Response to Reviewers:

We would like to thank all the reviewers for their insightful comments. We have made major revisions to address their comments. Below are the item-by-item responses to each reviewer’s comment. For clarity, in the following we include each review comment in black font and then our response in blue font. In the manuscript, we also highlight the revised/added texts in blue font.

**Reviewer: 1**

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

Comments:  
Thank you to the authors for the revised version of their paper.  I still believe that this paper probably makes a valuable contribution, but there are still some substantial clarifications and corrections needed before the paper will be ready for publication. The presentation and interpretation of the method and results still need to be clarified in particular.  I am recommending another Major Revision as a result.

My comments on the revised version of the paper are provided below.

1. The related work is greatly improved with the inclusion of work in 2D handwriting verification tasks, and I was glad to see work by Liwicki, Jain, Jawahar, and Schomaker included here, as is appropriate. Content matching in Section 3.2 should be moved earlier to the Related Work section; this is not part of the studied method, so it is confusing to introduce this in Section  
   3.2, and not earlier in Section 2.2.

We appreciate the suggestion. We have moved content matching (original Section 3.2) to the end of Section 2.2. Please see line 6, right column, page 4 of the revised manuscript. We have modified the text as follows:

*Several literatures focus on content matching. For instance, online handwriting recognition has been studied in pattern recognition community since 1980’s, or similarity comparison based on handwriting recognition. Online handwriting recognition can achieve an accuracy of more than 85% on pure cursive writings or 95% on others [26], [32], [38], [46], while similarity comparison can achieve a higher accuracy, since it does not require the specific recognition of each letter, but a confidence score that shows similarities between two data. Furthermore, recent study [20] achieved a recognition accuracy of 97.59% for in-air English character recognition, with data from two depth sensors, Kinect and Leap Motion Controller. The focus of this paper is to study the biometric built on handwriting motion, and will utilize prior work on content matching. In particular, for security purpose, we introduce challenge-response mechanism to the MoCRA. The user writes the content that the system provides (the Challenge). After receiving the input from the user (the Response), the system checks if the input content is the same as the system expected. We call this step as content matching.*

1. In Section 3, one cannot simply claim that 'existing technologies' can be used to recognize whether the word drawn has been correctly written - there will be an associated error profile added in this case. More concerningly, the paper cited is from 1990, and is introduced in such a way as to suggest that this is a solved problem, which is not. Better integration with the other related work in the paper is needed here, and/or the issue should be addressed differently.

The problem was fixed by adding more related work and this part is moved to Section 2.1 as the reviewer suggested in previous item.

*… on others [26], [32], [38], [46],… and*

1. In Section 3, the authors should acknowledge the assumption that the user writes what is shown, and not just words they know will be easiest to imitate a chosen victim's style for, and that people do not 'team up' so that one repeatedly draws in a simple, easily repeated way that an unsubscribed user can imitate.

We apologize for the confusion. At the bottom of right column of page 4, we added the following sentences to clarify this problem and added the assumption.

*The first step is not the focus of this paper and can be accomplished by utilizing the prior work explained in Section 2. In particular, we introduce challenge-response mechanism to the MoCRA for security purpose. The user writes the content that the system provides (the Challenge). After receiving the input from the user (the Response), the system checks if the input content is the same as the system expected. As such, an attacker cannot `replay’ handwriting performed in the past. Based on the assumption that the replay attack would not be an issue for MoCRA, …*

1. In classification (Section 5.4): details of how SVM parameters were tuned should be provided explicitly (e.g., what values for penalty parameter C were tried, that a grid search over vars X,Y,Z was used to optimize, etc.). These are \*not\* minor details, as they affect and limit other people's ability to replicate results in the paper. Related to this, another motivation to use SVM is its documented efficacy when only small training samples are available.

In the right column of page 10, at the end of Section 5.4, we added the following paragraph on the details of tuning SVM parameters:

*In our experiments, we choose the (Gaussian) radial basis function (RBF) kernel. We use grid search method to optimize RBF SVM parameters C and Gamma [9] and other parameter are set as default. Finally, we set Gamma as 0.01 and the penalty parameter C as 5000.*

1. In Section 6.1, how many times did each participant write the 19 paragraphs?

*Once.* We have added this information to line 5 of Section 6 (page 10).

1. Is this the '30 samples' on page 11? This is an important detail.  The explanation of how data samples were split, in what specific ratios, or if cross-fold validation was used needs to be made clear in 6.1.

The 19 paragraphs include the 30 training samples, and include more samples from each subject used for testing. We did not use cross validation, as we prefer to simulate a practical system where the first batch of enrolled data are usually used for training. We clarify this in Section 6 (page 10, the last paragraph before Section 6.1).

*We use the first few paragraphs for enrollment and the rest for testing without cross validation.*

Following it, we added the following sentence to clarify the specific number of samples used for training and testing:

*We get 3256 samples in total. Among them, 720 samples are for training, and others are used for testing. Then, we collect data from 7 subjects who act as observing attackers. In total, 84 attack samples are collected.*

1. 6.1.2: Each classifier is binary, I believe. The indication that classifiers are trained with 'data from 24 subjects' makes unclear whether this is 24 classifiers, one per subject, or something else (despite the explanation in 6.1.1).

We apologize for the confusion and the reviewer’s understanding matches with our algorithm. To clarify, we have modified the descriptions in 6.1.2 to the following:

*Twenty-four classifiers are trained using the training data from 24 subjects and are tested using the remaining data. For each classifier, one subject act as the victim (positive training samples) and the rest 23 ones act as attackers (negative training samples).*

1. In figures 11 and 12, what sentences/cases tended to lead to TP/FP errors? Is the worst performance in each case for the person with a style closest to the 'average writer' (i.e., average co-occurrence matrix)?

The worst cases are mainly caused by some of its samples that show similar styles (i.e., similar co-occurrence matrices) to a few other subjects, instead of having average co-occurrence matrix.

we add discussed in 6.1.2 to the following:

*The worst cases are mainly caused by some of their samples that show similar styles (i.e., similar co-occurrence matrices) to a few other subjects.*

1. In the 'insider' attacks, how are the 'insiders' random? Isn't every participant used to attack every other?  That isn't random by definition - it's exhaustive given the available data.  I think what you mean by 'Random' vs. 'Observing' is 'Non-Observing' vs.  'Observing' - one variable, indicating whether or not the attacker is trying to imitate the victim, rather than just  
   write in their own handwriting style.

We appreciate the suggestion, and we have changed the “Random” attacker into “*Non-Observing*” attacker in the revision,

1. I'm afraid that the difference between 'imposters' and 'insiders' is still unclear, assuming that each SVM is binary, trained for a single user. Is the 'insider' problem a 24-class problem, and 'random imposter' repeated instances of two class problesm?  For the 'insiders,' given there are 24 participants, what is 'The remaining data...for testing'? This doesn't make sense to me. It  
   looks like you may be testing on training data, which as Reviewer 2 points out is a serious concern for pattern recognition tasks; we don't want to 'take the test' after already 'seeing the answers,' as it then says very little about patterns learned for the input space as a whole.

We apologize for the confusion. To clarify it, we rewrote this part from the aspect of each victim subject, which corresponds to one ROC curve in the result figure.

***Non-Observing Insiders. (Paragraph 1 of Section 6.1.2)***

*Twenty-four binary classifiers are trained using the training data from 24 subjects. For each classifier, one subject act as the victim (positive training samples) and the rest 23 ones are acted as attackers (negative training samples). To test the classifiers, the data from the 24 subjects that not used for training are used for testing.*

***Non-Observing Impostors. (At middle of left column, page 12)***

*For each subject (one of the 24 subjects, si) that acts as a victim, we assume another subject (one from the rest of 23 subjects, sj ) is an impostor without knowing any information of the victim. We label the rest 22 subjects as srest. For each victim si, we train 23 classifiers and show average result as one ROC curve in Figure 12. To train classifiers for si, we use si’s training data as positive samples, the srest’ training data as negative samples, while sj ’s data are not included. In other word, the 23 classifiers for si have the same positive samples from si, but negative training samples are partly different as the srest sets are different since different sj is chosen as an impostor. To test classifiers of si, we label all data from the attacker sj as negative samples, and the testing data from the victim si are positive samples. The result of each curve shown in Figure 12 are averaged on testing results to the 23 classifiers.*

Methods:

1. - Reviewer 2 raises an important issue: while users are asked to write a word shown, if I understand correctly, a savvy attacker could write some word - \*any word\* likely to match the style of another user; provided the style features are consistent with the other users' distribution, they would pass the challenge. This would be precisely because, as pointed out in the Conclusion, that the system matches:

" 'how they write' instead of 'what they write'

This is an important limitation and significant vulnerability of the system, which needs to be clearly identified and pointed out as important future work, including in the Conclusion of the paper.

We agree with the reviewer, and added the following sentences in Section 7.1:

*Essentially, the implemented MoCRA in the paper authenticates based on `how a user writes’ instead of both ‘how a user writes’ and `what a user writes’. As a direction for future work, it is important to investigate content matching and its impact on the accuracy of the system. We suspect that content matching could improve the security, since our evaluation results do not reject the attack samples that have the same hand writing style distribution as the legitimate one but formed from different content, such cases should be rejected as the earlier stage -- content matching.*

We also added the following sentences in Conclusion:

*Our scheme authenticates users based on `what they write' and `how they write'. In this paper, we focus on ` how they write’, and `what they write’ will be discussed in the future work.*

1. - Some content on the extent to which \*what\* people write influences \*how\* feature/code distributions are defined for users should be briefly discussed in the paper. Obviously if a user's features come from just one or two physically similar phrases, the user's writing will be easier to imitate by an attacker; this independence between 'what' and 'how' in the paper needs to be carefully qualified/explained.

We appreciate the insightful suggestion. The writing styles are independent of `what’ because writing styles are extracted from stroke segments instead of letters. Thus, as long as enough stroke segments are used for training and testing, MoCRA can authenticate a user reliably. In terms of an attacker, the observation is that an attacker cannot mimic every stroke segments well, and cannot produce a consistent writing style. We added the following paragraph in Section 5.4 to clarify this issue:

*We split the input data into two parts: training data and testing data. The training data are used to model the writing style of a user, and a testing sample is used to test a user's writing style, thus a sample should include writing style representations of the user. Since the writing styles are extracted from stroke segments instead of letters. Thus, as long as enough stroke segments are used for training and testing samples, MoCRA can authenticate a user reliably. We assume the content of the challenge has a fair distribution, so the length of a sample (i.e., how many letters or how many frames in a sample) indirectly shows the representations of the writing style. We will examine appropriate length in Section 6.*

1. - The normalization of features needs to be clarified on page 9 – how is single feature distribution forced to be univariate Gaussian? Are the mean and standard deviation computed, after which every feature is assumed to be normally distributed?

We agree with the reviewer. We added the following sentence at the end of Section 5.1 (page 9) to clarify it:

*For each dimension, we first calculate a mean value and its standard deviation, then the values are converted to new values that fit into a standard Gaussian distribution.*

Typos and Other Corrections:

1. - Page 1: Unclear: 'where a user obtained the secret' - what secret?

We updated it to:

*where a user obtained the secret* *(i.e., the response)*

1. - 'HandWritings' -> 'handwriting samples' (p. 1)

Fixed.

1. - page "Our system can not only exclude insiders but also reject imposters" – this is not a contribution of a paper - this is a requirement for the verification task!

We deleted this statement.

1. - page 3: "utilize [a] more sophisticated features" - remove 'a'.

Fixed.

1. - page 5: '(widely used in the literature) - provide a list of example references;  
     don't force the reader to find these on their own.

We added several references:

*Instead of using the histogram or probability density functions (PDFs) of the individual stroke segment indices [7], [27], [43], ...*

1. - Move Table 1 closer to where it is mentioned in the text.

Fixed.

1. - In Section 5.3 - 'AS the sample length, increases...the shape of the co-occurrence matrices remain similar' - either refer forward to supporting evidence, or perhaps better, omit this statement for discussion in the results portion of the paper.  Abstract:

We omitted this and the result section shows the trend.

1. - Figure 12: Zoom in on the upper-left corner; this is unreadable. (e.g. from 0.8 - 1.0 on the y-axis; 0 -0.2 on the x-axis). Similarly for Fig. 11.

We zoomed in both figures into 0 to 0.2 on the x-axis and 0.8 to 1.0 on y-axis.

Reviewer: 2

Recommendation: Author Should Prepare A Minor Revision

Comments:

1. [2D-3D capture] The article has improved but still the difference with 2D data capture should be highlighted. You should add explanations about the « unique starting patterns » ( starting parts ?) and tell that starting parts are kept. This makes a difference with tablet data where the recording starts when you touch the screen. With your 3D capture system, when does the data recording start? do you first stay motionless before starting to write the 3D word « when » ? In fig. 6b how do you process the last points after word « comes » (frames between 900 and 1000)? Do these points belong to the next word? what happens when you reach the last word ?

We appreciate the suggestion. We added the following paragraph at the end of Section 4.3 to clarify this issue:

*MoCRA sets hotkeys for starting a recording and stopping a recording. Before starting, finger(s) stay motionless and start writing after the press of the start hotkey. MoCRA introduces a few data frames before the intent writing and includes these frames in the first word of a recording. The recording will be ended if any of the two scenarios happen: the finger(s) are out of the recording area or the stop hotkey is pressed. Thus, it is possible that extra frames are added to the last word before the recording ends completely.*

1. In Fig. 6 c word « it » : I do not understand how this word was drawn :  especially it seems that the point above « i » was drawn before the « i » itself.

We checked with the subject, he did draw the point before drawing the lower part of letter “i”. In the revision, we clarified this point in the caption of Fig. 6.:

*Note that the subject draw the point before the lower part of the letter `i'*.

1. [training-testing data] p. 11, lines 29-32 : « Classifiers are trained using the training data from 24 subjects and tested using the remaining data ». Which remaining data? Do you mean testing data from the 24 subjects?

We apologize for the confusion. We clarified it by revising the first paragraph of Section 6.1.2 to:

*Twenty-four classifiers are trained using the training data from 24 subjects. For each classifier, one subject act as the victim (positive training samples) and the rest 23 ones are acted as attackers (negative training samples). To test the classifiers, the data from the 24 subjects that not used for training are used for testing.*

1. please provide details about number of words or samples for training, and testing respectively.

We agree with the reviewer. At the end of the paragraph 1 of Section 6, we added the following sentence:

*We get 3256 samples in total, and include 10 words in a sample on average. Among them, 720 samples are for training, and others are used for testing.*

1. « we only utilize the data from the 24 subjects » : but there are only 24 subjects.

We apologize for the confusion, but we collected 7 subjects’ data as observing attacks, we clarify this at the beginning of Section 6 (i.e., at the end of paragraph 4 in Section 6) by changing “…*only utilize*…” to: “*To evaluate performance under Non-Observing insiders, we utilize the data from the 24 subjects and do not include data from subjects who act as observing attackers.”*

1. p. 12, lines 32-33 « and HIS samples are never part of the training sets ». I would add : (except for building the vocabulary).

Fixed.

1. [ROC curves] fig. 11 and 12 are not legible. You should enlarge the left-top corner.

We appreciate the suggestion. I have modified the figure by zooming in both figures into 0 to 0.2 on the x-axis and 0.8 to 1.0 on y-axis.

1. [Figure 8] Figure 8 should be revised : In the third block (style level feature), it is not so clear now (less clear than in the previous version) that f\_s is an index.  In addition, the sample level feature is the matrix, not the histogram vector and you cannot derive the co-occurrence matrix from the histogram vector. So that the arrow is confusing. I would also remove the « 2nd sample » since you do not associate it with a co-occurrence matrix.

We agree with the reviewer. In the revision, we removed the histogram to improve the clarity.

1. [Typos] page 8, line 12 : left end ?

several times : relay attack instead of replay attack

imposer : imposter

We have fixed all these typos and proofread the whole paper.