November 2022

3D ADVERSARIAL FACE TARGETS

Visa

3D ADVERSARIAL FACE TARGETS

VISA

INVENTOR:

SUNPREET SINGH ARORA

TECHNICAL FIELD

[0001] The present subject matter relates to face recognition systems and, more particularly, to designing and fabrication of three-dimensional adversarial face targets for testing face recognition systems.

BACKGROUND

[0002] The global Artificial Intelligence (AI) space has seen a rapid growth due to increased utilization of machine learning and deep neural networks across industries. However, recent research has demonstrated the vulnerability of machine learning-based systems against adversarial machine learning techniques. More specifically, adversarial machine learning, a technique that attempts to fool models with adversarial examples, is a growing threat in the AI research community. The most common reason to generate adversarial example is to cause a malfunction in an AI model.

[0003] An adversarial example is an input sample that is intentionally designed to cause a machine learning model to make a mistake in its predictions. In other words, an adversarial example is an input provided by an attacker with specifically crafted perturbation that causes an AI model to make mistakes. Often the perturbations are minimalistic such that a human observer cannot distinguish adversarial example from a benign input example. In general, adversarial examples are crafted either in digital domain or physical domain specifically with the aim of circumventing an AI model and presenting of such adversarial examples is referred to as an adversarial attack. As such, attacks conducted with adversarial examples in digital domain (e.g., on a software machine learning model) are referred to as digital attacks and attacks where adversarial examples are input directly to a system's physical interface, for example, providing adversarial examples to a face recognition system via a sensing mechanism is referred to as physical attacks. Such adversarial attacks might entail presenting an AI model with test input which is significantly different from training data distribution, or such maliciously designed data may be used to deceive an already trained model.

[0004] Conventionally, adversarial example generation methods focus on generating adversarial examples in the digital domain. The assumption in this case is that the adversary can provide a digital input to the AI model that he/she intends to attack. While digital attacks

can be used to attack open machine learning APIs, they cannot be used to attack end-to-end systems (*e.g.,* face recognition systems) that are deployed in real world. This is because deployed systems often rely on open capture or sensing mechanisms that obtain input from the physical world. Furthermore, deployed systems inherently employ mechanisms to safeguard digital transmission and storage channels from digital attacks, such as man-in-the-middle attack. Physical attacks on face recognition systems, on the other hand, directly provide adversarial input via the sensing mechanism and thus can be used to attack end-to-end systems.

[0005] In view of the susceptibility of face recognition systems to physical adversarial attacks, it is important to comprehensively test the security of real-world face recognition systems under various adversarial configurations before deployment.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] The accompanying drawings, which are incorporated in and constitute a part of this disclosure, illustrate exemplary embodiments and, together with the description, serve to explain the disclosed principles. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The same numbers are used throughout the figures to reference like features and components. Some embodiments of device or system and/or methods in accordance with embodiments of the present subject matter are now described, by way of example only, and with reference to the accompanying figures, in which:

[0007] FIG. 1 illustrates a simplified representation of an environment related to face recognition for implementing embodiments consistent with the present disclosure.

[0008] FIG. 2 illustrates a system for generating 3D digital adversarial face targets.

[0009] FIG. 3 depicts a flowchart illustrating a process for generating 3D target faces from a plurality of images.

[0010] FIG. 4 depicts a flowchart illustrating a process for generating 3D digital adversarial face targets.

[0011] FIG. 5 depicts a flowchart illustrating a process for manufacturing a 3D adversarial face target.

[0012] FIG. 6 illustrates a block diagram of an exemplary computer system for implementing embodiments consistent with the present disclosure.

[0013] The figures depict embodiments of the disclosure for purposes of illustration only. One skilled in the art will readily recognize from the following description that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the disclosure described herein.

DESCRIPTION OF THE DISCLOSURE

[0014] In the present document, the word "exemplary" is used herein to mean "serving as an example, instance, or illustration." Any embodiment or implementation of the present subject matter described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other embodiments.

[0015] While the disclosure is susceptible to various modifications and alternative forms, specific embodiment thereof has been shown by way of example in the drawings and will be described in detail below. It should be understood, however that it is not intended to limit the disclosure to the particular forms disclosed, but on the contrary, the disclosure is to cover all modifications, equivalents, and alternative falling within the spirit and the scope of the disclosure.

[0016] The terms “comprises”, “comprising”, or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a setup, device or method that comprises a list of components or steps does not include only those components or steps but may include other components or steps not expressly listed or inherent to such setup or device or method. In other words, one or more elements in a device or system or apparatus proceeded by “comprises… a” does not, without more constraints, preclude the existence of other elements or additional elements in the device or system or apparatus.

[0017] The terms "an embodiment", "embodiment", "embodiments", "the embodiment", "the embodiments", "one or more embodiments", "some embodiments", and "one embodiment" mean "one or more (but not all) embodiments of the invention(s)" unless expressly specified otherwise.

[0018] The terms "including", "comprising", “having” and variations thereof mean "including but not limited to", unless expressly specified otherwise.

[0019] The term ‘digital adversarial face target’ as used herein refers to an electronically synthesized face structure that is manufactured with, intentional feature perturbations. These feature perturbations may be achieved by intentionally introducing known adversarial patterns in synthesized face structures that may cause a machine learning model to make a false prediction. The digital adversarial face target may be manufactured using printing techniques to generate a physical structure referred to herein as ‘adversarial face target’. In general, these adversarial face targets are adversarial examples that are used to conduct physical adversarial testing of end-to-end face recognition systems in a repeatable manner. Generating adversarial face targets with known ground truth adversarial configurations (i.e., known adversarial patterns) as will be explained in detail with reference to FIGS. 1-4.

[0020] **FIG. 1** illustrates a simplified representation of an environment 100, in which at least some example embodiments of the disclosure can be implemented. The environment 100 exemplarily depicts a face recognition system 108 that controls access to a service/location, for example, access to medical record room of hospital. It shall be noted that the face recognition system 108 depicted in FIG.1 is for exemplary purposes and the face recognition system 108 may indeed control access to personal devices, vehicles, and several areas/premises such as, offices, residences, military establishments and government organizations.

[0021] The face recognition system 108 is configured to identify or verify a person from an image or a video frame. Accordingly, an imaging apparatus 104 is configured to capture one or more images of a user requesting access to the service/location. For example, the user 102 may intend to access the service/location and as such, the imaging apparatus 104 captures an image of the user 102. It shall be noted that although the imaging apparatus 104 is shown as a separate entity for exemplary purposes, the imaging apparatus 104 may be integrated within the face recognition system 108. For example, the face recognition system 108 may be implemented within a smartphone for controlling access of the smartphone or different software applications installed on the smartphone and the imaging apparatus 104 may be embodied within the smartphone to capture images.

[0022] The imaging apparatus 104 is configured to forward captured images to the face recognition system 108. The face recognition system 108 is configured to identify or verify the

identity of the user 102 based on one or more facial features determined from the captured images. In an identification scenario, the face recognition system 108 compares the facial features with a plurality of facial features stored in a database 112 to identify the user 102 and in a verification scenario, the facial features are mapped to a corresponding profile of a user 102 in the database 112 to verify the identity of the user 102. Accordingly, the database 112 is populated with a plurality of profiles of users and corresponding facial features. As such, the face recognition system 108 utilizes one or more AI models trained to authenticate an identity of the user 102. In an example, machine learning models are trained to identify/verify the identity of the users. In another example, the deep neural networks, for example, convolutional neural networks, are trained to identify/verify the identity of the user 102. In other words, the facial features populated in the database 112 may be used to train the face recognition system 108 to authenticate the identity of the person requesting access to a location/service.

[0023] In an example scenario, during the training phase, the face recognition system 108 may be susceptible to an adversarial attack that create data variations in the training data i.e., facial features of the user 102. The term ‘adversarial attack’ as used herein refers to a malicious attempt in which an adversarial pattern is intentionally introduced to cause the face recognition system 108 to make a mistake in identification/verification. The adversarial attack may be a digital attack which implements modification in the image captured by the imaging apparatus 104 or may be a physical attack which modifies the physical appearance of the face of the user 102 prior to capturing the image of the user 102 by the imaging apparatus 104. In one example, the digital attack may be performed during the training phase, in which carefully crafted perturbations, called adversarial patterns may be used to modify the training data provided to the machine learning model/deep neural network. In another example, the adversarial patterns may be introduced in the images captured by the imaging apparatus 104 which modifies the extracted features of the user 102. As such, the face recognition system 108 is susceptible to attacks that cause the face recognition system 108 to make an error in authenticating the identity of the user 102.

[0024] Various embodiments of the present invention disclose a system 200 employing a method for generating 3D digital adversarial face targets. More specifically, the system 200 generates 3D digital adversarial face targets by manipulating viewpoints and introducing adversarial perturbations in a specific location of an electronically generated 3D target face. In general, the 3D digital adversarial face targets are efficiently crafted by introducing known

ground truth adversarial configurations (i.e., known adversarial patterns) in electronically generated 3D target faces. Moreover, the 3D digital adversarial face targets m ay be fabricated to generate adversarial face targets which are physical structures used to systematically test the face recognition system 108 under adversarial configurations for physical adversarial attacks. In general, such synthesized adversarial face targets may be used to comprehensively test the security of real-world face recognition systems under various adversarial configurations before deployment. The system 200 for generating digital adversarial face targets is explained in detail next with reference to **FIG. 2**.

[0025] **FIG. 2** illustrates the system 200 for generating 3D digital adversarial face targets. In an embodiment, the system 200 may be a stand-alone computer system or processor capable of processing a plurality of images for generating digital adversarial face targets. In another embodiment, the system 200 may be embodied within the face recognition system 108 and may be configured to generate digital adversarial face targets. Further, it shall be noted that the face recognition system 108 is shown for exemplary purposes and embodiments of the present invention may be practiced with systems other than a face recognition system, for example, an object recognition system that verifies/identifies objects. As such, the system 200 may be configured to generate 3D adversarial object targets. However, for exemplary purposes, the description is limited to the 3D adversarial face targets.

[0026] The system 200 is depicted to include a processor 202, a memory 204, an Input/Output module 206, and a communication interface 208. It shall be noted that, in some embodiments, the system 200 may include more or fewer components than those depicted herein. The various components of the system 200 may be implemented using hardware, software, firmware or any combinations thereof. Further, the various components of the system 200 may be operably coupled with each other. More specifically, various components of the system 200 may be capable of communicating with each other using communication channel media (such as buses, interconnects, etc.). It is also noted that one or more components of the system 200 may be implemented in a single server or a plurality of servers, which are remotely placed from each other.

[0027] In one embodiment, the processor 202 may be embodied as a multi-core processor, a single core processor, or a combination of one or more multi-core processors and one or more single core processors. For example, the processor 202 may be embodied as one or more of

various processing devices, such as a coprocessor, a microprocessor, a controller, a digital signal processor (DSP), a processing circuitry with or without an accompanying DSP, or various other processing devices including, a microcontroller unit (MCU), a hardware accelerator, a special-purpose computer chip, or the like. The processor 202 includes a 3D target face generator 210, a viewpoint generator 212, an adversarial input generator 214, and a texture manipulator 216 which are explained in detail later.

[0028] In one embodiment, the memory 204 is capable of storing machine executable instructions, referred to herein as instructions 205. In an embodiment, the processor 202 is embodied as an executor of software instructions. As such, the processor 202 is capable of executing the instructions 205 stored in the memory 204 to perform one or more operations described herein. The memory 204 can be any type of storage accessible to the processor 202 to perform respective functionalities, as will be explained in detail with reference to FIGS. 2 to 5. For example, the memory 204 may include one or more volatile or non-volatile memories, or a combination thereof. For example, the memory 204 may be embodied as semiconductor memories, such as flash memory, mask ROM, PROM (programmable ROM), EPROM (erasable PROM), RAM (random access memory), etc. and the like.

[0029] In an embodiment, the processor 202 is configured to execute the instructions 205 for: (1) synthesizing a set of 3D target faces based on a plurality of images, (2) capturing a set of 2D viewpoint configurations for each 3D target face based on a set of 2D viewpoints, (3) generating at least one adversarial pattern for each 2D viewpoint configuration of the set of 2D viewpoint configurations, and (4) generating a set of 3D digital adversarial face targets based on the set of 2D viewpoint configurations and the at least one adversarial pattern.

[0030] In an embodiment, the I/O module 206 may include mechanisms configured to receive inputs from and provide outputs to peripheral devices such as, an operator of the system 200. The term ‘operator of the system 150’ as used herein may refer to one or more individuals, whether directly or indirectly, associated with testing the quality of the face recognition system 108. To enable reception of inputs and provide outputs to the system 200, the I/O module 206 may include at least one input interface and/or at least one output interface. For examples, adversarial patterns may be provided by the operator of the system 200 using the keyboard/mouse and digital adversarial targets may be displayed on a liquid crystal display (LCD) display. Examples of the input interface may include, but are not limited to, a keyboard,

a mouse, a joystick, a keypad, a touch screen, soft keys, a microphone, and the like. Examples of the output interface may include, but are not limited to, a display such as a light emitting diode display, a thin-film transistor (TFT) display, a liquid crystal display, an active-matrix organic light-emitting diode (AMOLED) display, a microphone, a speaker, a ringer, and the like.

[0031] In an embodiment, the communication interface 208 may include mechanisms configured to communicate with other entities in the environment 100. In other words, the communication interface 208 is configured to access the plurality of images corresponding to a plurality of users. In an example, the plurality of images may be stored in the database 112 and the communication interface 208 may access the plurality of images for generating digital adversarial face targets. In another example, the plurality of images may be images captured by the imaging apparatus 104. In yet another example, the plurality of images may correspond to images scanned using the I/O module 206, for example, scanner. The images are then stored in the database 112. Further, in some example embodiments, the communication interface 208 may be configured to receive adversarial patterns from the operator of the system 200. In an embodiment, the communication interface 208 may receive a target portion selected by the operator of the system 200 for introducing the adversarial pattern. As such, the plurality of images, target portion and the adversarial pattern are received by the communication interface 208 and sent to the processor 202 which performs one or more operations described herein to generate digital adversarial face targets.

[0032] The system 200 is depicted to be in operative communication with a database 220. In one embodiment, the database 220 is configured to store known ground truth adversarial configurations (i.e., known adversarial patterns) that were generated for each digital adversarial face. These ground truth adversarial configurations may be used by the face recognition system 108 for evaluating performance of the face recognition system 108 before deployment. Further, the database 220 is configured to store the digital adversarial face targets that are artificially synthesized by the system 200. In general, the database 220 stores profiles of digital adversarial face targets with information corresponding to set of viewpoints and adversarial patterns that may be used for evaluating performance of the face recognition system 108. Further, the database 220 may be configured to store the synthesized target faces, different target face configuration based on viewpoints, adversarial patterns/ selected portions, texture information and the like.

[0033] The database 220 may include multiple storage units such as hard disks and/or solid-state disks in a redundant array of inexpensive disks (RAID) configuration. In some embodiments, the database 220 may include a storage area network (SAN) and/or a network attached storage (NAS) system. In one embodiment, the database 220 may correspond to a distributed storage system, wherein individual databases are configured to store custom information, such as synthesizing rules, facial features related data, adversarial pattern data, mask specifications, etc.

[0034] In some embodiments, the database 220 is integrated within the system 200. For example, the system 200 may include one or more hard disk drives as the database 220. In other embodiments, the database 220 is external to the system 200 and may be accessed by the system 200 using a storage interface (not shown in FIG. 2). The storage interface is any component capable of providing the processor 202 with access to the database 220. The storage interface may include, for example, an Advanced Technology Attachment (ATA) adapter, a Serial ATA (SATA) adapter, a Small Computer System Interface (SCSI) adapter, a RAID controller, a SAN adapter, a network adapter, and/or any component providing the processor 202 with access to the database 220.

[0035] As already explained, the communication interface 208 is configured to receive a plurality of images. The communication interface 208 forwards the plurality of images to the processor 202. The modules of the processor 202 in conjunction with the instructions in the memory 204 are configured to process the plurality of images to generate the set of digital adversarial face targets. The processor 202 is configured to forward the plurality of images to the synthetic target face generator 210.

[0036] The 3D target face generator 210 in conjunction with the instructions in the memory 205 is configured to synthesize a set of 3D target faces from the plurality of images. The operations of the 3D target face generator 210 for generating a set of target faces is explained as process steps with reference to **FIG. 3**.

[0037] **FIG. 3** depicts a flowchart 300 illustrating a process for generating 3D target faces from a plurality of images. The steps of the flowchart 300 may be performed by the 3D target face generator 210 of the processor 202.

[0038] At 302, the plurality of images $I_1, I_2, \ldots, I_n$ are received by the 3D target face generator 210.

[0041] At 304, a plurality of shape components and a plurality of texture components corresponding to a plurality of faces detected from the plurality of images are determined. In an embodiment, the target face generator 210 is configured to detect faces of people from the plurality of images $I_1, I_2, \ldots, I_n$. For example, faces $F_1, F_2, F_3, \ldots, F_m$ may be detected from the plurality of images $I_1, I_2, \ldots, I_n$. It shall be noted that each image may include more than one face and as such, the target face generator 210 detects one face from each image. The faces $F_1, F_2, F_3, \ldots, F_m$ may be detected using methods such as, knowledge-based, feature-based, template matching, or appearance-based. Some examples of algorithms that may be used for detecting faces from images include, but not limited to, Principal Component Analysis (PCA), Linear Discriminant Analysis(LDA), Independent Component Analysis(ICA), Elastic Bunch Graph Matching (EBGM), Fisherfaces, Eigen face based algorithm, Viola Jones algorithm, and the like. In one example, feature-based detection techniques locate a face in an image by identifying and extracting structural features of the face, for example, distinctive features on the surface of a face, such as the contour of the eye sockets, nose, and chin.

[0042] Further, the 3D target face generator 210 is configured to determine the plurality of shape components S (i.e., $S = s_1, s_2, \ldots, s_m$). More specifically, features of the detected face may be extracted and used to modify a 3D generic face model (i.e., 3D generic face mesh) to generate a 3D face model corresponding to the detected face (e.g., face F_1). As such, the shape components in S comprise vertices and edges of the 3D face model corresponding to the detected face F_1. As such, the plurality of texture components T (i.e., $T = t_1, t_2, \ldots, t_m$) correspond to an original texture of the plurality of faces $F_1, F_2, F_3, \ldots, F_m$. For example, texture component t_1 corresponding to face F_1 would have colour values (i.e., r, g, b values) for each vertex of the 3D face model. In general, the texture component maps each vertex in the 3D face model to a colour value and such associated is represented as texture component t_1 for the detected face F_1.

[0043] At 306, a first distribution (D_1) based on the plurality of shape components S and a second distribution (D_2) based on the plurality of texture components T are plotted. The distribution plots (i.e., first distribution D_1 and second distribution D_2) provide a visual

representation of different facial features extracted from the plurality of faces F_1, F_2, F_3, …, F_m and different texture t_1, t_2, …, t_m distributions.

[0044] At 308, a set of 3D target faces are generated based on sampling the first distribution D_1 and the second distribution D_2 using a sampling function. More specifically, a first set of parameters α are determined by sampling the distribution D_1 and a second set of parameters β are determined by sampling the distribution D_2 using the sampling function. Some examples of the sampling function f include, but not limited to, uniform function, gaussian function, and the like. In one example, the plurality of faces F_1, F_2, F_3, …, F_m (hereinafter indicated as 'F', where F = F_1, F_2, F_3, …, F_m) are represented using the plurality of shape components S and the plurality of texture components T(i.e., $F = \{S, T\}$). Assuming the distributions of the plurality of shape components S and the plurality of texture components are D_1 and D_2, respectively. As such, the first set of parameters α are determined from the distribution D_1 and the second set of parameters β may be parameterized from the distribution D_2. Accordingly, 'X' 3D targets faces (including shape components and texture component may be generated by sampling from the two distributions D_1 and D_2 using the sampling function *f* as shown below by equation (1).

$$X = f(D1(\alpha), D2(\beta))$$ ---------- Equation (1)

[0045] In general, each 3D target face in the set of 3D target faces are generated using a parameter sampled from D_1, for example, facial features from the shape component in the distribution D_1 and a parameter sampled from D_2, for example, texture information from the distribution D_2. Such, sampling ensures a wide variety of population may be accounted for in designing the set of 3D target faces F_S (i.e., $F_S = F_{S1}, F_{S2}, \ldots, F_{Sl}$, where '*l*' denotes a number of 3D target faces) to ensure comprehensive testing of the security of real-world face recognition systems under various adversarial configurations before deployment. Moreover, synthesizing 3D target faces electronically provisions options for controlling various facial features and textures, for example, the face shape and skin texture, so that they are representative of real-world population demographics.

[0046] Referring back to **FIG. 2**, the 3D target faces generated by the 3D target face generator 210 are forwarded to the viewpoint generator 212. The viewpoint generator 212 in conjunction with the instructions 205 stored in the memory 204 is configured to capture a set of 2D

viewpoint configurations V (V = V_1, V_2, …, V_n) from each 3D target face for different face configurations C (C = C_1, C_2, …, C_n). These different face configurations C (C = C_1, C_2, …, C_n) correspond to each 3D target face (for example, face F_{S1}) in a different viewpoint. The term 'viewpoint' as used herein refers to a perspective view of the 3D target face. In addition, different viewpoints of the 3D target face may also be achieved by variation of other parameters such as, but not limited to, distance from the imaging apparatus, orientation with respect to optical axis of the imaging apparatus, in-plane, and out-of-plane rotation, lighting conditions, and the like.

[0047] In one example, a viewpoint may correspond to a 30° orientation of the 3D target face F_{S1} with respect to optical axis of an imaging apparatus for capturing a 2D viewpoint configuration V_1 of the 3D target face F_{S1} and another viewpoint corresponds to 10° orientation of the 3D target face F_{S1} with respect to the optical axis imaging apparatus for capturing a 2D viewpoint configuration V_2 of the 3D target face F_{S1}. In another example, a viewpoint corresponds to a 30° orientation of the 3D target face F_{S1} with respect to optical axis of the imaging apparatus and a distance of 30 metres from the imaging apparatus for capturing a 2D viewpoint configuration V_1 of the 3D target face F_{S1} whereas another viewpoint may correspond to 10° orientation of the 3D target face F_{S1} with respect to the optical axis of the imaging apparatus and distance of 20 metres from the imaging apparatus for capturing a 2D viewpoint configuration V_2 of the 3D target face F_{S1}. In general, a position, angle, stance, lighting and other parameters may be varied to capture the different 2D viewpoint configurations for each 3D target face. It shall be noted that the above examples of some parameters (i.e., orientation and distance) considered for capturing the 2D viewpoints for each 3D target face are exemplary and a number of 2D viewpoints may be captured for each 3D target face based on a wide variety of parameters and combinations.

[0048] In an embodiment, a projection function *p* may be used for mapping each 3D target face to a 2D viewpoint for generating a 2D viewpoint configuration V1. For example, the projection function *p* may be used for mapping the 3D target face F_{S1} to the set of 2D viewpoint configurations V (i.e., V = V_1, V_2, …, V_n) based on the set of 3D face configurations C_1, C_2, …, C_n corresponding to various viewpoint for testing. If the projection function is *p*, and the set of viewpoints created for the 3D target face is mapped based on the set of 3D face configurations C_1, C_2, …, C_n corresponding to various viewpoint, then the 2D viewpoint

configuration for each 2D viewpoint configuration is determined based on the below equation 2.

$$V = p(C(Ti)) \quad \text{---------- (Equation 2)}$$

[0049] More specifically, the projection function *p* enables mapping the 3D locations and textures of the 3D target face to corresponding 2D viewpoint configurations based on viewpoint. An example of the projection function is a geodesic distance preserving function such as, multi-dimensional scaling (MDS). The set of 2D viewpoint configurations for each 3D target face is sent to the adversarial input generator 214.

[0050] The adversarial input generator 214 in conjunction with the instructions 205 in the memory 204 is configured to synthesize adversarial perturbation (also referred to herein as adversarial patterns) for each 2D viewpoint configuration in the set of 2D viewpoint configurations. As already explained, these adversarial patterns are generated to modify at least a portion of the 2D viewpoint configuration corresponding to the 3D target face. More specifically, these adversarial patterns modify/perturb one or more pixels intentionally in the 2D viewpoint configuration for deceiving face recognition systems such as, the face recognition system 108. The adversarial patterns may be generated for impersonation of a particular subject (i.e., human) or obfuscation to evade recognition. Accordingly, at least one adversarial pattern A_1 is synthetically generated by the adversarial pattern generator 214 for each 2D viewpoint configuration (for example, the 2D viewpoint configuration V_1). In an embodiment, one or more 2D synthetic faces are generated for each 2D viewpoint configuration of the corresponding 3D target face (for example, 3D target face) for impersonation. The 2D viewpoint configuration V (i.e., $V = V_1, V_2, \ldots, V_n$) corresponding to each 3D target face (for example, 3D target face F_{s1}) may be generated may be enrolled as a synthetic identity for a user, for example, the user 102. As such, at least one 2D adversarial pattern (for example, A_1) is generated to impersonate a synthetic identity for each 2D viewpoint configuration (in set V) using a state-of-the-art adversarial pattern generation method. As already explained, the adversarial pattern is introduced or integrated within a targeted portion of the 2D viewpoint configuration, for example, forehead or cheeks. In one example, the adversarial pattern A_1 may be integrated in the jaw for the viewpoint configuration V_1 and adversarial pattern A_2 may be integrated in the cheek of the viewpoint configuration V_2. More specifically, the portion in which the adversarial pattern in introduced may be determined based

on the evaluation objective. For example, if a convolutional neural network is utilized by the face recognition system 108 for verifying identity of an individual, the evaluation objective of the CNN and evaluation technique to verify identity are assessed prior to determining the portion in which the adversarial portion has to be introduced. Some examples of adversarial pattern generation techniques include, but not limited to, projected gradient descent, Carlini, Wagner and the like. The adversarial patterns generated by the adversarial input generator 214 is forwarded to the texture manipulator 216.

[0051] The texture manipulator 216 in conjunction with the instructions 205 stored in the memory 204 is configured to perturb a 3D target texture based on adversarial pattern introduced in the 2D viewpoint configuration. The adversarial patterns generated for each 2D viewpoint configuration are mapped to the corresponding 3D target texture and used to perturb an original texture. Initially, expectation is taken over adversarial pattern set A (assume expected adversarial perturbation accounts for different configurations in *C*). Let it be *E(A).*

$$T_p = T_c + \mu(N,E(A)) \text{ ---------- (Equation 3)}$$

[0052] Here μ is the perturbation function and *N* is the set of normals *i.e.* set of perpendicular directions at each vertex in a 3D digital adversarial target T_i, T_c is the original texture for 3D digital adversarial face target T_i and T_p is texture of the adversarial face target that is to be printed on the adversarial face target. With this perturbation each 3D digital adversarial face target T_i is now a synthetically generated 3D digital adversarial face target. The process is repeated for generating each 3D digital adversarial face target T_i.

[0053] In an embodiment, the communication interface 208 is configured to send the set of 3D digital adversarial face targets T_i to a 3D printer 230. The 3D printer 230 is a high-fidelity color 3D printer. The 3D printer 230 fabricates each 3D digital adversarial face targets on skin-like material, for example, silicone to generate adversarial face targets. Some examples of the 3D printer 230 include, but not limited to, Stratasys Objet500 Connex3, and the like.

[0054] In an embodiment, an error ‘err’ in replicating texture T_{cp} for each 3D digital adversarial face target T_i that is induced because of manufacturing the adversarial face targets using the 3D printer 230 is determined. More specifically, the 3D adversarial face target is analysed to determine difference between expected perturbation and observed perturbation after

manufacturing the 3D adversarial face target. As such, ‘err’ indicates the deviation of the adversarial perturbation as observed from the 3D adversarial face target that was applied to a 2D viewpoint configuration. For each 3D digital adversarial face target T_i, the error is corrected before manufacturing based on equation (4).

$$T_{cp} = T_p + \mu(N, err(T_c)) \quad \text{---------- (Equation 4)}$$

T_{cp} is the texture to be printed, and T_p is the texture after perturbation from Equation 3. The process flow of generating the 3D digital target adversarial faces is explained next with reference to **FIG. 4**.

[0055] **FIG. 4** is a flowchart illustrating a method 400 for generating digital adversarial face targets. The method 400 depicted in the flow diagram may be executed by, for example, the system 200 shown and explained with reference to FIGS. 2-3. Operations of the flow diagram, and combinations of operation in the flow diagram, may be implemented by, for example, hardware, firmware, a processor, circuitry and/or a different device associated with the execution of software that includes one or more computer program instructions. The operations of the method 400 are described herein with help of the system 200. It is noted that the operations of the method 400 can be described and/or practiced by using one or more processors of a system/device other than the system 200. The method 400 starts at operation 402.

[0056] At operation 402 of the method 400, a plurality of images are received by a system such as, the system 200 explained with reference to **FIGS. 2-3**.

[0057] At operation 404 of the method 400, a set of 3D target faces are synthesized from the plurality of images using a sampling function. Synthesizing the set of 3D target faces from the detected faces in the plurality of images is explained with reference to FIG. 3 and is not explained herein for the sake of brevity.

[0058] At operation 406 of the method 400, a set of 2D viewpoint configurations corresponding to each 3D target face of the set of 3D target faces are captured based on a projection function. The set of 2D viewpoint configurations are generated from each 3D target face for different face configurations. These different face configurations correspond to each 3D target face in a

different viewpoint. Accordingly, the 3D target face is mapped to a 2D view plane to generate the 2D viewpoint configuration for different viewpoints. Generally, the 2D viewpoint configuration is face configuration based on the viewpoint and the projection function may be used for mapping each 3D target face to a 2D viewpoint for generating a 2D viewpoint configuration.

[0059] At operation 408 of the method 400, at least one adversarial pattern is generated in relation to each 2D viewpoint configuration of the set of 2D viewpoint configurations. The adversarial patterns may be generated for impersonation of a particular subject (i.e., human) or obfuscation to evade recognition. Some examples of adversarial pattern generation techniques include, but not limited to, projected gradient descent, Carlini, Wagner and the like.

[0060] At operation 410 of the method 400, a set of 3D digital adversarial face targets are generated by perturbing an original texture of 3D target face based on the set of 2D viewpoint configurations and the adversarial pattern. The adversarial patterns generated for each 2D viewpoint configuration are mapped to the corresponding 3D target texture and used to perturb an original texture. Evaluating performance of a face recognition system based on 3D adversarial face target is explained next with reference to **FIG. 5**.

[0061] **FIG. 5** is a flowchart illustrating a method 500 for testing efficiency of a face recognition system under various adversarial configurations before deployment.

[0062] At operation 502 of the method 500, a plurality of images are received by a system such as, the system 200 explained with reference to **FIGS. 2-4**. As already explained, the system 200 may be embodied within a face recognition system 108 configured to authenticate identity of users or maybe a remote standalone server or processor that generates the digital adversarial face targets. The plurality of images may correspond to a plurality of people. More specifically, the plurality of images capture facial features of the plurality of people. In an example, the plurality of images include images of differing demographics including age, gender, ethnicity, race, etc.

[0063] At operation 504 of the method 500, a set of 3D target faces are synthesized based on the plurality of images. In general, parametric 3D face shape and texture models of faces detected in the plurality of images are used to synthesize the set of 3D target faces in the digital domain. More specifically, statistical distribution of face shape and face texture of the plurality

of faces in the plurality of images are sampled to synthesize the set of 3D target faces using a sampling function. The use of sampling function to synthesize the set of 3D target face is explained in detail with reference to **FIG. 3** and is not explained herein for the sake of brevity. It shall be noted that such synthesis of 3D target face electronically ensures face shape and face texture may be controlled explicitly to represent real-world population demographics.

[0064] At operation 506 of the method 500, a set of 3D digital adversarial face targets corresponding to each 3D target face of the set 3D digital adversarial face targets is generated based on a set of 2D viewpoint configurations and corresponding at least one adversarial pattern.

[0065] At operation 508 of the method 500, the set of adversarial face targets is manufactured using a 3D printer based on the set of 3D digital adversarial face targets. In an embodiment, a high-fidelity 3D color printer is used to fabricate the 3D adversarial face targets on skin-like material such as, silicone. Some examples of the high-fidelity 3D color printer include, but not limited to, Stratasys, Objet500, Connex3, and the like. In some embodiments, the system 200 determines an error in replicating texture for each 3D adversarial face target that is induced because of manufacturing the 3D adversarial face target. The error is used to rectify texture of the 3D adversarial face target that is to be manufactured using equation 4.

[0066] At operation 510 of the method 500, performance of a face recognition system is evaluated using the set of adversarial face targets. More specifically, the face recognition system 108 is subject to physical adversarial testing using the set of adversarial face targets. The set of adversarial face targets are presented in a repeatable manner to the face recognition system 108 during the evaluation phase (i.e, prior to deployment phase). Initially, the at least one target face for which the set of adversarial target faces are generated is enrolled in the face recognition system 108. During the evaluation phase, each adversarial face target of the set of adversarial face targets are presented a plurality of times to the face recognition system. More specifically, each adversarial face target is obfuscated/impersonated and presented to the face recognition system 108 with respect to a corresponding enrolled identity. In an embodiment, the performance of the face recognition system 108 is evaluated using an average attack presentation match rate (APMR) is used as a metric. The APMR is a metric that computes ratio of successful adversarial attacks over a fixed number of adversarial presentation attempts. As such, the APMR is used to benchmark adversarial security of different face recognition systems

It shall be noted that the face recognition system 108 may be presented with such adversarial face targets even after deployment to evaluate performance of the face recognition system 108 after deployment to check the robustness of the system to different kinds of adversarial attacks.

[0067] Computer System

[0068] FIG. 6 illustrates a block diagram of an exemplary computer system for implementing embodiments consistent with the present disclosure.

[0069] In an embodiment, the computer system 600 may be used to implement the system 200. The computer system 600 may include a central processing unit (“CPU” or “processor”) 602. The processor 602 may include at least one data processor for securing third-party identification. The processor 602 may include specialized processing units such as, integrated system (bus) controllers, memory management control units, floating point units, graphics processing units, digital signal processing units, etc.

[0070] The processor 602 may be disposed in communication with one or more Input/Output (I/O) devices (612 and 613) via I/O interface 601. The I/O interface 601 employ communication protocols/methods such as, without limitation, audio, analog, digital, monoaural, radio corporation of America (RCA) connector, stereo, IEEE-1394 high speed serial bus, serial bus, universal serial bus (USB), infrared, personal system/2 (PS/2) port, Bayonet Neill-Concelman (BNC) connector, coaxial, component, composite, digital visual interface (DVI), high-definition multimedia interface (HDMI), radio frequency (RF) antennas, S-Video, video graphics array (VGA), IEEE 802.11b/g/n/x, Bluetooth, cellular e.g., code-division multiple access (CDMA), high-speed packet access (HSPA+), global system for mobile communications (GSM), long-term evolution (LTE), worldwide interoperability for microwave access (WiMAX), or the like, etc.

[0071] Using the I/O interface 601, the computer system 600 may communicate with one or more I/O devices such as input devices 612 and output devices 613. For example, the input devices 612 may be an antenna, keyboard, mouse, joystick, (infrared) remote control, camera, card reader, fax machine, dongle, biometric reader, microphone, touch screen, touchpad, trackball, stylus, scanner, storage device, transceiver, video device/source, etc. The output devices 613 may be a printer, fax machine, video display (e.g., cathode ray tube (CRT), liquid

crystal display (LCD), light-emitting diode (LED), plasma, plasma display panel (PDP), organic light-emitting diode display (OLED) or the like), audio speaker, etc.

[0072] In some embodiments, the processor 602 may be disposed in communication with a communication network 609 via a network interface 603. The network interface 603 may communicate with the communication network 609. The network interface 603 may employ connection protocols including, without limitation, direct connect, ethernet (e.g., twisted pair 10/100/1000 Base T), transmission control protocol/internet protocol (TCP/IP), token ring, IEEE 802.11a/b/g/n/x, etc. The communication network 609 may include, without limitation, a direct interconnection, local area network (LAN), wide area network (WAN), wireless network (e.g., using Wireless Application Protocol), the Internet, etc. Using the network interface 603 and the communication network 609, the computer system 600 may communicate with a database 614, which may be the enrolled templates database 613. The network interface 603 may employ connection protocols include, but not limited to, direct connect, ethernet (e.g., twisted pair 10/100/1000 Base T), transmission control protocol/internet protocol (TCP/IP), token ring, IEEE 802.11a/b/g/n/x, etc.

[0073] The communication network 609 includes, but is not limited to, a direct interconnection, a peer to peer (P2P) network, local area network (LAN), wide area network (WAN), wireless network (e.g., using Wireless Application Protocol), the Internet, Wi-Fi and such. The communication network 609 may either be a dedicated network or a shared network, which represents an association of the different types of networks that use a variety of protocols, for example, hypertext transfer protocol (HTTP), transmission control protocol/internet protocol (TCP/IP), wireless application protocol (WAP), etc., to communicate with each other. Further, the communication network 609 may include a variety of network devices, including routers, bridges, servers, computing devices, storage devices, etc.

[0074] In some embodiments, the processor 602 may be disposed in communication with a memory 605 (e.g., RAM, ROM, etc. not shown in Fig. 6) via a storage interface 604. The storage interface 604 may connect to memory 605 including, without limitation, memory drives, removable disc drives, etc., employing connection protocols such as, Serial Advanced Technology Attachment (SATA), integrated drive electronics (IDE), IEEE-1394, universal serial bus (USB), fiber channel, small computer systems interface (SCSI), etc. The memory drives may further include a drum, magnetic disc drive, magneto-optical drive, optical drive,

redundant array of independent discs (RAID), solid-state memory devices, solid-state drives, etc.

[0075] The memory 605 may store a collection of program or database components, including, without limitation, user interface 606, an operating system 607, etc. In some embodiments, computer system 600 may store user/application data, such as, the data, variables, records, etc., as described in this disclosure. Such databases may be implemented as fault-tolerant, relational, scalable, secure databases such as Oracle or Sybase.

[0076] The operating system 607 may facilitate resource management and operation of the computer system 600. Examples of operating systems include, without limitation, AppleTM MacintoshTM OS XTM, UNIXTM, Unix-like system distributions (e.g., Berkeley Software Distribution (BSD), FreeBSDTM, Net BSDTM, Open BSDTM, etc.), Linux distributions (e.g., Red HatTM, UbuntuTM, K-UbuntuTM, etc.), International Business Machines (IBMTM) OS/2TM, Microsoft WindowsTM (XPTM, Vista/7/8, etc.), Apple iOSTM, Google AndroidTM, BlackberryTM operating system (OS), or the like.

[0077] In some embodiments, the computer system 600 may implement web browser 608 stored program components. Web browser 608 may be a hypertext viewing application, such as MicrosoftTM Internet ExplorerTM, Google ChromeTM, Mozilla FirefoxTM, AppleTM SafariTM, etc. Secure web browsing may be provided using secure hypertext transport protocol (HTTPS), secure sockets layer (SSL), transport layer security (TLS), etc. Web browsers 608 may utilize facilities such as AJAX, DHTML, AdobeTM Flash, JavaScript, Application Programming Interfaces (APIs), etc. In some embodiments, the computer system 600 may implement a mail server stored program component. The mail server may be an Internet mail server such as Microsoft Exchange, or the like. The mail server may utilize facilities such as ASP, ActiveX, ANSI C++/C#, Microsoft .NET, Common Gateway Interface (CGI) scripts, Java, JavaScript, PERL, PHP, Python, WebObjects, etc. The mail server may utilize communication protocols such as Internet Message Access Protocol (IMAP), Messaging Application Programming Interface (MAPI), Microsoft Exchange, Post Office Protocol (POP), Simple Mail Transfer Protocol (SMTP), or the like. In some embodiments, the computer system 600 may implement a mail client stored program component. The mail client may be a mail viewing application, such as Apple Mail, Microsoft Entourage, Microsoft Outlook, Mozilla Thunderbird, etc.

[0078] Furthermore, one or more computer-readable storage media may be utilized in implementing embodiments consistent with the present disclosure. A computer-readable storage medium refers to any type of physical memory on which information or data readable by a processor may be stored. Thus, a computer-readable storage medium may store instructions for execution by one or more processors, including instructions for causing the processor(s) to perform steps or stages consistent with the embodiments described herein. The term "computer-readable medium" should be understood to include tangible items and exclude carrier waves and transient signals, i.e., be non-transitory. Examples include Random Access Memory (RAM), Read-Only Memory (ROM), volatile memory, non-volatile memory, hard drives, Compact Disc (CD) ROMs, DVDs, flash drives, disks, and any other known physical storage media.

[0079] The described operations may be implemented as a method, system or article of manufacture using standard programming and/or engineering techniques to produce software, firmware, hardware, or any combination thereof. The described operations may be implemented as code maintained in a "non-transitory computer readable medium", where a processor may read and execute the code from the computer readable medium. The processor is at least one of a microprocessor and a processor capable of processing and executing the queries. A non-transitory computer readable medium may include media such as magnetic storage medium (e.g., hard disk drives, floppy disks, tape, etc.), optical storage (CD-ROMs, DVDs, optical disks, etc.), volatile and non-volatile memory devices (e.g., EEPROMs, ROMs, PROMs, RAMs, DRAMs, SRAMs, Flash Memory, firmware, programmable logic, etc.), etc. Further, non-transitory computer-readable media may include all computer-readable media except for a transitory. The code implementing the described operations may further be implemented in hardware logic (e.g., an integrated circuit chip, Programmable Gate Array (PGA), Application Specific Integrated Circuit (ASIC), etc.).

[0080] The illustrated steps are set out to explain the exemplary embodiments shown, and it should be anticipated that ongoing technological development will change the manner in which particular functions are performed. These examples are presented herein for purposes of illustration, and not limitation. Further, the boundaries of the functional building blocks have been arbitrarily defined herein for the convenience of the description. Alternative boundaries can be defined so long as the specified functions and relationships thereof are appropriately

performed. Alternatives (including equivalents, extensions, variations, deviations, etc., of those described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. Such alternatives fall within the scope and spirit of the disclosed embodiments. Also, the words "comprising," "having," "containing," and "including," and other similar forms are intended to be equivalent in meaning and be open ended in that an item or items following any one of these words is not meant to be an exhaustive listing of such item or items or meant to be limited to only the listed item or items. It must also be noted that as used herein, the singular forms “a,” “an,” and “the” include plural references unless the context clearly dictates otherwise.

[0081] Furthermore, one or more computer-readable storage media may be utilized in implementing embodiments consistent with the present disclosure. A computer readable storage medium refers to any type of physical memory on which information or data readable by a processor may be stored. Thus, a computer readable storage medium may store instructions for execution by one or more processors, including instructions for causing the processor(s) to perform steps or stages consistent with the embodiments described herein. The term “computer readable medium” should be understood to include tangible items and exclude carrier waves and transient signals, i.e., are non-transitory. Examples include random access memory (RAM), read-only memory (ROM), volatile memory, non-volatile memory, hard drives, CD ROMs, DVDs, flash drives, disks, and any other known physical storage media.

[0082] Finally, the language used in the specification has been principally selected for readability and instructional purposes, and it may not have been selected to delineate or circumscribe the inventive subject matter. Accordingly, the disclosure of the embodiments of the disclosure is intended to be illustrative, but not limiting, of the scope of the disclosure.

[0083] With respect to the use of substantially any plural and/or singular terms herein, those having skill in the art can translate from the plural to the singular and/or from the singular to the plural as is appropriate to the context and/or application. The various singular/plural permutations may be expressly set forth herein for sake of clarity.

ADVERSERIAL FACE TARGETS

ABSTRACT

The present disclosure relates to adversarial face targets that may be used to test performance of a face recognition system. As such, a plurality of images are received by a system. The plurality of images are processed to detect faces and a set of 3D target faces are synthesized. Further, a set of 2D viewpoint configurations corresponding to each 3D target face of the set of 3D target faces are captured based on a projection function. Adversarial perturbations are generated in relation to each 2D viewpoint configuration of the set of 2D viewpoint configurations. Thereafter, a set of 3D digital adversarial face targets are generated by perturbing an original texture of 3D target face based on the set of 2D viewpoint configurations and the adversarial pattern. The set of adversarial face targets is manufactured using a 3D printer based on the set of 3D digital adversarial face targets and performance of the face recognition system is evaluated using the set of adversarial face targets.

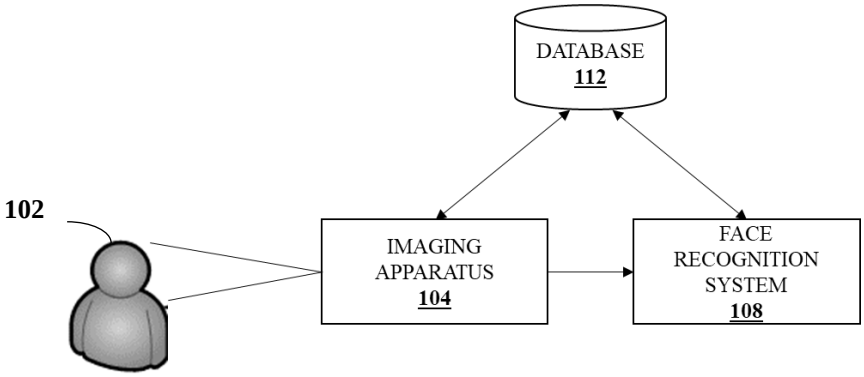

DATABASE
112
102
IMAGING
APPARATUS
104
FACE
RECOGNITION
SYSTEM
108

FIG. 1

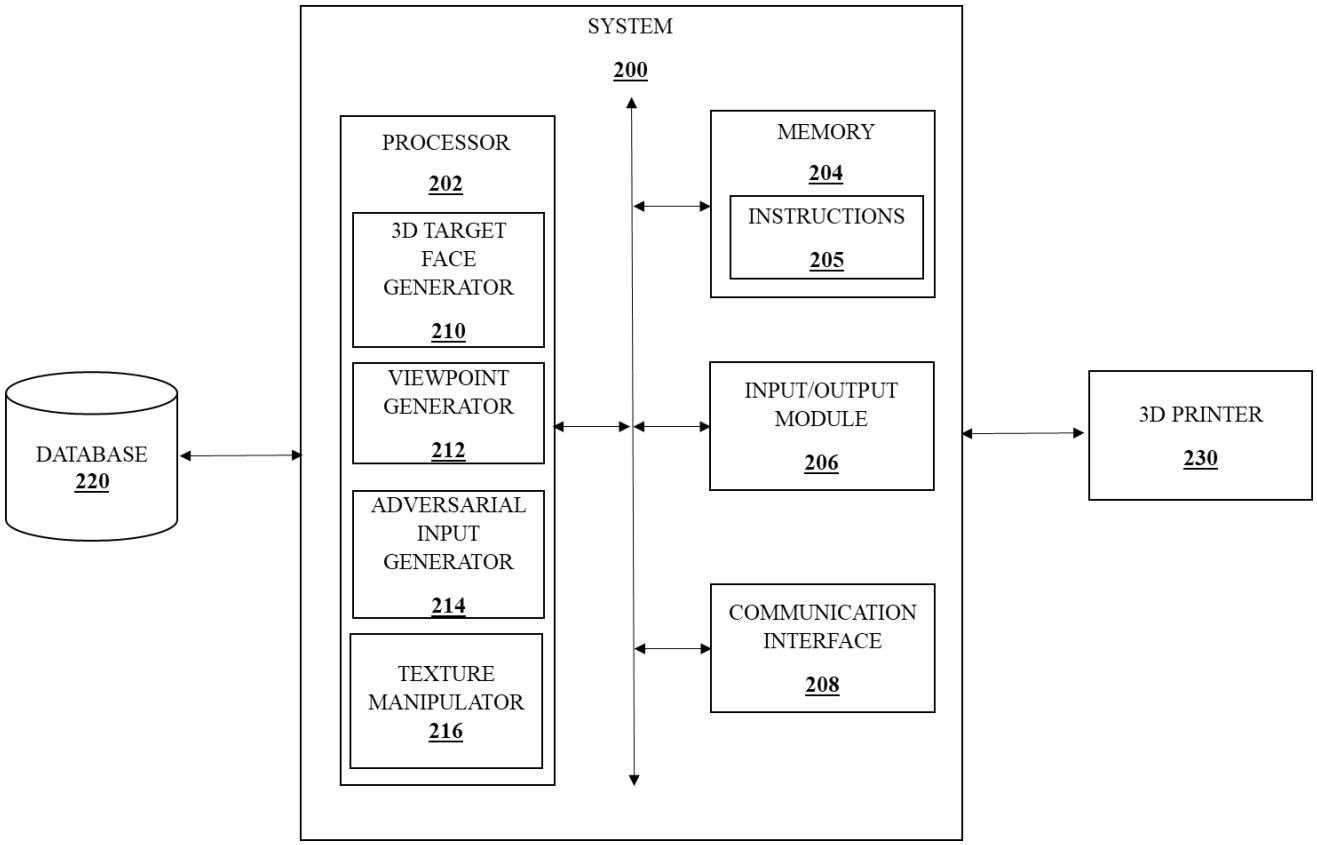

SYSTEM
200
PROCESSOR
202
3D TARGET
FACE
GENERATOR
210
VIEWPOINT
GENERATOR
212
ADVERSARIAL
INPUT
GENERATOR
214
TEXTURE
MANIPULATOR
216
MEMORY
204
INSTRUCTIONS
205
INPUT/OUTPUT
MODULE
206
COMMUNICATION
INTERFACE
208
DATABASE
220
3D PRINTER
230

FIG. 2

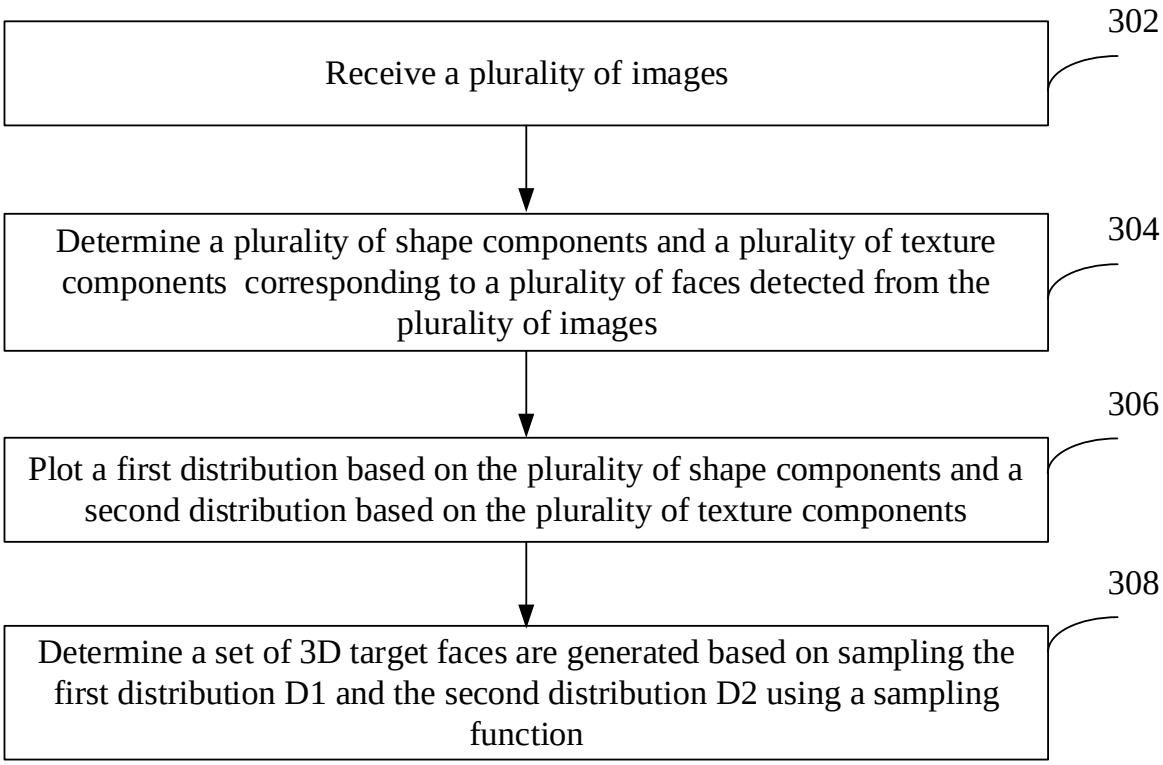
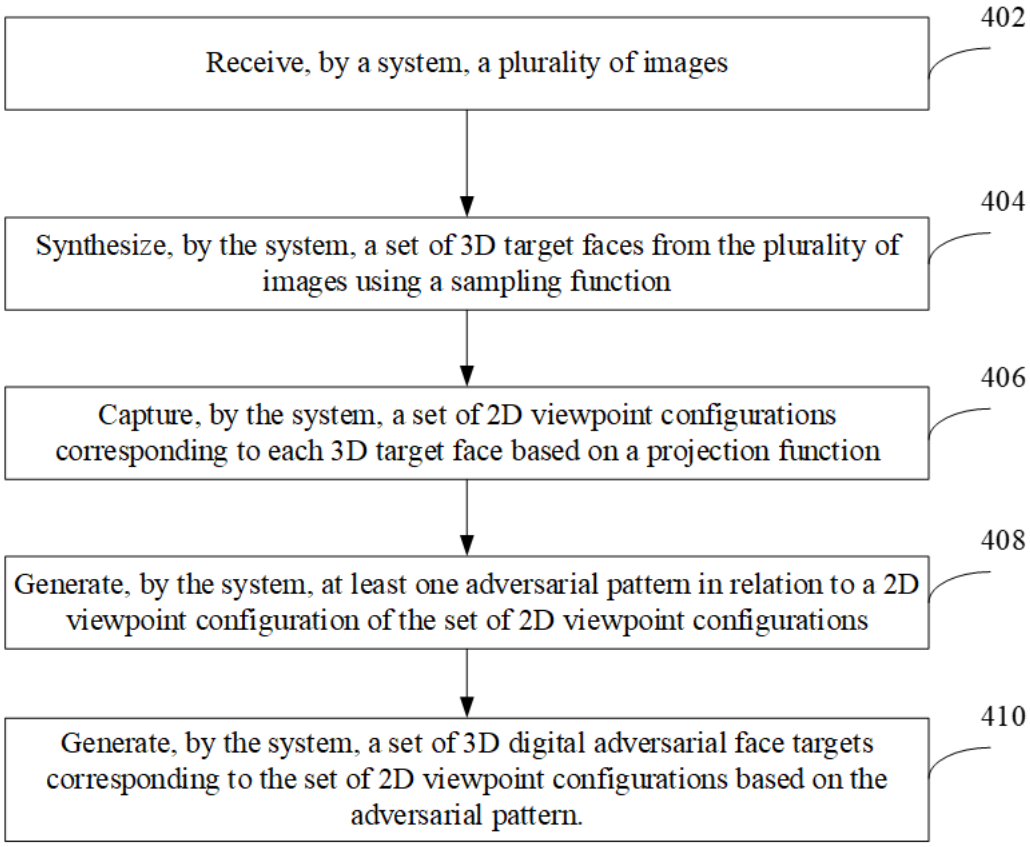
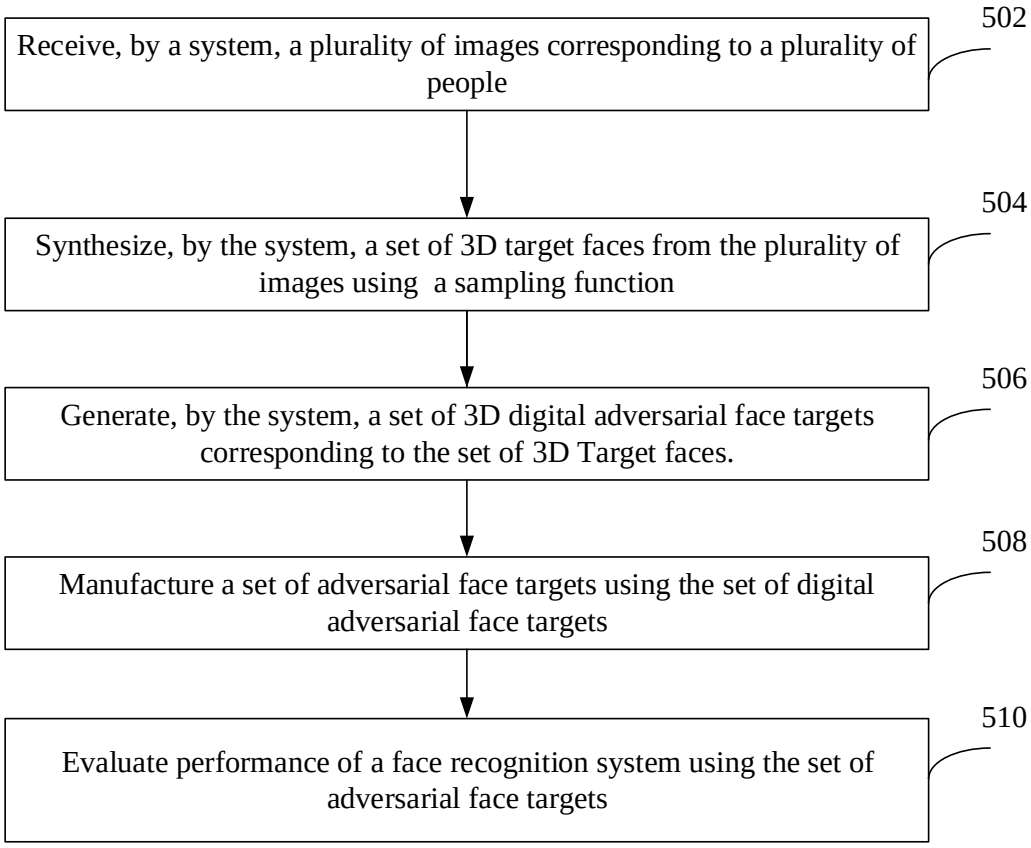

302
Receive a plurality of images
304
Determine a plurality of shape components and a plurality of texture components corresponding to a plurality of faces detected from the plurality of images
306
Plot a first distribution based on the plurality of shape components and a second distribution based on the plurality of texture components
308
Determine a set of 3D target faces are generated based on sampling the first distribution D1 and the second distribution D2 using a sampling function

FIG. 3


Receive, by a system, a plurality of images
402
Synthesize, by the system, a set of 3D target faces from the plurality of images using a sampling function
404
Capture, by the system, a set of 2D viewpoint configurations corresponding to each 3D target face based on a projection function
406
Generate, by the system, at least one adversarial pattern in relation to a 2D viewpoint configuration of the set of 2D viewpoint configurations
408
Generate, by the system, a set of 3D digital adversarial face targets corresponding to the set of 2D viewpoint configurations based on the adversarial pattern.
410


Receive, by a system, a plurality of images corresponding to a plurality of people
502
Synthesize, by the system, a set of 3D target faces from the plurality of images using a sampling function
504
Generate, by the system, a set of 3D digital adversarial face targets corresponding to the set of 3D Target faces.
506
Manufacture a set of adversarial face targets using the set of digital adversarial face targets
508
Evaluate performance of a face recognition system using the set of adversarial face targets
510

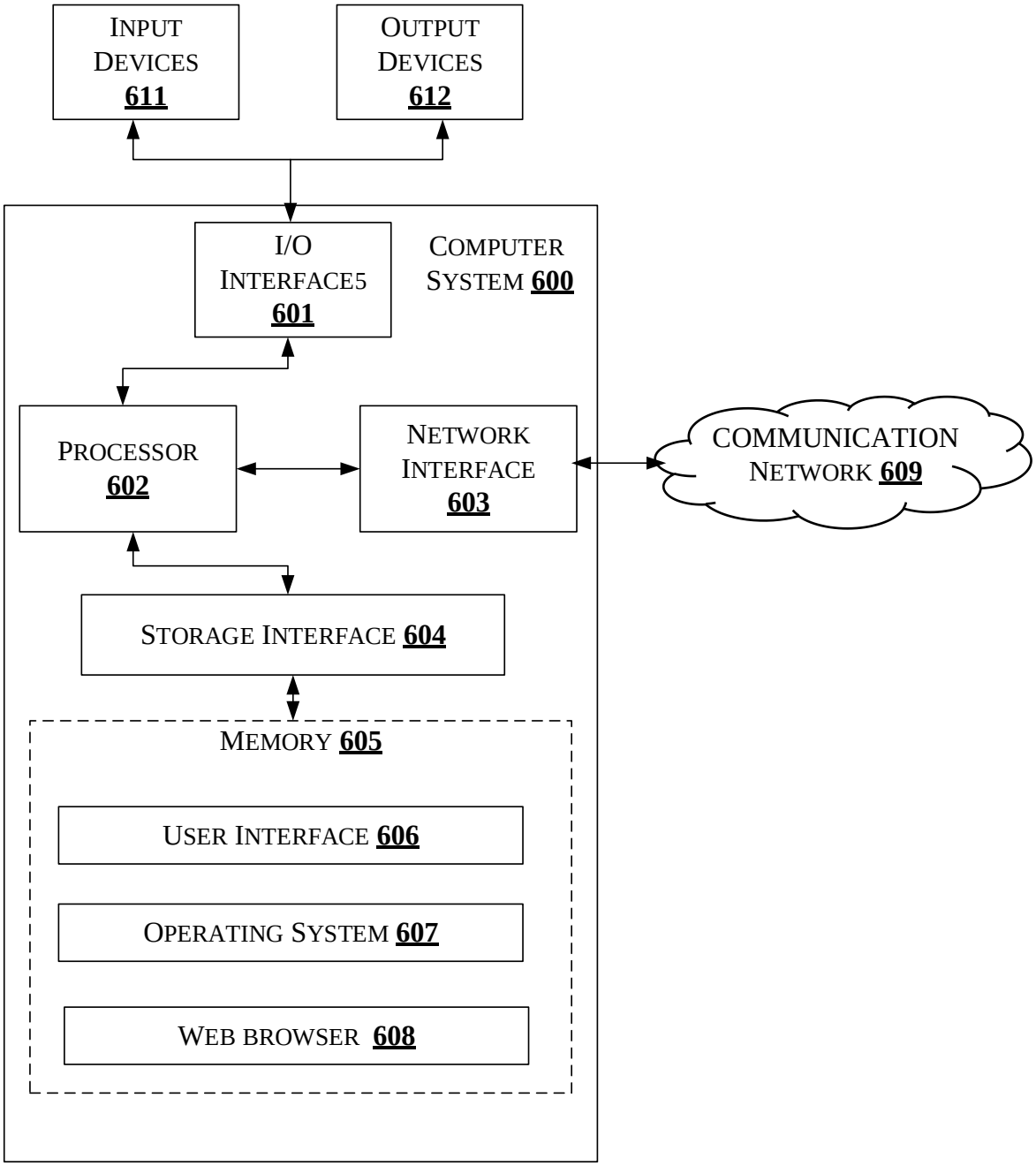

INPUT DEVICES 611
OUTPUT DEVICES 612
I/O INTERFACE5 601
COMPUTER SYSTEM 600
PROCESSOR 602
NETWORK INTERFACE 603
COMMUNICATION NETWORK 609
STORAGE INTERFACE 604
MEMORY 605
USER INTERFACE 606
OPERATING SYSTEM 607
WEB BROWSER 608

FIG. 6