

USING AN ENROLLED BIOMETRIC DATASET TO DETECT ADVERSARIAL EXAMPLES IN BIOMETRICS-BASED AUTHENTICATION SYSTEM

BACKGROUND

[001] Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool machine learning models through malicious input. Recent research has shown a susceptibility of deep neural networks to adversarial examples. Adversarial examples are inputs that are specifically crafted with a malicious intent. For a network trained to classify objects in an image, adversarial examples could be generated with the goal of misclassifying an object belonging to one category to a different category, such as one person modifying his/her image to impersonate another. Examples of adversarial examples include a machine-tweaked image of a dog that looked like a cat both to computers and to humans, and perturbing the appearance of a stop sign so that an autonomous vehicle will classify the stop sign as a merge or speed limit sign instead.

[002] Over the past few years, several different methods to generate adversarial examples for deep neural networks have been published. Techniques have been also developed to defend against adversarial examples crafted using some of these methods. Most of these defensive techniques, however, though can be broken easily given the knowledge of the defense method.

[003] Accordingly, it would be desirable to provide an improved system and method for detecting adversarial examples.

BRIEF SUMMARY

[004] Various embodiments provide methods and systems for improving security of a biometrics-based authentication system using an enrolled biometric dataset to detect adversarial examples. Aspects of the embodiments include receiving enrolled biometric samples of an enrolled user during an enrollment stage of the biometrics-based authentication system. Augmented biometric samples are created by adding learned perturbations to the enrolled biometric samples of the enrolled user. During a request for authentication, submitted biometric samples are received from a second user purporting to be the enrolled user. The submitted biometric samples of the second user are compared to the enrolled biometric samples and to the augmented biometric samples of the enrolled user based on predefined metrics. Based on the comparison it is determined whether the submitted biometric samples of the second user have been modified to impersonate the enrolled user.

[005] According to the method and system disclosed herein, the exemplary embodiments improve adversarial defense techniques of biometric authentication systems in that adding the learned perturbations to the enrolled biometric samples of the enrolled user increases the difficulty of an adversary creating adversarial biometric samples even in the case where the enrollment biometric samples are leaked to the adversary via an insider breach.

BRIEF DESCRIPTION OF SEVERAL VIEWS OF THE DRAWINGS

[006] Figure 1 is a diagram illustrating one embodiment of a system for using an enrolled biometric dataset to detect adversarial examples in a biometrics-based authentication system.

[007] Figure 2 is a diagram illustrating a generic authentication process performed by the biometric authentication system prior to activation of the adversarial defense system.

[008] Figure 3 is a diagram illustrating an impersonation attack performed on a generic biometric authentication system.

[009] Figure 4 is a flow diagram illustrating one embodiment of a process for improving security of a biometrics-based authentication system.

[010] Figure 5 is a diagram illustrating detection of an impersonation attack on the biometric authentication system by the adversarial defense system in further detail.

[011] Figure 6 is a diagram illustrating the process used by the adversarial perturbation detector to detect perturbations in the submitted biometric samples.

[012] Figure 7 is a diagram illustrating embodiments performed by the adversarial perturbation detector to detect adversarial submitted biometric samples.

[013] Figure 8A is a diagram illustrating a process for training the adversarial perturbation detector prior to deployment.

[014] Figure 8B is a diagram illustrating an overview of the process of creating augmented biometric samples of the enrolled user (block 402 of Figure 4).

DETAILED DESCRIPTION

[015] The exemplary embodiments relate to using an enrolled biometric dataset to detect adversarial examples in biometrics-based authentication system. The following description is presented to enable one of ordinary skill in the art to make and use the invention and is provided in the context of a patent application and its requirements. Various modifications to the exemplary embodiments and the generic principles and features described herein will be readily apparent. The exemplary embodiments are mainly described in terms of particular methods and systems provided in particular implementations. However, the methods and systems will operate effectively in other implementations. Phrases such as "exemplary embodiment", "one embodiment" and "another embodiment" may refer to the same or different embodiments. The embodiments will be described with respect to systems and/or devices having certain components. However, the systems and/or devices may include more or less components than those shown, and variations in the arrangement and type of the components may be made without departing from the scope of the invention. The exemplary embodiments will also be described in the context of particular methods having certain steps. However, the method and system operate effectively for other methods having different and/or additional steps and steps in different orders that are not inconsistent with the exemplary embodiments. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features described herein.

[016] The disclosed embodiments relate to using an enrolled biometric dataset to detect adversarial examples in a biometrics-based authentication system. Applicants recognize that one way to design stronger defense mechanisms is to incorporate domain knowledge. According to the disclosed embodiments, domain knowledge is used in a biometric verification setting to design a strong defense mechanism from adversarial biometric samples. More specifically, security of a biometrics-based authentication system is improved by using enrolled biometric samples of a user to detect adversarial biometric samples crafted specifically to target the enrolled user's identity. It is assumed that the adversary crafting the adversarial biometric samples does not have access to the enrolled biometric samples. However, augmented biometric samples with learned perturbations (realistic or synthetic) are added to the enrolled biometric samples with the aim of increasing the difficulty of crafting adversarial biometric samples if the enrolled biometric samples are accessed by an adversary due to, for example, an insider data breach. The enrolled biometric samples with the augmented enrolled biometric samples are compared with the adversarial biometric samples based on different pre-defined metrics. Based on the comparison, a decision is made whether the adversarial biometric samples have been modified to impersonate the enrolled user.

[017] Figure 1 is a diagram illustrating one embodiment of a system for using enrolled biometric dataset to detect adversarial examples in a biometrics-based authentication system. The system 10 includes a biometric authentication system 12 in communication with the user devices 14 over a network 16, such as the Internet. Components of the biometric authentication system 12 may include one or more processors or servers 20, a

projection deep neural network (DNN) 22, an enrollment database 24, a similarity comparator 26 and an adversarial defense system 28.

[018] In one embodiment, the biometric authentication system 12 may be the front end for another system, such as a payment processing network 18, to authenticate users for a transaction. The payment processing network 18 may refer to an entity that receives transaction authorization requests from merchants or other entities and provides guarantees of payment, in some cases through an agreement between the transaction service provider and an issuer institution. The payment processing network supports and delivers payment related services (e.g., authentication services, authorization services, exception file services, and clearing and settlement services, etc.). Examples of a payment processing network may include a payment network provided by Visa®, MasterCard®, American Express®, or any other entity that processes credit card transactions, debit card transactions, and other types of commercial transactions.

[019] The biometric authentication system 12 acquires biometric samples of a first user (e.g., enrolled user A) during an enrollment stage. The enrollment stage may be performed through a software authentication application 34 provided by the biometric authentication system 12 that runs on either one of the servers 20 or a user device 14 (e.g., smartphone, PC, watch or tablet). During the enrollment process the authentication application 34 prompts the enrolled user to create enrolled biometric samples 30, which in one embodiment, are images of the user's face/head. In one embodiment, the images may be taken with the user device 14, and the authentication application 34 transmits the enrolled biometric samples 30 to the biometric authentication system 12 over the network

16, where they are received by one of the servers 20. The enrolled biometric samples 30 are processed by a projection DNN 22 into enrolled biometric dataset 32 and stored in the enrollment database 24, as explained further below.

[020] Subsequently, the biometric authentication system 12 receives an authentication request from a second user (user B). The authentication request transmitted from the second user includes submitted biometric samples 36. The request can be made by the enrolled user or another user, e.g., user B, or user B may be an attacker purporting to be the enrolled user. The submitted biometric samples 36 are processed by the projection DNN 22 into submitted biometric datasets (not shown) and the similarity comparator 26 compares the enrolled biometric datasets 32 of the enrolled user to the submitted biometric datasets. If there is a match within one or more thresholds, the user is authorized and gains access to the payment processing network 18. If not, authorization request of the second user is rejected.

[021] Before describing the adversarial defense system 28 that enhances security of the biometric authentication system 12, the authentication process performed by the biometric authentication system 12 and attacks thereon are first explained.

[022] Figure 2 is a diagram illustrating a generic authentication process performed by the biometric authentication system 12 prior to activation of the adversarial defense system 28, where like components from Figure 1 have like reference numerals. Referring to both Figures 1 and 2, the biometric authentication system 12 receives the enrolled biometric samples 30, represented as x_i , of the enrolled user. The enrolled biometric samples x_i may be one or many samples. The projection DNN 22 applies a mapping

function $f(\cdot)$ to the enrolled biometric samples 30 to generate a biometric template $f(x_i)$ 200, which can be a one dimensional feature vector or a high dimensional tensor. The enrolled biometric samples x_i and the biometric template $f(x_i)$ 200 are stored as the enrolled biometric dataset 32 in the enrollment database 24.

[023] Similarly, the biometric authentication system 12 receives the submitted biometric sample x' 36 (one or many) from the same or a second user. The projection DNN 22 applies the same mapping function $f(\cdot)$ to the submitted biometric samples 36 to generate a biometric template $f(x')$ 202. The mapping function $f(\cdot)$ projects x_i and x' to a common embedding subspace.

[024] The similarity comparator 26 then compares the biometric template $f(x_i)$ 200 with the biometric template $f(x')$ 202 in the embedding subspace. This is done by computing a distance between the biometric template 200 and biometric template 202. The distance may be represented as similarity score or distance score, such that $D = \sum_i d(f(x_i), f(x'))$. If the similarity comparator 26 determines that the distance D is greater than a first threshold (t), the authorization request is rejected. Otherwise, the authentication request is authorized.

[025] Referring again to Figure 1, the enrolled user may further have public biometric samples 38 that are easier for an attacker to obtain than the enrolled biometric samples 30 that are stored on the user device 14 and/or in the enrollment database 24. For example, the enrolled user may have images stored on social media websites that are publically available to others to see. As explained below, the presence of public biometric

samples 38 presents a security risk as an attacker can gain access to, and modify, the public biometric samples 38 to attack the biometric authentication system 12.

[026] Figure 3 is a diagram illustrating an impersonation attack performed on a generic biometric authentication system. Figure 3 assumes an attacking second user targets the enrolled user for impersonation and that the biometric information of the attacking second user is represented by biometric sample x' . The second user first obtains the public biometric sample 38 of the enrolled user. The second user then computes a similarity between the biometric template 200 of the enrolled user and a biometric template of the attacker (not shown). Based on the similarity, the attacker uses the public biometric sample 38 to create a perturbation $\Delta x'$ 300 that will be added to the attacker's biometric sample x' to generate a perturbed adversarial sample $x' + \Delta x'$ 302. The perturbed adversarial sample $x' + \Delta x'$ 302 of the attacker is designed to minimize the distance between perturbed adversarial sample $x' + \Delta x'$ 302 and the enrolled biometric samples x_i of the targeted enrolled user. The perturbation $\Delta x'$ 300 is computed based on standard attacking methods such as the Fast Gradient Signed Method (FGSM), projected gradient descent (PGD) attack, Carlini-Wagner (CW) loss function, and the like to minimize the distance: $d((f(x' + \Delta x'), f(x)))$ or $d((f(x' + \Delta x'), f(x_i)))$.

[027] The perturbation $\Delta x'$ 300 may be crafted based either on the public biometric samples 38 or the enrolled biometric samples 30 if obtainable (e.g. during a breach of the enrollment database 24 or the user device 14 of the enrolled user). Due to the transferability of adversarial perturbation, even though the perturbation $\Delta x'$ is crafted based on the public biometric samples 38 of the enrolled user, the perturbation $\Delta x'$ can

be used to attack the enrolled biometric samples 30. The perturbation $\Delta x'$ 300 may be physical or digital. In the example shown, the perturbation $\Delta x'$ 300 comprises cardboard glasses that are worn by the second user when taking biometric sample images to generate the perturbed adversarial sample $x' + \Delta x'$ 302.

[028] After the perturbed adversarial sample $x' + \Delta x'$ 302 is captured and submitted to the biometric authentication system, the projection DNN 22 applies the same mapping function $f(\cdot)$ to the perturbed adversarial sample $x' + \Delta x'$ 302 to generate a perturbed adversarial template $f(x' + \Delta x')$ 304. The similarity comparator 26 then compares the biometric template $f(x_i)$ 200 of the enrolled user with the perturbed adversarial template $f(x' + \Delta x')$ 304 in the embedding subspace. However, in this case since the perturbed adversarial template $f(x' + \Delta x')$ 304 is designed to be close in distance to the biometric template $f(x_i)$ 200, the similarity comparator 26 misclassifies and authorizes the attacking second user because the distance D or similarity score is less than the threshold t , i.e., $\sum_i d(f(x_i), f(x' + \Delta x')) < t$. Thus, in this approach, the attacker uses a well-crafted perturbed adversarial biometric sample 302 to impersonate the enrolled user to attack the generic authentication system and generate a fraudulent transaction once authenticated.

[029] Referring again to Figure 1, according to the disclosed embodiments, the adversarial defense system 28 is added to the biometric authentication system 12 (e.g., facial image, voice, fingerprint, etc.) to improve the security of the biometric authentication system 12 by defending against the use of adversarial samples of an attacker. As described herein, the adversarial defense system 28 uses the enrolled biometric samples 30 from the enrolled user to detect any fraudulently submitted biometric samples 36. In

one embodiment the biometric samples may comprise a facial image, but may represent other types of biometric data such as voice, fingerprints, and the like.

[030] In one embodiment, the adversarial defense system 28 may include an augmented biometric sample generator 42 that uses and modifies the enrolled biometric dataset 32 of the enrolled user to include: (i) the one or more enrolled biometric samples 30 of the enrolled user acquired during in the enrollment stage, which are difficult to obtain by the attacker, and (ii) augmented biometric samples 40 (can be both realistic or synthesized) created from the enrolled biometric samples 30 or the public biometric samples 38 that increase the difficulty of crafting perturbed adversarial samples 302 should the enrollment biometric dataset 32 be leaked due to an insider data breach. In one embodiment, the augmented biometric sample generator 42 adds learned perturbations 44 to the enrolled biometric samples 30 to generate the augmented biometric samples 40.

[031] In one embodiment, the projection DNN 22, the similarity comparator 26, and the adversarial defense system 28 are implemented as software components. In another embodiment, the components could be implemented as a combination of hardware and software. Although the projection DNN 22, the similarity comparator 26, and the adversarial defense system 28 are shown as separate components, the functionality of each may be combined into a lesser or greater number of modules/components. In addition, although one or more servers 20 are described as running the projection DNN 22, the similarity comparator 26, and the adversarial defense system 28, such

components may be run on any type of one more computers that have a non-transitory memory and processor.

[032] Both the server 20 and the user devices 14 may include hardware components of typical computing devices (not shown), including a processor, input devices (e.g., keyboard, pointing device, microphone for voice commands, buttons, touchscreen, etc.), and output devices (e.g., a display device, speakers, and the like). The server 20 and user devices 14 may include computer-readable media, e.g., memory and storage devices (e.g., flash memory, hard drive, optical disk drive, magnetic disk drive, and the like) containing computer instructions that implement the functionality disclosed when executed by the processor. The server 20 and the user devices 14 may further include wired or wireless network communication interfaces for communication. It should be understood that the functions of the software components may be implemented using a different number of software components than that shown.

[033] Figure 4 is a flow diagram illustrating one embodiment of a process for improving security of a biometrics-based authentication system. The process may begin by one or more servers 20 obtaining the enrolled biometric samples 30 of an enrolled user during the enrollment stage of the biometrics-based authentication system (block 400). Augmented biometric samples 40 are created by adding learned perturbations 44 to the enrolled biometric samples 30 of the enrolled user (block 402). In one embodiment, the augmented biometric sample generator 42 executing on one or more processors or servers generates and adds the learned perturbations 44 to the enrolled biometric samples 30 to create the augmented biometric samples 40.

[034] Thereafter, during a request for authentication by a second user purporting to be the enrolled user, the one or more servers 20 receive submitted biometric samples 36 from the second user (block 404).

[035] In response, the submitted biometric samples 36 of the second user are compared to the enrolled biometric samples 30 and to the augmented biometric samples of the enrolled user based on predefined metrics (block 406). In one embodiment, 46 may be implemented by the similarity comparator 26 executing on the one or more processors or servers 20.

[036] Based on the comparison, it is determined that the biometric samples of the second user have been modified with a perturbation 300 to impersonate the enrolled user (block 408). In one embodiment, block 408 may be implemented by the adversarial perturbation detector 28 executing on the one or more processors or servers 20. Referring again to Figure 1, in one embodiment, the adversarial defense system 28 may further comprises an adversarial perturbation detector 46 to determine if the submitted biometric samples 36 of the second user have been modified with a perturbation. In one embodiment, the adversarial perturbation detector 46 may include a transformation convolutional neural network (CNN) 48 and a classifier 50 (e.g. a neural network), as described further below.

[037] Figure 5 is a diagram illustrating detection of an impersonation attack on the biometric authentication system by the adversarial defense system in further detail. The process begins similar to the process described in Figure 3 in which the attacker (user B) obtains the public biometric sample 38 of the enrolled user; computes a similarity between

the biometric template 200 of the enrolled user and a biometric template of the attacker (not shown); and based on the similarity, the public biometric sample 38 is used to create a perturbation $\Delta x'$ 300 that will be added to the attacker's biometric sample x' to generate a perturbed adversarial sample $x' + \Delta x'$ 302. Detection of an impersonation attack involves detecting that the submitted biometric samples 36 of the second user comprises one or more perturbed adversarial samples $x' + \Delta x'$ 302.

[038] First, the submitted biometric sample 36 comprising the perturbed adversarial sample $x' + \Delta x'$ 302 is received by the biometric authentication system 12, and the projection DNN 22 applies the mapping function $f(\cdot)$ to the perturbed adversarial sample $x' + \Delta x'$ 302 to generate the perturbed adversarial template $f(x' + \Delta x')$ 304. The similarity comparator 26 then compares the biometric template $f(x_i)$ 200 of the enrolled user from the enrollment database 24 with the perturbed adversarial template $f(x' + \Delta x')$ 304 in the embedding subspace by calculating the distance score as described in Figure 2. However, in this case since the perturbed adversarial template $f(x' + \Delta x')$ 304 is designed to be close in distance to the biometric template $f(x_i)$ 200, the similarity comparator 26 authorizes the attacking second user.

[039] According to one aspect of the disclosed embodiments, the adversarial perturbation detector 28 is activated only in response to the request for authentication by second user being temporarily authorized by the biometric authentication system 12 via the similarity comparator 26 (block 500). According to one aspect of the disclosed embodiments, the adversarial perturbation detector 28 uses the enrolled biometric samples x_i 30 of the enrolled user, which are hidden/inaccessible by the second user, to

detect whether adversarial samples were submitted by the second user. During a training stage, the adversarial perturbation detector 28 uses the enrolled biometric samples 30 at least in part to create the augmented biometric samples 40 by adding learned perturbations 44 to the enrolled biometric samples 30. The augmented biometric samples 40 are then added to the enrolled biometric samples 30.

[040] In addition to the enrolled biometric sample x_i 30, the adversarial perturbation detector 28 receives as input the submitted biometric samples 36 of the second user, which comprise perturbed adversarial samples $x' + \Delta x'$ 302, and determines if any perturbation are detected (block 502). Responsive to detecting any perturbations in the submitted biometric samples 36, the adversarial perturbation detector 28 rejects the authorization request of the second user (block 504). Otherwise the authorization request is granted and the second user is authorized.

[041] Figure 6 is a diagram illustrating the process used by the adversarial perturbation detector 28 to detect whether the submitted biometric samples 36 have been modified with perturbations. In this embodiment, the submitted biometric samples 36 of the second user are adversarial in that they have been altered with perturbations and now comprise perturbed adversarial samples $f(x' + \Delta x')$ 282.

[042] The process may begin by the adversarial perturbation detector 28 accessing the enrolled biometric dataset 32 from the enrollment database 24 to aid in detecting adversarial samples. The enrolled biometric dataset 32 includes images with the learned perturbations 44. Thus the enrolled biometric dataset 32 comprises x_i samples/images,

where $i \in \{1:N\}$. Similarly, the perturbed adversarial samples $x' + \Delta x'$ 302 may include multiple perturbations, shown as $\sum \Delta x'$, where $\sum$ is a summation.

[043] According to a further aspect of the disclosed embodiments, the adversarial perturbation detector 28 further includes a transformation CNN 48 that has a learned mapping function $k(\cdot)$ configured to maximize the summation of distance $\sum_i d(k(x' + \Delta x') - k(x_i))$, $i \in \{1:M\}$. The adversarial perturbation detector 28 receives the enrolled biometric samples 30 comprising x_i of the enrolled user and the perturbed adversarial samples $x' + \sum \Delta x'$ 302 of the second user for transformation into a transformed subspace. In one embodiment, this may be done by using transformation CNN 48 to apply function $k(\cdot)$ to the enrolled biometric samples x_i 30 and the perturbed adversarial samples $x' + \sum \Delta x'$ 302 to generate biometric template $k(x_i)$ 600 and a biometric template $k(x' + \sum \Delta x')$ 602, which are in the transformed subspace. In embodiments, the mapping function $k(\cdot)$ can be decomposed into multiple projections: $k = g_1 \circ g_2 \circ g_3 \dots$, where g_1 can be layer in a convolutional neural network, g_2 can be another layer in the convolutional network, and g_3 can be a non-linear projection in a support vector machine (SVM).

[044] The adversarial perturbation detector 28 then computes a distance (similarity score) between biometric template $k(x_i)$ 600 and a biometric template $k(x' + \sum \Delta x')$ 602 in the transformed subspace (block 604). In one embodiment the distance F can be computed as: $F = \sum_i d(k(x_i), k(x' + \sum \Delta x'))$. The classifier 50 determines that the biometric template $k(x' + \sum \Delta x')$ 602, and therefore the submitted biometric samples 36 of the second user, is adversarial when the computed distance F is greater than a second threshold t' , indicating that one or more perturbations have been detected in the perturbed

adversarial biometric samples 302 (block 608). Otherwise the second user is authorized (block 606).

[045] Figure 7 is a diagram illustrating embodiments performed by the adversarial perturbation detector 28 to detect adversarial submitted biometric samples when the enrolled biometric dataset x_i 32 includes augmented biometric samples x_i' such that the enrolled biometric samples $x_i = [x_i, x_i']$.

[046] In one embodiment, the adversarial perturbation detector 28 may be deployed using a parallel model 700 (also shown in Figure 5) in which the x_i and x_i' of the enrolled biometric samples 30 are input to the transform CNN $k(\cdot)$ in parallel along with the perturbed adversarial samples $x' + \Delta x'$ 302 of the second user to compute $k(x_i)$ and $k(x' + \Delta x')$. The adversarial perturbation detector 28 then computes the distance score and inputs the distance score to the classifier 50 to make a final decision of whether a perturbation as detected.

[047] In another embodiment, the adversarial perturbation detector 28 may be deployed using a sequential model 702 in which the x_i and learned perturbations x_i' of the enrolled biometric samples 30 are input sequentially to the transform CNN 48. In the sequential model 702, z_i is inputted to the transform CNN 48 with the perturbed adversarial samples $z' + \Delta z'$ 302 to compute $k(z_i)$ and $k(z' + \Delta z')$. The adversarial perturbation detector 28 then computes the distance score and uses the classifier 50 to determine whether a perturbation as detected. If no perturbation is detected, then the learned perturbations z_i' of the enrolled biometric samples 30 is input to the transform CNN 48 with the perturbed adversarial samples $z' + \Delta z'$ 302 to compute $k(z_i')$ and $k(z' +$

$\Delta z'$). The adversarial perturbation detector 28 then computes the distance score and uses the classifier 50 to make a final decision of whether a perturbation is detected.

[048] Figure 8A is a diagram illustrating a process for training the adversarial perturbation detector 28 prior to deployment. The training process may begin by generating training examples using adversarial training (block 800) in one embodiment, the training examples include training enrolled biometric samples x_i , training public biometric samples x , and training perturbed adversarial samples $x' + \Delta x'$.

[049] Next, the process learns parameters for the transform CNN $k(\cdot)$ 46 and classifier 50 by inputting the training examples to the transform CNN $k(\cdot)$ 46, where $k(\cdot)$ has learnable parameters θ (block 802). The process generates a training enrolled biometric template $k(x_i)$, a training public biometric template $k(x)$, and a training perturbed adversarial template $k(x' + \Delta x')$. In embodiments, the mapping function $k(\cdot)$ can be decomposed into multiple projections: $k = g_1 \circ g_2 \circ g_3 \dots$, where g_1 can be a layer in a convolutional neural network, g_2 can be another layer in the convolutional network, and g_3 can be a non-linear projection in a support vector machine (SVM).

[050] Next the classifier 50 with learnable parameters σ is used to classify the training perturbed adversarial template $k(x' + \Delta x')$ as a 1 and training public biometric template $k(x)$ as 0 based on training enrolled biometric template $k(x_i)$ (block 804). In embodiments, the classifier may be the same or different than classifier 50. The classifier may be a deep neural network or a threshold-based classifier: $\sum_i d(k(x_i), k(x' + \sum \Delta x')) < \text{third threshold } t'$, it depends on the output of $k(\cdot)$. In either case, the classifier needs to contain an aggregation function (e.g., $\sum_i$) to merge the results of x_i . Based on a result of the

classification, the learnable parameters θ and σ are updated and process may repeat as necessary.

[051] Another operation of the training process is to create the augmented biometric samples 40 by adding learned perturbations 44 to the enrolled biometric samples 30 of the enrolled user as described in block 402 of Figure 4.

[052] Figure 8B is a diagram illustrating an overview of the process of creating augmented biometric samples 40 of the enrolled user (block 402 of Figure 4). In one embodiment, the process of creating the augmented biometric samples 40 is performed after enrollment of each new user, but before the adversarial perturbation detector 28 is deployed to authenticate the new enrolled user.

[053] The process may include by retrieving the enrolled biometric data set x_i 32 for the enrolled user from the enrollment database 24 and performing an operation $q(\cdot)$ on x_i to generate variations of the learned perturbations Δx_i (block 900). Operations in $q(\cdot)$ may increase the intra-class distances between x_i , such as by generating random noise on the enrollment images. In an alternative embodiment, an image may be selected from the public biometric samples x 38, and generate adversarial images based on x_i to slightly increase the distance: $d(k(x), k(x_i + \Delta x_i))$ to obtain Δx_i . The learned perturbations Δx_i are then added to the enrolled biometric data set 902 for the enrolled user (block 902).

[054] According the disclosed embodiments, adding the learned perturbations Δx_i that are added to x_i may increase the difficulty of creating adversarial biometric samples even in the case where the enrollment set is leaked to an adversary during an insider breach so that:

$$\sum_i d(k(\Delta x' + x') - k(x_i + \Delta x_i)) > \sum_i (k(\Delta x' + x') - k(x_i)).$$

[055] A method and system for using an enrollment set to detect adversarial examples in biometrics-based authentication system has been disclosed. The present invention has been described in accordance with the embodiments shown, and there could be variations to the embodiments, and any variations would be within the spirit and scope of the present invention. For example, the exemplary embodiment can be implemented using hardware, software, a computer readable medium containing program instructions, or a combination thereof. Accordingly, many modifications may be made by one of ordinary skill in the art without departing from the spirit and scope of the appended claims.

CLAIMS

We Claim:

1. A computer-implemented method for improving security of a biometrics-based authentication system, comprising:

receiving, by one or more servers, enrolled biometric samples of an enrolled user during an enrollment stage of the biometrics-based authentication system;

applying, using a projection deep neural network, a first function $f(\cdot)$ to the enrolled biometric sample, x_i , to generate a first enrolled biometric template $f(x_i)$ and to project the enrolled biometric samples, x_i , to a common embedding subspace.

creating, by the one or more servers, augmented biometric samples by adding learned perturbations to the enrolled biometric samples of the enrolled user;

receiving, by the one or more servers, during a request for authentication, submitted biometric samples from a second user purporting to be the enrolled user;

applying, using the projection deep neural network, the first function $f(\cdot)$ to the submitted biometric sample, x' , to generate a first submitted biometric template $f(x')$ and to project the submitted biometric samples, x' , to the common embedding subspace;

comparing, by the one or more servers, the submitted biometric samples of the second user to the enrolled biometric samples and to the augmented biometric samples of the enrolled user based on predefined metrics, wherein the comparing comprises computing a distance in the common embedding subspace between the first enrolled biometric template $f(x_i)$ and the first submitted biometric template $f(x')$; and

based on the comparison, determining, by the one or more servers, that the submitted biometric samples of the second user have been modified to impersonate the enrolled user.

2. The method of claim 1, further comprising: storing the enrolled biometric samples x_i , the first enrolled biometric template $f(x_i)$ and the augmented biometric samples as an enrolled biometric dataset.

3. The method of claim 1 or claim 2, comprising:
comparing the distance to a threshold; and
if the distance is greater than the threshold, rejecting the request for authentication.

4. The method of any one of the preceding claims, wherein comparing the submitted biometric samples of the second user to the enrolled biometric samples and to the augmented biometric samples of the enrolled user based on predefined metrics, further comprises, responsive to determining that the distance is less than a first threshold, temporarily authorizing the request for authentication.

5. The method of claim 4, wherein determining that the submitted biometric samples of the second user have been modified is activated only in response to the request for authentication being temporarily authorized.

6. The method of claim 5, further comprising:

responsive to the request for authentication being temporarily authorized, receiving, by an adversarial perturbation detector, the enrolled biometric samples x_i of the enrolled user and the submitted biometric samples of the second user, where the submitted biometric samples comprise perturbed adversarial samples $x' + \Delta x'$; and

responsive to detecting any perturbed adversarial samples, rejecting the request for authentication of the second user, and otherwise, granting the request for authentication and authorizing the second user.

7. The method of claim 6, wherein determining that any perturbed adversarial samples are detected further comprises:

applying a second function $k(\cdot)$ to both the enrolled biometric samples x_i and the perturbed adversarial samples $x' + \Delta x'$ to generate a second biometric template $k(x_i)$ and a second submitted biometric template $k(x' + \Delta x')$, which are in a transformed subspace;

computing a distance F between the second biometric template $k(x_i)$ and the second submitted biometric template $k(x' + \Delta x')$ in the transformed subspace; and

determining that the second submitted biometric template $k(x' + \sum \Delta x')$ is adversarial when the distance F is greater than a second threshold t' indicating that one or more perturbations have been detected.

8. The method of claim 7, wherein the enrolled biometric samples x_i include augmented biometric samples x_i' such that the enrolled biometric samples $x_i = [x_i, x_i']$, and wherein the x_i and x_i' of the enrolled biometric samples are input to the second function $k(\cdot)$ using one of a parallel model and a sequential model.

9. The method of claim 7 or claim 8, further comprising:
- generating training enrolled biometric samples, training public biometric samples, and training perturbed adversarial samples;
 - inputting the training enrolled biometric samples, the training public biometric samples, and the training perturbed adversarial samples to the second function $k(\cdot)$, where $k(\cdot)$ has learnable parameters θ to generate a training enrolled biometric template, a training public biometric template, and a training perturbed adversarial template;
 - classifying, by a classifier with learnable parameters σ , the training perturbed adversarial template as a 1 and the training public biometric template as 0 based on the training enrolled biometric template; and
 - based on a result of the classification, updating the learnable parameters θ and σ .
10. The method of any one of the preceding claims, wherein the enrolled biometric samples comprise images of the enrolled user's face or images of the enrolled user's head.
11. A non-transitory computer readable medium having stored thereon software instructions that, when executed by a processor, cause the processor to perform the method of any one of the preceding claims.
12. A system, comprising:
- a memory;
 - a processor coupled to the memory; and

a software component executed by the processor that is configured to perform the method of any one of claim 1 to claim 10.