HANDWRITTEN DIGITS RECONSTRUCTION FROM UNLABELLED EMBEDDINGS

Thomas Thebaud, Gaël Le Lan, Anthony Larcher

HANDWRITTEN DIGITS RECONSTRUCTION FROM UNLABELLED EMBEDDINGS

Thomas Thebaud[⋆†] *Gaël Le Lan*[⋆] *Anthony Larcher*[†]

[⋆] Orange Labs, France
[†] LIUM - Le Mans University, France

ABSTRACT

In this paper, we investigate template reconstruction attack of touchscreen biometrics, based on handwritten digits writer verification. In the event of a template database theft, we show that reconstructing the original drawn digit from the embeddings is possible without access to the original embedding encoder. Using an external labelled dataset, an attack encoder is trained along with a Mixture Density Recurrent Neural Network decoder. Thanks to an alignment flow, initialized with Linear Discriminant Analysis and Procrustes, the transfer function between the output space of the original and the attack encoder is estimated. The successive application of transfer function and decoder to the stolen embeddings allows to reconstruct the original drawings, which can be used to spoof the behavioural biometrics system.

***Index Terms*—** Alignment flow, Mixture Density Network, Behavioral biometrics, Template reconstruction attack

1. INTRODUCTION

The generalisation of biometric authentication rises concerns about personal data protection. Most recent biometric systems [1] encode biometric data, such as gait sequences [2], voice recordings [3], face images [4], fingerprints [5] or handwritten digits [6–8], into representations commonly named embeddings, thanks to deep neural networks.

For a given user, during the enrollment phase, a set of distinct embeddings is generated from his first interactions with the biometric system, constituting his templates. Then, for any posterior authentication attempt, the user is challenged again, and the newly generated embedding(s) are compared to the enrolment embeddings (templates) to determine if the user should be granted access. The storage and transfer of those embeddings could be vulnerable to theft, and represent a potential breach in the system's security.

In the event of a template database theft, unlabelled embeddings alone are not sufficient to spoof the biometric system. Recent approaches [4,9,10] succeed with at least a black box access to the original embedding encoder.

In this paper, we investigate deep template reconstruction attack for touchscreen handwritten digits writer verification, without access to the original encoder. Due to the nature of the data, we suppose an attacker able to find or handcraft another database of handwritten digits, to train his own attack encoder and compute an attack set of embeddings. It is supposed to help him infer private information from the stolen embeddings, using density matching.

First investigations in this area [11] show that inferring the most represented digits values in clusters of stolen embeddings is possible by statically aligning them to the attack set of labelled embeddings.

This paper investigates further in the context of a One-Time-Password authentication system [6]. We consider here that encryption and template protection mechanisms are out of the scope for this study. We make the hypothesis that the original encoder is known to be based on Long Short Term Memory (LSTM) cells [12] (standard architecture for that kind of sequence data [8]).

We propose to enhance the method of [11] to initialize training of an alignment flow of the distributions of the stolen and attack embeddings in full dimension. Using methods from [13–15], a LSTM based Mixture Density Recurrent Neural Network (MD-RNN) decoder is also trained to reconstruct the attack set of embeddings into their original drawings. Finally, the combination of alignment flow and decoder allows to reconstruct the original sequences from the stolen embeddings and to spoof the biometric system. The main contributions of this paper are :

1. Digit value inference for unlabelled embeddings of handwritten digits in a writer verification system.
2. Alignment flow estimation between the spaces of stolen and attack embeddings.
3. Combination of alignment flow and MD-RNN based decoding to reconstruct strokes sequences from the stolen embeddings.
4. Proof that the statistical alignment is key for the success of the template reconstruction attack.

In section 2, we expose related works about template reconstruction attacks and drawings reconstruction using RNNs. Section 3 presents the attack scenario. Section 4 presents the data used for evaluation. In section 5 we present the experiments and their associated results. Finally, section 6 concludes and presents our future works.

2. RELATED WORK

2.1. Template reconstruction attacks

Modern biometric authentication systems rely on neural network based embeddings to encode the identity of a user, as well as other features. [9] [10] [4] expose vulnerabilities of such biometric authentication systems by reconstructing fingerprints, iris and faces images from templates.

Here we focus on the [4] method, that uses stolen templates and the associated black-box embedding extractor. An artificially generated set of face templates is used to train a decoder to reconstruct faces images from embeddings. Used on the stolen face embedding templates, this decoder succeeds to approximate the original face images. Such images used in a spoofing setting got up to 67% of True Acceptance Rate (TAR) at the Equal Error Rate (EER) threshold (around 1% on the Face recognition system).

Our work differs from [4] as we have no access to the encoder, but a white box access to a handcrafted attack encoder of similar architecture but different weights. We also investigate behavioral biometrics based on handwritten digits writer recognition, that has been shown [8] to have a higher EER (for 1 versus 1 digit comparison) than iris, fingerprints or faces analysis [16]. With a suitable alignment function between the sets of stolen/attack embeddings and an efficient handcrafted decoder, the template reconstruction conditions should be similar to [4, 9, 10].

2.2. Sketch synthesis

Iris, fingerprints and faces images are physiological biometrics: images that can be computed from templates using deconvolutional networks, while the handwritten digits used here are sequences of strokes and represent a behavioral biometry, thus Recurrent Neural Networks (RNNs) should be used for reconstruction. [13–15] proposed methods synthesise sketches or handwriting with RNNs.

Graves [14] proposes a recurrent decoder architecture associated with mixture density networks, for handwriting synthesis. We choose to keep this architecture and losses for our digit reconstruction, with a few modifications to fit the digit characteristics, which will be detailed in section 3.3.

3. PROPOSED ATTACK SCENARIO

We propose a template reconstruction attack of a biometric authentication system based on touchscreen handwritten digits [8]. Let **U** be the set of unlabelled stolen embeddings.

U is composed of N embeddings of dimension 256. It results from the penultimate layer of a bi-LSTM classifier designed to process 2D stroke sequences taken from handwritten digits. The classifier is trained beforehand to predict digit value and identity [8] of users distinct from those of the stolen embeddings. We suppose the attacker able to find or handcraft

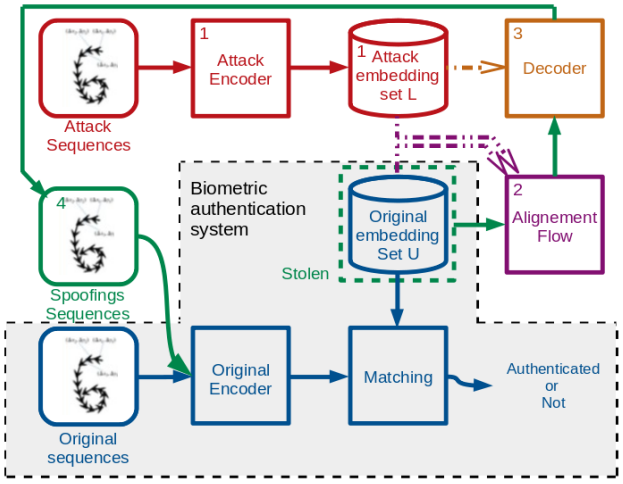

Fig. 1. Illustration of the attack method (best viewed in color)

his own dataset of handwritten digits. This data can be used to compute a set of statistically comparable embeddings and to train an attack decoder.

Our scenario is a 4 steps attack illustrated in Figure 1 where numbers refer to the following steps:

1. Train an attack embedding encoder to get an attack set of embeddings **L**.
2. Find the alignment flow minimizing the distance between the statistical distributions of **U** and **L**, as explained in subsection 3.2.
3. Decode the flow-projected **U** embeddings (in the L space) into 2D sequences of strokes to reconstruct the initial sketches, thanks to a MD-RNN decoder trained on **L**, as detailed in subsection 3.3.
4. Pass the decoded sequences through the original encoder to spoof the system.

3.1. Digit value estimation

[11] proposes a method to infer the most represented digits in clusters of unlabelled stolen embeddings of handwritten digits. An alignment flow matches the distribution of PCA-reduced stolen embeddings with that of PCA-reduced handcrafted labeled distinct embeddings. The flow is designed to pair the 10 clusters composing **U** with the 10 classes of **L**.

Because in this paper, the embedding extractor is trained to predict writer identity additionally to the digit value, we replace the PCA by a Linear Discriminant Analysis (LDA) [17]. K-means clustering is used to divide **U** in 10 clusters that are then used to train a LDA. This LDA is then applied on the same data, to project the embeddings from **U** in a more discriminant space. This different dimension reduction technique focused on the digit clusters and digit classes will allow the [11] method to work identity and digit value trained encoders, as shown in section 5.1.

3.2. Alignment flow

3.2.1. Definition

The distribution of embeddings from each set (GMM_U and GMM_L) is fitted using a Gaussian Mixture model. We define a linear normalizing flow $\mathbf{W} \in \mathbb{R}^{256\times256}$ whose objective is to maximise the likelihood of the projected embeddings $\mathbf{UW}$ given GMM_L and of the projected embeddings $\mathbf{LW^T}$ given the GMM_U [18]. The loss is minimized on couples of mini-batches of $\mathbf{U}$ and $\mathbf{L}$ with gradient descent.

The alignment flow is optimised on the sum of 3 losses :

- $|\log(\det(\mathbf{W}))|$ to target a determinant of 1
- $||\mathbf{U} - \mathbf{W^T W U}||_2$ to keep $\mathbf{W}$ orthogonal
- $\log \mathcal{N}_{GMM_{L\_256}}(\mathbf{UW}) + \log \mathcal{N}_{GMM_{U\_256}}(\mathbf{LW^T})$

3.2.2. Initialization

As pointed out in [18], initialization is key for such alignment problems: the alignment flow often needs some clues. The process of subsection 3.1 provides an estimation of the most represented digit in 10 K-means clusters of $\mathbf{U}$. Thus a Procrustes analysis [19] can be used to compute the optimal rotation matrix between centers of the digit-value related clusters of $\mathbf{U}$ and those of the digit-value labelled classes of $\mathbf{L}$ as pivots.

A Procrustes analysis finds the rotation $\mathbf{W}$ that minimizes the Wasserstein distance between two sets of points $\mathbf{C_U}$ and $\mathbf{C_L}$ in D dimensions (here the centers of the clusters of $\mathbf{U}$ and the classes of $\mathbf{L}$) by solving the equation 1.

$$\min_{\mathbf{W}\in\mathbb{R}^{D\times D}} \|\mathbf{C_U W} - \mathbf{C_L}\|_2^2 \tag{1}$$

If solving this equation in $D = 256$ dimensions for 10 clusters gives an unique computational solution (which will be used for the rest of the paper), it is important to note that from a mathematical point of view, defining a rotation with more dimensions than pivot point gives infinite potential rotations.

3.3. Embedding to stroke sequence decoder

Graves [14] proposed an MD-RNN architecture based on a LSTM decoder using Mixture Density outputs and sequence length prediction for handwriting prediction. We use that architecture for our decoder. For training, the attack encoder is frozen and we add a weak loss constraint on the embedding reconstruction which aims at minimizing the Mean Squared Error (MSE) between an embedding and its forgery (obtained after a decoding-encoding pass). Once trained, the goal of this decoder is to reconstruct the sequences of the projected stolen embeddings $\mathbf{UW}$. It is trained on the attack embeddings $\mathbf{L}$ and their corresponding sequences.

4. DATA

Experimental data is taken from three different datasets: two produced by the University of Madrid [6, 7] and one handcrafted by Orange Labs, containing data from respectively 217, 93 and 66 users. They contain respectively 8460, 7430 and 5850 stroke sequences of variable length, in 2 dimensions, representing digits from 0 to 9.

Data collection was performed in two sessions apart from at least two weeks for each user. Users are divided in 4 groups of equal proportion, each containing a randomized quarter of the users. The first (resp. second) group serves as train and validation data for the original encoder (resp. attack encoder). Validation data for the attack encoder is also used to validate the MD-RNN decoder. The two remaining groups serve entirely for embedding inference and evaluation of the attack. The third group of users constitutes the stolen embeddings $\mathbf{U}$ computed from the original encoder. The fourth constitutes the attack embeddings $\mathbf{L}$ extracted from the attack encoder. Experiments are conducted on 100 randomized data splits of the users, for cross validation purpose.

5. EXPERIMENTS

5.1. Clusters labelling

The method of [11] consists in inferring the most represented digit value in 10 K-means clusters of $\mathbf{U}$. In this work, the embedding extractor was only trained to predict digit values. In the current paper, the embedding extractor is now trained to discriminate between writers and digits at the same time. For this reason, the strict application of the cited method accurately labels the clusters in only 27% of our cross validated experiments. Replacing the PCA by a Linear Discriminant Analysis (LDA) for dimension reduction of embeddings allows us to correctly infer the cluster's digit in 100% of our tests. Training of the LDA requires labeled data which is the case for $\mathbf{L}$ embeddings. For $\mathbf{U}$, we use the labels provided by a K-mean clustering.

5.2. Alignment flow

Following the optimisation method proposed in subsection 3.2, we train an alignment flow $\mathbf{W}$ to project $\mathbf{U}$ in the output space of the attack encoder. The alignment flow is trained using Adam optimizer, with a learning rate of 5.10^{-3}.

To assess the alignment efficiency, the projected $\mathbf{UW}$ embeddings are passed through the digit value prediction layer of the attackers encoder. Digit prediction accuracy is computed according to the ground truth labels of $\mathbf{U}$.

Digit inference accuracy after alignment is presented in the first two lines of table 1. Digit values of $\mathbf{U}$ are predicted with an average accuracy of 96,71% (line 2), similar to the digit value prediction accuracy of the attackers classifier on $\mathbf{L}$ (96.67%, line 4 of the table). This shows that the estimated

#	Configuration	Acc. (%)	EER (%)	TAR (%)	Comment
1	$\mathbf{UW_{procrustes}}$	96.60	20.44	-	Stolen digits inference
2	$\mathbf{UW_{flow}}$	96.71	20.32	-	
3	$\mathbf{L_{reforged}} = Enc_L(Dec(\mathbf{L}))$	85.29	21.69	87.48	Black box access attack
4	$\mathbf{L}$ (*oracle*)	96.67	20.18	100.00	(conditions similar to [4])
5	$\mathbf{U_{reforged}} = Enc_U(Dec(\mathbf{UW_{procrustes}}))$	67.67	49.53	10.42	No encoder access attack
6	$\mathbf{U_{reforged}} = Enc_U(Dec(\mathbf{UW_{flow}}))$	84.77	33.44	21.07	
7	$\mathbf{U}$ (*oracle*)	96.78	20.28	100.00	

Table 1. Digit value accuracy (acc.), EER and TAR at EER threshold for diverse evaluated scenarios

alignment is precise enough to preserve the digit information of the embeddings. It can also be noted that the alignment preserves writer separability: the Equal Error Rate computed within $\mathbf{UW_{flow}}$ (20.32%) is close to that of $\mathbf{U}$ (20.28%, line 7 of the table). It is expected as the flow is forced to remain close to a rotation.

5.3. Embedding decoder training

The MDN-RNN decoder is trained on the $\mathbf{L}$ embeddings and their associated sequences. It is trained for 10 Gaussian distributions, a learning rate of 0.002, no dropout, and uses the Adam optimizer [20]. Performances of the decoder are evaluated in the table 1, line 3. It allows to conserve a reasonably high digit detection accuracy (85,29%, line 3) on the reconstructed sequences compared to performances on the attack embeddings $\mathbf{L}$ (96,67%, line 4). Embeddings $\mathbf{L_{reforged}}$ encoded from reconstructed sequences of $\mathbf{L}$ have an EER (21,69%), similar to the EER on the original $\mathbf{L}$ (20,18%).

5.4. Template reconstruction attack

Once trained, both alignment flow and decoder are successively applied to the stolen embeddings in order to forge spoofing stroke handwritten digits, despite the lack of access to the original embedding extractor. An illustration of the forged digits is shown in figure 2 : the first (resp. second) row being the original (resp. forged) sequences.

Performances of the final reconstruction attack are presented in lines 6 and 7 of Table 1. The most successful configuration (line 6) combines the alignment flow with the decoder, to achieve a TAR of 21.07%. This shows that the reforged digit drawings have in average slightly more chances than a random draw from the general population to be accepted by the biometric system at the EER threshold, the ERR of the system (thus False Alarm Rate) being of 20.28% (line 7).

The contrastive experiment using only Procrustes alignment (line 5) shows that the proposed flow (line 6) is able to preserve some writer identity characteristics: writer separability among the flow-reforged embeddings $\mathbf{U_{reforged}}$ is somewhat preserved, with an EER of 33.44% against 49.53% for the Procustes-only alignment approach.

Obviously, there is room for improvement regarding alignment estimation, since the black-box access based attack exploiting the exact same decoder (line 3) achieves a

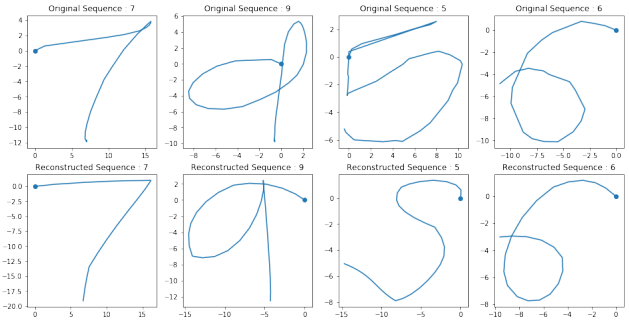

Fig. 2. illustration of original and forged sequences of U

TAR of 87.48%. A *proofing* experiment fine tuning the alignment flow in the whole pipeline (with frozen decoder/encoder in $\mathbf{U_{reforged}} = Enc_U(Dec(\mathbf{UW_{flow}}))$) using MSE objective between $\mathbf{U}$ and $\mathbf{U_{reforged}}$ shows that the TAR can improve up to 73.38% from 21.07% (not shown in table 1). This proves that alignment is critical for the attack success.

6. CONCLUSION AND FUTURE WORK

This paper introduces a template reconstruction attack on a handwritten digit writer verification system, without access to the original embedding extractor, in the event of a template database theft. The attack aims to reconstruct drawn digits from a stolen embedding set in order to spoof the system.

Thanks to a handcrafted set of attack handwritten digits, combination of alignment flow and decoder allows to reconstruct the original sequences from the stolen embeddings and to spoof the biometric system with a TAR of 21.07%, the EER being of 20.28%. In comparison, a standard embedding extractor inversion attack with black box access to the encoder gives a TAR of 87.48%.

These results are promising and show a progression margin related to optimal alignment problem. Such alignment is key for the attack to succeed, due to the lack of access to the original encoder. The proposed method also allows to infer the digit values of the stolen embeddings with an accuracy of 96.71% in average for 100 cross-validated experiments.

Future works will be dedicated to improve the alignment flow, investigate other biometric modalities and other personal data leakage in the event of a biometric database breach.

7. REFERENCES

[1] Anil K Jain, Karthik Nandakumar, and Abhishek Nagar, "Biometric template security," *EURASIP Journal on advances in signal processing*, vol. 2008, pp. 1–17, 2008.

[2] Lily Lee and W Eric L Grimson, "Gait analysis for recognition and classification," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*. IEEE, 2002, pp. 155–162.

[3] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain, "On the reconstruction of face images from deep face templates," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 5, pp. 1188–1202, 2018.

[5] Wencheng Yang, Song Wang, Jiankun Hu, Guanglou Zheng, and Craig Valli, "Security and accuracy of fingerprint-based biometrics: A review," *Symmetry*, vol. 11, no. 2, pp. 141, 2019.

[6] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia, "Incorporating touch biometrics to mobile one-time passwords: Exploration of digits," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[7] Ruben Tolosana, Ruben Vera-Rodriguez, and Julian Fierrez, "Biotouchpass: Handwritten passwords for touchscreen biometrics," *IEEE Transactions on Mobile Computing*, 2019.

[8] Gaël Le Lan and Vincent Frey, "Securing smartphone handwritten pin codes with recurrent neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2612–2616.

[9] Raffaele Cappelli, Dario Maio, Alessandra Lumini, and Davide Maltoni, "Fingerprint image reconstruction from standard templates," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 9, pp. 1489–1503, 2007.

[10] Javier Galbally, Arun Ross, Marta Gomez-Barrero, Julian Fierrez, and Javier Ortega-Garcia, "Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1512–1525, 2013.

[11] Thomas Thebaud, Gaël Le Lan, and Anthony Larcher, "Unsupervised labelling of stolen handwritten digit embeddings with density matching," in *International Workshop on Security in Machine Learning and its Applications (SiMLA)*, 2020.

[12] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] Ruben Tolosana, Paula Delgado-Santos, Andres Perez-Uribe, Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales, "Deepwritesyn: On-line handwriting synthesis via deep short-term representations," *arXiv preprint arXiv:2009.06308*, 2020.

[14] Alex Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[15] David Ha and Douglas Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.

[16] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang, "Biometric recognition using deep learning: A survey," *arXiv preprint arXiv:1912.00271*, 2019.

[17] Suresh Balakrishnama and Aravind Ganapathiraju, "Linear discriminant analysis-a brief tutorial," in *Institute for Signal and information Processing*, 1998, vol. 18, pp. 1–8.

[18] Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig, "Density matching for bilingual word embedding," *arXiv preprint arXiv:1904.02343*, 2019.

[19] John C Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.