Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study (Review)

Tejas Gaikwad, Dept. of CSE, IIT Jodhpur

I. INTRODUCTION

This paper attempts new version of adversarial attack which they termed as Adversarial light projection attack. They had discussed the ways to attempt this attack along with the restrictions, results and the logistics requirement to attempt this attack.Their experiment results demonstrate the vulnerability of face recognition systems to light projection attacks in both white box and black box attacks. In White box attacks are the adversary(person who wants to fake the model) knows the internal details of the model including the architecture and trained weight parameters whereas in black-box attack, the attacker knows the output of the model or more inputs.

The papers has discussed two methods of faking the face recognition systems and those are impersonation(method used for impersonating different enrolled users) and obfuscation(Dodging the recognition). They(authors) have used a camera, a projector and a pre-trained face recognition model FaceNet and SphereFace to perform the attack. As shown in Figure 1., a digitally generated projectable adversarial light pattern is given to a projector to project it on attackers face to fake the model. The camera, light projector, area of the face and location on which the light should fall are calibrated too to make this attack successful. The use case for such type of attacks are useful when the attacker intends to object access to a personal device protected with a target's face. Wherein obfuscation is the goal of an adversary blacklisted law enforcement agencies who wants to evade recognition in scenarios such as border crossing.

Major contribution by this paper is Investigation of real-time adversarial light projection attacks using off-the-shelf camera-projector setup on state-of-the-art face recognition systems. An efficient transformation-invariant adversarial pat- tern generation method suitable for conducting real-time adversarial light projection attacks and Demonstration of vulnerability of state-of-the-art face recognition systems to adversarial light projection at- tacks in both white-box and black-box settings.

II. OVERALL ARCHITECTURE AND SYSTEM DESIGN

A. Assumptions

Following assumptions are made in the paper.

1. Attacker is assumed to have black box or white box access.
2. The attacker uses an open-source algorithm to generate adversarial patterns to attack the black box system.
3. The attacker have access to the image of the target(Social Media can play important role in this).
4. Attacker has access to the camera to capture attacker's own face image in order to compute adversarial light pattern.
5. The attacker have either access to or have a reasonable prior knowledge of the environment where the face recognition system is to be deployed.

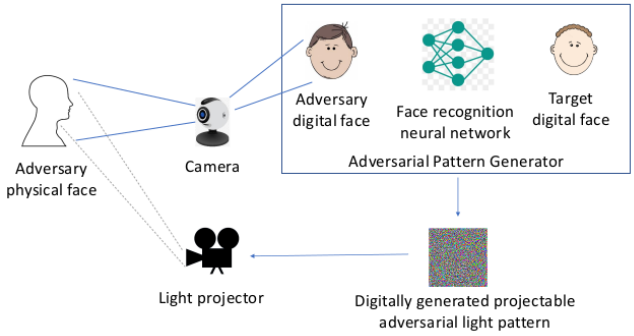

Fig. 1. Overall Architecture.

B. Practical Consideration

The adversarial light projection attacks are inherently challenging because of their unconstrained nature. There are some sections that are required to be considered for having a successful test. The areas are Environmental Factors(ambient and positional lighting, and their interplay with projected light), Intra-Adversary facial variations(due to slightly movements of the attacker) and Intra-Target variations(as attacker would not typically have access to the enrolled images of the target in the deployed face recognition system.

C. Attack Setup Calibration

There are 2 main calibration steps important for success of the attack. Those calibrations are position calibration and color calibration. As the attacker should be in view of both camera ad the projector, some positional calibration are required to be performed. These are further of 2 types, manual calibration(the attacker manually annotates the some small number corresponding points between the two views) and automatic (a facial landmark algorithm is used to detect the facial landmarks from the two view.

Whereas in color calibration, the main aim is to generate a relationship between digital and physical domain.

Below Algorithm summarizes the proposed transformation-invariant pattern generation method. The method takes as input the image x of the adversary and the image of the target y, and outputs the adversarial pattern x adv . The transformations used depend on the invariance objective (e.g., affine, perspective, photo-metric or others).A brightness term sampled from a normal distribution is used during each iterative update for obtaining invariance to slight illumination changes. Furthermore, a binary mask can be used to constrain the facial region for which the adversarial pattern is generated. In below equation, x^{adv} is the adversarial digital image patch which the attacker is targeting.

$$\mathrm{x}^{adv} = argmin \sum_{i=0}^{k-1} (w_i((\tau_i(x) + \Delta x, y)), s.t. \|\Delta x\|_p)$$

III. Experiment

Impersonation

A different face image of the target (also obtained from the web or a database) is assumed to be enrolled in the face recognition system to be attacked. Impersonation attempts are made using different subject pools in two scenarios, those are fixed Target (23 and 21 out of 25 samples on FaceNet and SphereFace) and on selected target 14 and 12 attempts were successful out of 15 on FaceNet and SphereFace).

Obfuscation

All 10 obfuscation attempts were successful for whitebox attack on FaceNet and SphereFace. wehre as in blackbox attack 7 out of 10 went successful.

IV. Results

Fig. 2. Impersonating Results

V. Conclusion

The primary results of the preliminary experiments conducted by them using two open-source and one commercial face recognition system on a pool of 50 subjects demonstrate the vulnerability of face recognition systems to light projection attacks in both white-box and black-box attack settings.