**Challenge-Response Authentication using *In-Air Handwriting Style Verification***

Jing Tian, Yu Cao, Wenyuan Xu, Song Wang

**Response to Reviewers**

We would like to thank the reviewers for their comments. Their feedback has assisted in clarifying the materials and improving the presentation of the paper. For easy reference, text that has been added or substantially revised appears in blue in the submitted manuscript. Figures with blue captions are updated images. Specific replies to the reviewers’ comments appear below. The original review text appears in Times New Rome, the response appears in blue, and the text added in the submitted paper appears in *italic and underline*.

**Reviewer: 1**

[ Title and Abstract ]

1. The authors should retitle the paper to make the verification method/task explicit in the title, e.g., "Challenge-Response Authentication using In-Air Handwriting Verification."

**Response:** We appreciate the suggestion and have changed the title to “*Challenge-Response Authentication using In-Air Handwriting Style Verification”*

[ Introduction ]

1. There is too much jargon without explanation (e.g. 'man-in-the-middle attack'), and not enough intuition in this Section. This should be rewritten for a more general Computer Science audience, and the basic idea that in-air writing is being used for verification should be mentioned in the first paragraph, to establish the main contribution and focus of the paper clearly and early.

**Response**: We have added explanations on a few terms, including ‘man-in-the-middle attacks’, ‘stroke segment’, ‘sample’, and ‘primitive’. We have added the following content in the introduction:

*…man-in-the-middle attacks (aka., an attack whereby attackers intercept the communication and could alter the information between the sender and receiver),…*

*To characterize handwriting styles, we introduce the concept of stroke segment: a small segment of finger-tip trajectories that is small enough to serve as writing style element.*

*We use a sample, e.g., a piece of handwriting (e.g., a trajectory recorded in 5 seconds,) to represent a user’s handwriting style.*

*Primitives in the codebook (i.e., vocabulary) are local elements that extracted from writing data*

We also added brief introduction to handwritings and explained why in-air writing is chosen. Please refer to *italic and underline part* in Response to Question (10).

[ Literature Review]

1. The paper would benefit from a more detailed review of techniques and features used to capture handwriting styles. For my taste, there is too much emphasis upon which modalities have been used for verification in the 'Related Work' section, and not enough on techniques related to the handwriting features and pattern analysis used in the paper. In particular, Section 2.3 is a description of the system, rather than discussing related approaches in the literature. The authors should find recent papers (e.g. in ICFHR 2014, ICDAR 2015, TPAMI, Pattern Recognition journal) and dig into the surveys and other papers cited already, and then provide a more detailed picture of how their work relates to the state-of-the-art for handwriting verification.

If the handwriting verification approach is not very novel on its own, the application is still of interest, but the authors need to establish where this work fits within the handwriting recognition literature more firmly. Also, how do the features differ from previous gesture recognition models? Are there common mathematical underpinnings in their models, or are they quite different?

**Response:** We appreciate the suggestion, and we have added additional literatures on the topics of handwriting writer identification and handwriting verification. The writing identification mostly focused on 2D handwriting instead of 3D, yet 3D introduces additional challenges. For instance, writing in the 3D space will result in no obvious stroke information or pen-up/down points, and the touch free input method provides little feedback while writing in-air. Thus, it may result in in-consistency between trials and make it harder to identify a user reliably. Most handwriting verification methods depend on writing content for verification while ours is content-independent. We have added the following content.

*Research on handwriting style has been used for identifying the person who wrote a document or determining whether multiple documents are written by the same person. Handwritings could be obtained offline, i.e., scanned images of handwriting [8], online by a digitizing tablet [19], or in the 3D space [47]. Traditionally, handwriting styles mostly focused on off-line handwritings. Features to be proved to be effective can be categorized into two classes: textural features such as directionally and curvature of patterns in handwritten images, or allographs extracted from local handwritten patterns, i.e., shapes [40]. To extract handwriting styles, feature study techniques fall into two categories: statistical- and codebook- based feature extraction. For statistical method, Bulacu et al. proposed edge based directional probability distributions as features [8]. Schomaker et al. proposed joint probability distribution of angle combination of two ‘hinged’ edge fragments [6] and extended to by [17]. The codebook-based features are derived from Bags of (Visual) Words from computer vision [28]. Primitives in the codebook (i.e., vocabulary) are local elements that extracted from writing data. Then histogram of primitives referred from the codebook as characteristic for a user [7]. Schomaker et al. used the COnnected-Component COntours as the basic elements to capture features of the pen-tip trajectory [41] and then extended to ink-blob shapes [7]. These methods do not necessary work well in our problem because our handwriting is dynamic and contain temporal information (e.g., speed). In addition, existing writer identification methods used a large amount of testing text for testing, and are not suitable for authentication where userability is the key. For instance, [29] used 80 words for a single test and [27] used one paragraph of about 40 Chinese and/or English characters for a single test.*

*On-line handwritings (e.g., handwriting recorded by a tablet) contain temporal information such as the velocity of the pen movements. For content-dependent application, 2D online handwritings are widely used for signature verification. For content-independent applications, the writing style analysis is applied in writer identification. Liwicki et al. presented an on-line writer identification system for Smart Meeting Rooms, used features at the point level and stroke level extracted from text line [29]. Namboodiri et al. used low level shape-based features and Li et al. used the stroke level at probability distribution [26] and then extended to use temporal sequence codes for speed and pressure changes and shape codes for direction [27]. These work analyzed handwriting in 2D space and ours focused on 3D, which we believe introduces additional challenges due to the un-intended issues in the 3D space.*

*Compared to 2D handwritings, in-air handwriting style modeling is challenging, as trajectories are continuously recorded in 3D space, which results in no obvious stroke information or pen-up/down points. In addition, the touch free input method provides less feedback while writing in-air and thus may result in in-consistency between trials. To address it, we utilize both spatial and temporal information of the in-air writing recordings, adopt the concept of the stroke segment, and extract a vocabulary for writing-style-element representation. Instead of only using histograms of primitives, we introduce a co-occurrence matrix that quantifies the transition information between connecting stroke segments, and our results show that the co-occurrence matrix can achieve a better performance than histogram-based methods in our systems.*

1. Having the 'Background' and 'Related Work' sections separated in the paper is confusing. I personally prefer to have the Background material in Section 2, but wherever it appears, it would be clearer to join all discussions of related work together, and then place the system overview in a separate section.

**Response:** We appreciate the suggestion, and we combined the background and literature sections together.

*Please refer to Section 2 BACKGROUND AND RELATED WORK*

[ Figures and Tables ]

1. Figures showing the handwriting are difficult to read (e.g. Fig. 2). Perhaps show the written word at the bottom of each handwritten pattern - the writing is very cramped, and hard to read without this.

**Response:** We appreciate the suggestion. We have added printed explanation to the captions of Fig. 2 and Fig. 7.

1. Fig. 3 - if fewer examples can be provided, with clearer emphasis of key similarities and differences, it will be clearer for the reader.

**Response:** We deleted three rows out of five rows in Fig. 3. We added additional explanation on ‘p’ and ‘b’ part, in caption of Fig. 3 as follows.

*Stroke segments (denoted by different colors on the characters) of ‘b’ and ‘p’ from the same user (either User 1 or User 2) show similar patterns, i.e., similar sequences of primitive indices, but the stroke segment sequences from the different users of the same letter ‘b’ show distinguished patterns.*

1. Fig. 7 - Make User 1/2 a major axis in the table - it is not clear which frames are from which without conscious effort - it should be easy to see using headers and lines to separate the groups. Again, writing out the written words below each image will help the reader greatly here.

**Response:** We appreciate the suggestion. We have added a new column to indicate whether the samples are from User 1 or User 2, and we added the texts of each writing word, which indeed have made the table clearer.

1. Figure 8 - The 'sample level' features are confusing - '2,3,87...' appears to be time indices, although here I believe they represent integer identifiers for codes. I appreciate that the authors are trying to illustrate the bottom-up composition of features here, but some work is needed to make this clearer. The 'Vocabulary' (codebook?) node is also has a confusing location.

**Response:** We apologize for confusing expression in Fig. 8, we represented the style-level features using the general symbols and eliminated the confusing illustrative numbers in histograms. We use rectangular with dash lines to represent various types of features and use an ellipse with solid lines to represent ‘vocabulary / codebook’. To illustrate examples, we show co-occurrence matrixes with example indices and added the following caption.

*Fig. 8 An illustration of constructing a style-level feature, i.e., a co-occurrence matrix. The rectangular with dash lines represent various levels of features and an ellipse with solid lines represent ‘vocabulary.’*

1. Because of the nature of the experiments, Table 2 is confusing; this is a series of single-parameter experiments, but the columns make it look like only 30 samples are used with L\_s = 2000, etc.

**Response:** It is difficult to illustrate all results in a table, and thus we only change one parameter while using the default ones for the rest in each set of experiment. To clarify the issue, we added the default parameters to the caption of Table 2.

*We varied one parameter at a time and use the following default parameters: one sample contains no more than 2,000 frames; Each experiment uses 30 samples per subject for training and the rest for testing; Each stroke segment contains 12 frames; Each experiment has a vocabulary size of K = 200.*

[ Methodology ]

1. The motivation for not using 2D writing is a bit unclear - cleanliness is cited as an issue, but if a user is authenticating from their own touch device, is that a serious concern? Ideally, there would be some comparison between results obtained for 2D (e.g. on-tablet) and 3D finger writing. It would be of interest to see whether the 3D makes it easier or harder to verify a user relative to 'normal' handwriting verification.

**Response:** The purpose of this paper is to investigate the feasibility of using 3D writings and extract writing style from it for authentication. The application scenarios are mainly for contactless environment in the public (e.g., gate access). We wish we could compare our approach with 2D writings. However, there are several differences that make it unfair to compare both. 1) the current applications of 2D verification are mostly content-depended yet ours is content-in depended, 2) even though some existing applications are content-independent, they require heavy training for an improved verification rate and thus demand time and enrolling samples, not so user-friendly, 3) meaningful testing results should use similar data sets. Although we have thousands in-air handwritings samples, they are not the same one from any existing 2D signature varication database, or 2D online text recognition database. Thus, we did not compare 2D writing with 3D writing.

We added the underlying reasons on why choosing in-air gestures in the introduction section:

*We envision to design an in-air-based authentication system with the following features. First, it works in a contactless manner, which has the benefit to elliminate hygiene concerns and is resilient to smudge attacks [4], i.e., finger smudges (due to oily residues) on a touch screen can reveal passwords. Second, it can be used as a complementary method to be used when traditional biometrics-based authentication is inapplicable. For instance, fingerprints have limitation and are inapplicable in several scenarios, e.g., fingerprints are unsuitable to users with dirty, greasy, or worn-out fingerprints due to their professions (e.g., miners). Third, it should be applicable to the public scenarios and thus should be resistant to shoulder surfing (i.e., observing) attacks, and acoustic environmental noises.*

*We propose to use in-air handwriting style, which is unique to a user, as a new biometrics for authentication. We utilize a depth motion sensor, Leap Motion controller [15], to record in-air fingertip writings as shown in Figure 1. Handwritings, in the form of scanned images or recorded by tablet, have been used as signature verification or writer identification for authorization or forensics purpose [40]. Compared to writings on tablet, the in-air writings introduce a larger amount variability which we believe contain a richer set of biometric yet introduce extra intra-variance for multiple trial of writings from the same writer, and thus impose challenges for authentication. Despite of the challenge, we choose in-air handwriting because it has the following advantages. First, it does not require contact to any device. Second, handwriting style is essentially a behavioral biometrics and thus has limited privacy issue. Third, given the rich combination of letters and numbers, the continuously written words are challenging to be synthesized because we believe that imitating arbitrary handwritings in the 3 dimensional (3D)-space is ambitious. Thus, it should be resistant to shoulder surfing.*

1. More needs to be said about the 'Content matching' step that insures the words asked for have been written - as the authors note, existing recognizers are imperfect, but the effect of this on the usability of the method is not discussed in the paper. My sense is that it may not even be required for the approach chosen (i.e. using 'global' temporal writing features). The authors should explain more clearly why this is or is not the case.

**Response:** We regret about the confusion. One of the design goals of our method is to eliminate replay attacks and thus we design the challenge response method, where a user is required to write a given content. In fact, our approach does not have to verify whether a user has written what was asked for accurately, it should be varied that the handwriting is not a replay. To clarify and discuss related issues, we added the following content in the introduction and discussion session. for the usability concerns while introducing Challenge-response (by content matching).

In introduction:

*Compared with authentication using biometrics, adding the procedure of challenge-response will introduce an extra amount of time to verify the response, nevertheless, we believe the challenge-response can reduce the possibility of relay attacks.*

In discussion:

*Content Matching. In this paper we focused on how to apply the in-air handwriting style for authentication. For security concerns, we introduced challenge-response procedure. We believe that the matching between the challenge and the response (i.e., handwriting content matching) can utilize approaches designed for handwriting recognition. Much work has been proposed to recognize contend from off-line writings, on-line writings (2D), and in-air writings (3D). These content matching approached can be added to fulfill the proposed method. In addition, as a direction for future work, it is worth performing usability studies to understand the trade-off between security benefits and extra effort imposed by the introducing of CR mechanism.*

1. The terms 'Component' and 'Component-Level' are unclear. These are stroke segment features, correct? I would call these 'Stroke Segments' instead. For people in document analysis and recognition, 'component' suggests 'connected component,' i.e. a connected group of black pixels. 'Sample-Level' might also be clearer as 'Global' or 'Style.' The individual frames are technically also samples, whereas here the authors intend to mean 'all samples/the whole sample,' but this isn't clear on a first read.

**Response:** We appreciate for the suggestion, and we agree that the “stroke segments” and “style-level” are a better choice than ‘component’ and ‘sample-level’. We have changed them in the paper.

We clarified the “sample” in the introduction section.

*We use a sample, a piece of handwriting (e.g., a trajectory recorded for 5 seconds), to represent a user’s handwriting style.*

[ Evaluation ]

This is one of the weakest parts of the paper.

1. First, the system is not compared to any other techniques or baselines (e.g., other techniques for handwriting style verification). The authors should make some comparison with existing methods and/or simple controls to better characterize the behavior of their verification system, and better place the performance, strengths and weaknesses of their approach within the wider research literature. Also, some comparison with a 2D-based verification (e.g. using writing on a tablet) would be appropriate, and help readers better understand the strengths and limitations of the proposed technique, both algorithmically, and in terms of user familiarity and effort.

**Response:** We appreciate the suggestion. Please refer to the answers to Question (10) for the reason that we didn’t compare what we proposed with the existing 2D-writing-based methods.

We did compare the method that uses features on histogram (widely used in literatures) to the proposed co-occurrence matrix methods. The results are shown in Table 2.

In the related work, we discussed the relationship and differences between our work and the previous work. We didn’t compare the work numerically, because we believe that the error rates are not directly comparable if only different data types and data sources are available. Thus, we have to focus on exploring the feasibility of using in-air handwriting style for authentication.

1. In Section 5.1.2, the authors mention they vary one parameter at a time in their experiments, but then do not provide the values used for parameters that were not varied. For example, when trying to identify the ideal sample length. It isn't clear why the stroke segment ('component') lengths are tried with values 8, 12 and 20 - this is neither a linear nor logarithmic set of parameters. As a result, the effect of different parameters on performance, and their interpretation is very difficult to make.

**Response:** We appreciate the suggestion. We have added default parameters to the caption of the table as below, and we added one extra experiment result where a stroke segment length of 16 was used to the table, to make segment lengths increase linearly.

*We varied one parameter at a time and use the following default parameters: one sample contains no more than 2,000 frames; Each experiment uses 30 samples per subject for training and the rest for testing; Each stroke segment contains 12 frames; Each experiment has a vocabulary size of K = 200.*

1. 'Random Insider' does not seem correct as used in Section 5.1. This seems instead to be a within-subscribed user test, for all 24 users. I would remove 'Random' here, as it suggests either a random sampling protocol (not used), or a 'random' outside user trying to imitate a user.

**Response:** We regret about the confusion. The term “random” does not mean an attacker that comes randomly out of nowhere but an insider that does not intend to learn or have learned useful information from the victim. Thus, he/she randomly writes.

1. For 'Random Imposters,' the way in which users are selected seems inconsistent with the previous experiment. Why split in half, when you can simply rotate which user acts as the attacker instead, similar to the previous experiment? This shift in data-splitting method is confusing, and not well-motivated. It would be better to make this similar to the previous experiment, as it is simpler, more consistent, and easier to compare to the previous experiment.

**Response:** We appreciate the suggestion and agree with the reviewer, we re-did the experiments as suggested and have updated the results. We note that the result does not affect the final conclusion of the paper. We added the following to update the experiment setup.

***Random Impostors****. Classifiers are trained with training data from 23 subjects and tested using the data from the remaining subject only. In this scenario, the attackers’ samples are not part of the negative training samples. Among 24 subjects, we choose one of the subject as a random impostor and their samples are never a part of the training sets. We use the remaining 23 subjects as the legitimate users for training, and then collected the result assuming the chosen subject as an imposter. We repeat the experiment 24 times by rotating the impostor role. Note that we call such an attacker as random imposter because he/she is not part of the trained user and does not contain any useful information of the insiders.*

1. For the 'shoulder surfing' experiments, the amount and nature of writing performed seems unnatural (the paragraph repeated three times), and the reason for choosing the 4 victims is not explained - one can imagine biases resulting from this. It is also unclear why one would want to omit the victim's data from the training set. The claim that observation cannot help attackers imitate a victim does not feel well-supported, and seems unlikely, in part again because how users are sampled has shifted again without explanation or justification. Forgetting the average performance metrics for a minute, in Figure 13 for one user there \*does\* appear to be a benefit for observation as well, and I wonder if this has to do with how similar the challenge text excerpt is to that in the training set. More needs to be explained, and likely the experiment revised and repeated so that the outcome and analysis are clearer.

**Response:** We tried to let an attacker to mimic all users. However, after mimicking four victims, he started to feel confused about the writing style and was too tired to perform the attacks vigorously. To keep the attacker energetic while learning, we decided to limit the attack to 4 victims. In our experiments, without loss of generosity, we randomly chose 4 victims. Note that an alternative is to choose 4 random victims for each attacker. However, we did not to change attackers and victims, since it is better to change one factor in each experiment. Based on the results, we can see one we could conclude that the “observing” may not always help the attacker, and sometimes the learning opportunity will help the mimic attacks slightly. We have added the following to clarify the conclusion.

*In this experiment, out of the 24 subjects, we randomly select 4 as victims, and invite 7 attack subjects.*

*From the results, we observe that learning the writing content does not necessary always improve the impersonate attacks and may have limited help in impersonating users in MoCRA.*

1. The defined robot arm behavior seems arbitrary and fragile. Because co-occurrences are used as features, it is not surprising that just 'joining' characters from a fixed vocabulary of character strokes performs poorly. Further, Hidden Markov and other generative models are more likely to produce representative samples that will be more representative of the victim, and closer to a 'high resolution attack.'  Restricting the robot to a fixed vocabulary of characters, some of which may not be from the same word and so inconsistent in style will make the generated samples likely to be 'unnatural.'

**Response**: We appreciate the suggestion and believe that the robot attack scenario causes too much confusion and is a specific yet rare scenario. Since the paper focuses on human attacks, we have deleted the discuss.

1. Also, note the strong effect of large vocabulary size in the first experiment - does this 'protect' users in Table 2 simply because with a larger vocabulary users are less likely to have similar code sequences? This seems to contradict the claim that the writing 'style' features are text-independent.

**Response:** The larger vocabulary size represents a larger amount of variability and we believe the vocabulary can be obtained by enrolling enough users. The text-independent method does depend on the primitives extracted from the training text, but does not depend on the content of the testing text. Thus we believe the primitives represent the writing elements of text input but not the contend of the text input.

**Reviewer: 2**

Recommendation: Author Should Prepare A Major Revision For A Second Review

Comments:

1) The first task of the system would consist in providing a challenge word and recognizing whether the user has written the challenge word or not. It is very disturbing since you do not address this task in this article (called 'matching the content'). Moreover, it is not clear how this task is connected to the user authentication task addressed in this paper. Is the word written according to the challenge word (the response in fig. 4.A) also the input for the authentication task (Fig. 4 b)?

**Response:** Please refer to the answer to Question 11 of reviewer 1.

2) Your approach for creating writing components from samples, is simple (just concerning a fixed numbers of frames). Extracting fragments of writing could also be done according to directional changes or local minima in the position. This should be addressed in your future work.

**Response:** We appreciate for the suggestion. Extracting fragments of writing could be very interesting and we have added the future work section as below.

*Handwriting Segmentation. We used a fix number of continuous sampling frames to represent the primitives of handwritings. However, based on the trajectories’ inner variations, one can extract segments according to a few features (e.g., curvature, direction, speed) so that each segment will contain unique style information of a user, while the length of each segment is not necessarily the same. Thus, as a direction for future work, it is worth exploring new segmentation methods for enhancing writing style modeling.*

3) Fig. 3 « component index » : I assume you mean « primitive index» .

**Response:** We appreciate the suggestion and we have corrected it.

4) Codebook : You should illustrate more by showing components which belong to the same codebook element.  The main point is that your clustering is adapted to the users. A separate training set should be used just for building the codebook. The training set for the codebook should include users distinct from training and test users. Since in real scenarios, you do not retrain the whole system when a new user is added: you only train a new SVM for this specific user.

**Response:** We appreciate the suggestion on using extra database for codebook training. As far as practice is concerned, it is ideal to use a separate database, given that we have enough users enrolled. In our experiments we collected samples from 24 subjects, and thus our training data set is not large enough to create databases for both generating effective codebook for capturing a variety of writing styles, or maintaining a statistically significant results for authentication evaluation. Thus, extracted vocabulary based on training data prior to classifier training, and we believe it suits for a relatively small user group. To clarify the issue, we have added the following in the discussion section.

***Vocabulary Generation.*** *We extracted a vocabulary based on the training data prior to classifier training. Ideally, pre-trained vocabulary with a separate database can avoid new vocabulary training whenever a new user enrolls. For scenarios of a large group of users, the pre-training on the vocabulary in a separate database can improve the computation performance at the training stage. However, our experiments involved 24 subjects on campus, and is not large enough to generate a representative vocabulary for various writing styles while maintaining a statistically significant results of authentication evaluation. We note that this limitation can be eliminated by extra data collection*.

5) The test impostors should not be known from the system (i.e.,. the codebook, the SVMs) while training. Thus to my opinion, the « random impostors » scenario only is valuable, not the 'random insiders' scenario.  In the 'random impostors 'case, it is not clear if the actual impostor samples are used for building the codebook or not.

**Response:** We apologize for the confusion. We use 'random insiders' to refer to the attacker who is a legitimate user (his/her writing style is already trained in the model) and he/she wants to impersonate another user from the group.

The codebook was built during the training session and used the training data. Thus, the actual impostor’s samples were used for codebook building.

6) Training/test protocols are not sufficiently described. The training condition of the SVM should be described, especially which fraction/session of the user data and which fraction/session of the other user data (impostor data). You should provide a complete description about the use of these data since it seems that impostors that you use for testing, are also used for training the SVM (random insider scenario). Even in the case of « random impostor scenario », the impostors are distinct but the codebook may have used the impostor samples during its construction.  Thus for avoiding mixing train and test data, and for a better generalization across users, the data used for building the codebook (you call it « vocabulary ») should be built from data, distinct from the training and test data. (see Bulacu et al., Text-independent writer identification and verification using textural and allography features, IEEE PAMI, 2007.). A distinct database could be used.

**Response:** To avoid the confusion of training and testing data, we added a briefly introduction to explain each experiment setting.

**Random Insiders.** *Classifiers are trained using the training data from 24 subjects and tested using the remaining data.*

**Random Impostors.** *Classifiers are trained with training data from 23 subjects and tested using the data from the remaining subject.*

**Results for Observing Attacks.** *We use classifiers of 4 victims which are trained in* Random Insiders *part and tested by 7 attackers’ data, whose data never be part of training.*

About the codebook, the listed reference has writings from hundreds of subject and yet we only have 24 subjects. Please refer to the answer to Question (4) for pre-computing issues.

7) The segmentation of the data into words (word segments) is not clear at all. How do you discard the first movements just before « writing » the first word.

Fig. 6 shows f.i. points after the last word. Why isn't there extra points before the first word?

**Response:** We admit that the starting part of writing pieces contain both normal strokes and the unique starting patterns. Luckily, there are only a small number of the starting part and thus we did not remove these part.

Fig. 6 is used for illustrating the transition segmentation and thus we only showed the transition segmentation.

8) The fact that you collect data during different sessions is a positive point since data may vary from one session to the other, even when collected from the same user. However, you do not detail which sessions are used for training, and which are used for testing. Describe more precisely training and test conditions in a Table: length of the training/test samples (in sec), which session(s) where used. etc...

**Response:** To avoid the confusion of training and testing data, we added a briefly introduction before each experiment setting, and please refer to the response to Question (6) for detailed information. Note that the length of the training/test samples were fixed once all parameters are chosen.

9) what is the use of the overlap?

**Response:** The overlap of different stroke segment (component) is to eliminate the need to align the beginning of each stroke precisely. For instance, the peak would be ignored if the previous segmentation ends at the peak, and the next segmentation starts at this peak (without overlap).

10) Revise Table 2: It would be useful to add the total writing time for each column (this is of practical use). What is the unit of L\_s: number of frames? In this table, you can also omit the term « varied ».

**Response:** We have added total writing time for each column (*Writing Time for Training (Total))*.

11) There is a contradiction between saying that the SVM is binary and applying a threshold. (line 24, page 10). The existence of this threshold should be explained (distance...).

**Response:** We apologize for the confusion. “Binary” does not mean that the direct output of SVM is either true or false, instead the classification result is binary, i.e., the handwriting is written by a given user. The output of the SVM is a numerical number and we obtain classification results by comparing the numerical number with a threshold. By varying the threshold, we are able to obtain various classification results, which produce an ROC curve. Note that the ROC curve is a good method to guide threshold chosen, and an EER is a standard measurement for security performance, which is the equal error rate of false positive rate and false negative rate while varying threshold.

12) ROC curve: make a figure with an y axis greater than 1 in order to better visualize the curve.

**Response:** We showed ROC curves with axes in the range no larger than 1, because the range of ROC curves is between 0 to 1. To avoid misleading illustration where ROC is larger than 1, we didn’t use a y-axis larger than 1. We note that the idea position is (0% FPR, 100% TPR) point.

13) Please provide details about the « noisy » environment mentioned in line 45, page 12, and how you address it.

**Response:** We regret about the confusion. We mean acoustic noises instead of measurement noises. We have clarified the issue in the paper.

14) There is a confusion between « authentication » and « identification » see Fig. 4 b)

the output of SVM classification is not user identity but a binary reply (user or impostor)

**Response:** We appreciate the suggestion. We have replaced identification with <authentication> and sticked to a binary reply in Fig.4.

15) Please provide illustrated samples of robot attack.

Response: We appreciate the suggestion and believe that the robot attack scenario causes too much confusion and is a specific yet rare scenario. Since the paper focuses on human attacks, we have deleted the discuss.

16) theta\_yz appears in Fig. 9 but not in the list of features.

**Response:** We have updated the Fig.9.

17) English has to be checked. Many typos. "authentication", 'have attract' etc...

**Response:** We appreciate the suggestion and have gone through the paper a few times to eliminate as many typos as possible.