between 800 to 1000 nm. This spectrum was ideal as they found that most digital cameras operate between 200 to 1100 nm, while humans can only see 380 to 780 nm. Because of the perceived noise generated the facial recognition system was unable to detect a face. This research showed an intuitive way to fool FR however, the current reliability of the work could be scrutinized as facial detection and cameras have improved dramatically since 2012.

Rather than facing the IR light toward the camera Zhou et al. [3] took an approach that projected IR light on to the face. The work of Zhou et al. [3] launched two possible types of adversarial attacks on facial recognition systems. Both forms of attack used IR LED's placed inconspicuously in the brim of a hat (Fig. 2). The IR light was projected onto the attacker's face to either *dodge* the recognition system or *impersonate* someone else. Their method used white-box testing so that light spots (perturbations) could be optimized through the loss function based on the variables of; shape, colour, size and brightness. For the *dodge* attack the aim was to increase the loss between the attackers feature vectors and the threshold to classify said features. For the *impersonate* attack the aim was more detailed. The system initially searched a database to find an adversary (face) that can be impersonated by the attacker (70% chance of success at this stage). The optimizer then used the technique previously talked about to calculate the necessary perturbations, with the goal to bring the attackers feature vectors into the adversaries. The researchers noted the method was not as reliable under circumstances where the attack was moving their face around. Nevertheless, the findings revealed strong results that this form of attack can fool FR.

Fig. 1 Wearable IR LED's glasses also known as 'Privacy Visor'. The lack of squares surrounding the face in (b) denotes a lower confidence level on the system. [3]

Fig. 2 Wearable IR LED hat. The light is projected onto the face from the three LED's on the brim of the hat. [4]

B. FR Systems

DNN are large neural networks organized into layers of neurons, corresponding to successive representations of the input data [5]. Each neuron represents an individual computing unit that applies an activation function to the input before passing it onto the next neuron. The architecture of a network relies on weights and biases that characterize strengths of relationships between neurons. A convolutional neural network (CNN) is a type of DNN specific for computer vision tasks. A popular type of CNN for FR is Google's FaceNet system. The system promises excellent results with an accuracy of 99.63% on the LFW dataset [2]. As stated by Schroff et al. [2], "*FaceNet is a system that directly learns a mapping from face images to a*

compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented using standard techniques with FaceNet embeddings as feature vectors". Once the L2 distance between two faces has been measured the system employs a triplet loss function. This minimizes the distance between a positive anchor while maximizing distances to a negative anchor (Fig. 3). The aim is to make the squared distance between similar faces smaller and the squared distance between dissimilar faces larger inside the feature space. In other FR systems the loss encourages all faces of one identity to be projected onto a single point in the feature space. Whereas, triplet loss enforces margins between each pair of faces from one identity thus there is stronger discriminability to other identities. Further details on the implementation of the FaceNet system corresponding to the project is addressed in section III. Research Progress.

Fig. 3 Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity. [2]

III. RESEARCH PROGRESS

From the literature it was established that an IRH would be built to fool a facial recognition system. The Research Proposal document outlines the reasoning behind this choice of adversarial attack. The initial design concept for the IRH expands on the work done by Zhou et al. [3] (Fig. 2), in an attempt to make this device more robust the attackers' movement.

During the initial design phase, a theoretically logically idea was decided on for the IRH. Zhou et al. [3] device struggled to lower the confidence score of the FR system when the attacker moved their face.

In order to maximize the adversarial abilities of the IRH against attacker's movement, two new LED's were proposed. These are primarily positioned to target the sides of the face, under the assumption that this is where Zhou et al. [3] design was at its weakest. Figure 4 shows the positioning of these new LEDs on Zhou et al. [3] original design. Incorporating Velcro on the brim of the hat will allow for position adjustments of the LEDs during the refinement stage.

Fig. 4 Initial design of the IRH. Two extra LED's added to target the side of

Digital cameras can perceive wavelengths between 200 nm to 1100 nm, while the IR spectrum is above 700nm [3]. For five LEDs there was three possible choices 850 nm, 950 nm or 1050 nm wavelength. The 850 nm was chosen because its peak luminescence balanced well across RGB on a digital camera sensor (Fig. 5). The operating point of the 850 nm LED is a $V_{\text{forward}} 1.5 \sim 2 \text{ v}$ and $I_{\text{Recommended}} = 1400 \text{ mA}$. For this, one 9 V battery will support all five LED's as $V_{\text{forward}} = 1.7 \text{ v}$. The circuit design is shown in Figure 6. The circuit has been implemented in series with a current limiting resistor of 0.33Ω [6]. This will provide each LED's with the same voltage and current resulting in equal luminescence. These design elements can be adjusted during the refinement stage in conjunction with the results returned from the FR system.

Fig. 5 Camera sensor wavelength of different colours VS Quantum Efficiency (the efficiency of converting electrons to photons) [4]

Fig. 6 LT Spice circuit diagram for IRH.

In order to determine the adversarial positioning of the light spots on the attacker's face, a perturbation optimizer is designed. The aim of the LPO is to take an initial input photo of the attacker with the IRH off, from here the LPO calculates the optimal positioning for each light spot in order to minimize the confidence score of the detection. After this the attacker positions the LED's on the IRH to match the results from the LPO. To run this system in black-box conditions the attacker assumes no knowledge of internal workings of the system being used (FaceNet FR) and can only access the returned confidence scores. Figure 7 demonstrates how to use the LPO.

Fig. 7 LPO working diagram. The LPO returns the centre $[px, py]$ of each light spot on the face.

The attackers aim is to ultimately lower their confidence score as an untargeted attack on the FaceNet FR system. To do this without the internal knowledge of the system or its specifically trained model is difficult. Instead, an assumption is made that if we can lower the confidence score of detecting a face then it would also deter recognition of a face. For the project two systems are setup; the first is a FaceNet system trained to recognize the attacker, the second is a general OpenCV system trained to detect faces.

The FaceNet system is the FR. This is built in Python by *Aras Futura*. The system is trained on a random selection of 1150 subjects from the LFW dataset and the attacker (me). Due to limitations with the computing environment it was not possible to train on more subjects as processing time increased dramatically.

The OpenCV system is the FD. This is built in Python by *Adrian Rosebrock*. The system uses a pretrained model that detects faces. This FD is the system that the LPO is going to be built on top of.

The idea is to achieve adversarial attack black-box transferability from the FD to FR. The LPO uses the confidence scores of the FD to find the optimal positioning of light spots to render an attack. The attacker then adjusts the LEDs on the IRH to mimic the LPO's results and attempts the attack on the FR system. Figure 8 visualizes this process.

Fig. 8 IRH Testing Process.

Due to time constraints of the project the variable parameters were kept minimal. On the IRH the $[px, py]$ of the light spots is the only adjustable component. In research by Zhou et al. [3], the brightness and size of light spots were variables that could be altered as well. Currently the brightness and light size on the IRH stay constant and the LPO only determines positioning.

For the LPO, a perturbation spot model is devised with three factors to be as close to the real one as possible; colour, shape and opacity. The RGB colour is obtained from a sample photo of the light on the attacker's face. The shape is difficult as light contours to the surface its displayed on, for this an estimated shape from a series of samples is devised. For opacity, estimated guesses are made until the model appears like the sample photo.

Within the LPO the attacker's face is initially cropped using OpenCV tools and resized to fit the frame. The point of this step is to bring all faces to the same size no matter how far away the attacker is standing. The perturbation spot model is then applied to the face. The new photo is fed into the FD and the confidence score is returned. The LPO continues

using a simple brute force search to shift the light spot and obtain the confidence score for all possible combinations. This method is similar to what Su et al. [7] performed for the one-pixel attacks. The final step of the LPO is returning the lowest confidence score and corresponding center $[px, py]$ of each light spot.

In order to make the IRH more robust to the attacker's movements two of the five LEDs specifically target the sides of the face. For this the LPO takes in three photos, one front on, one of the left side and one of the right side of the face. The front on photo allows the LPO to use three perturbation spot models. The side on photos allow the LPO to use one perturbation spot models each. Figure 9 demonstrates the overall process of the LPO system.

Fig. 9 LPO full process.

IV. INITIAL TESTING

Building the IRH was a smooth process with limited issues. There were two adjustments made to the during the construction, current limiting resistor size and added lenses. Due to limitations with available resistor sizes the 0.33Ω was changed for a 0.47Ω , this produced $V = 1.7 \text{ v}$ and $I = 1060 \text{ mA}$ through each LED [6]. According to the LED data sheet [7] this lower current ($I_{\text{Recommended}} = 1400 \text{ mA}$) was still practical except relative luminesce decreased by $\sim 5 - 10 \%$. There was some initial concern with the angle of light emitted from the LED's as light distribution emits at 140° [7]. To concentrate this a 15° lens was added to each LED, this improved the brightness and telescoped each light spot. The final design is show in Figure 10, note that the battery is attached at the back of the hat.

Fig. 10 The IRH initial build.

Preliminary testing of the IRH produced conflicting results show in Table 1. The LEDs were randomly positioned for the photo (MacBook Pro inbuilt camera) as there was no intention of a specific attack. Very positive results came from the FD system, showing a decrease in confidence by 24.65% ~ IRH [on]. However, for the same photo the FR system increased in confidence by 15.45% ~ IRH [on]. From this, it's possible that the previous assumption of black-box transferability between the FD to FR systems may not be achievable as the results contradict. However, until official testing with the LPO is established the original assumption remains true until proven otherwise.

TABLE I
PRELIMINARY TESTING OF IRH ON FD AND FR

	IRH Off	IRH On	Confidence Difference
OpenCV FD	Confidence = 92.12%	Confidence = 67.47%	24.65%
FaceNet FR	Confidence = 77.20%	Confidence = 92.65%	15.45%

There were two observations from this test that could be noted.

- There are limitations on the amount of surrounding light. Testing must be performed in low light environments as the CMOS image sensor used on the MacBook Pro includes an Infrared Cut Filter (ICF) that attenuates infrared wavelengths until light diminishes below a certain level [8].
- The bounding box appears to shrink on both FD and FR systems when IRH [on]. The reason for this is unknown but it is possible that where the light is focused (predominantly forehead area) tightens the bounding box around it.

V. FUTURE WORK AND CONCLUSION

The research being undertaken has the objective to build an infrared device that can fool a FR system. Creating an adversarial attack by adding physical light perturbations to an attacker's face was the starting point of the design phase. An initial IRH has been constructed in its prototype phase, small adjustments are yet to be made before formal testing and verification can be undergone. Successful implementation of the device will progress the project onto its next stage, to optimize its adversarial abilities.

Another important step in the implementation and optimization of the IRH is the LPO system. This system produces a model of the optimal attack position for light on the face. The design of this has been laid out and the build is in progress. Once this is completed the final stage of continual data collection and refinement on the IRH/LPO can evolve.

Overall, it is expected that the project outcomes will be achieved and the IRH will be more robust than previous work. Data will be collected on the IRH's capacity to fool the FD when the attacker is moving. Assuming black-box transferability the FR system will experience similar outcomes. If results are undesirable the LPO system will be remodelled before possible changes to the IRH design. If results are desirable further research into white-box testing

on the FD system to increase the efficiency of the LPO will be studied. This research plays an important role in exposing weaknesses of DNN's. The more we try and exploit these systems the more other researchers can explore prevention techniques for the future.

Refer to the next paper 'Adversarial Attacks on Facial Recognition using Visible Light' by 'Morgan Frearson' to see the continuation of the work discussed in this paper.

REFERENCES

- [1] I. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *ICLR*, San Diego, 2015.
- [2] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, 2015.
- [3] T. Yamada, S. Gohshi and I. Echizen, "Use of Invisible Noise Signals to Prevent Privacy Invasion through Face Recognition from Camera Images," in *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, 2012.
- [4] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu and K. Zhang, "Invisible Mask: Practical Attacks on Face Recognition with Infrared," Cornell University Library, 2018.
- [5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy*, Saarbrücken, 2016.
- [6] Hobby-Hour, "LED Series Resistor Calculator," [Online]. Available: <https://www.hobby-hour.com/electronics/ledcalc.php>.
- [7] Shenzhen Shining Opto-electronic Co. Ltd, "SN-IR-850nm-5W-04 High Powered LED," [Online]. Available: <https://www.rcscomponents.kiev.ua/datasheets/sn-ir-850nm-5w-4chips.pdf>.
- [8] Mistral, "The Advantage of a Mechanical IR-Cut Filter Switch," 2018. [Online]. Available: <https://www.mistralsolutions.com/articles/advantage-mechanical-ir-cut-filter-switch/>.
- [9] T. Brown, D. Mané, A. Roy, M. Abadi and J. Gilmer, "Adversarial Patch," *31st Conference on Neural Information Processing Systems*, 2018.
- [10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, T. Kohno and D. Song, "Physical Adversarial Examples for Object Detectors," 2018.
- [11] S.-M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.
- [12] J. Su, K. Sakurai and D. Vasconcellos Vargas, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, VOL. 23, NO. 5, pp. 828-841, 2019.
- [13] S. Thys, W. Van Ranst and T. Goedeme, "Fooling automated surveillance cameras: adversarial patches to attack person detection," KU Leuven EAVISE, Belgium, 2019.
- [14] A. Fawzi, O. Fawzi, P. Frossard and S.-M. Moosavi-Dezfooli, "Universal adversarial perturbations," Cornell University Library, 2017.
- [15] Z.-A. Zhu, Y.-Z. Lu and C.-K. Chiang, "Generating Adversarial Examples By Makeup Attacks on Face Recognition," in *IEEE International Conference on Image Processing (ICIP)*, Taipei, 2019.
- [16] W. Ding, L. Wang, R. Huang, X. Jin and K. Y. C. Lui, CA, US Patent CA3043809A1, 2020.
- [17] J. Zhang, F. Araujo, T. Taylor and M. P. Stoecklin, US, CN Patent US20200005133A1, 2018.
- [18] R. Khandelwal, "Computer Vision — A journey from CNN to Mask R-CNN and YOLO -Part 2," 27 July 2019. [Online]. Available: <https://towardsdatascience.com/computer-vision-a-journey-from-cnn-to-mask-r-cnn-and-yolo-part-2-b0b9e67762b1>.
- [19] Dallmeier Electronic GmbH & Co. KG, *How does the IR (infrared) cut filter of a CCTV camera work?*, <https://www.youtube.com/watch?v=2BKtzpKk5Gk>, 2011.
- [20] T. Luedtke and S. Pingree, "ISHN," 2013 January 3. [Online]. Available: <https://www.ishn.com/articles/94815-dangers-of-overexposure-to-ultraviolet-infrared-and-high-energy-visible-light>.