**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.  
  
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 Setion  
3.2, and not earlier in Section 2.2.

Thanks for your suggestion. We have moved the section 3.2 to the end of 2.2.

*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. However, the focus of this paper is to study the biometric built on handwriting motion, while several literatures can be utilized for the content matching. For instance, MoCRA can utilize online handwriting recognition that 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 [25], [31], [37], [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.*

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 is addressed differently.

*The first step is not the focus of this paper, as explained in Section 2.*

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 add the assumption of the system would not be attacked by relay attack because of the introduction of the challenge-response mechanism.   
*The first step is not the focus of this paper, as explained in Section 2. However, the integration of the first step is necessary as the handwriting style might be learned from fully skilled attackers. For example, an attacker could learn the style of some given contents from a secret video tapping, then relay the content. The challenge-response mechanism addresses this concern by asking the user to write the content the system given, i.e., random content shown from a screen. Based on the assumption that the relay attack would not be an issue for MoCRA.*

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.

We have tuned the SVM parameters. The penalty parameter was tested from a very small positive number to large ones, by zoom in spe the 5000 is chosen, as it does not affect our   
  
In Section 6.1, how many times did each participant write the 19 paragraphs? [once]  
Is this the '30 samples' on page 11? [the 19 paragraphs include the 30 training samples, and also include more samples from each subject used for testing] This is an important detail.  The explanation of how data samples were split, in what specific ratios, or if  
cross-fold validation [cross validation are not made, as we prefer the consider the practical system of using the first batch of enrolled data for training] was used needs to be made clear in 6.1.

We add specific number of samples in the paper:

*We get 3256 samples in total. Among them, 720 samples are for training, and others are used for testing.*  
  
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 modified the descriptions in 6.1.2.

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

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-occurence matrix)?

The worst cases mainly because some of its samples shows similarities styles to a few subjects, instead of having average co-occurrence matrix. Seems its writing style is not as unique or we may guess the subject does not accommodate to the writing input method (in-air) thus does not form its own writing style.  
  
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.

Thanks for your suggestion, we changed the “Random” attacker into “Non-Observing” attacker.

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.

To clarify it, we rewrite this part from the aspect of each victim subject, which corresponds one ROC curve in the result figure.

***Non-Observing Insiders.***

*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.*

***Non-Observing Impostors.***

*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:  
  
- 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.

Thanks for your pointing out.

*The future work of the content matching will also increase the accuracy of the system, as our evaluation results does not reject the attack samples that have the same hand writing style distribution as the legitimate one but formed from different content, which should be rejected as the earlier stage -- content matching.*

*Our scheme authenticates users based on `what they write' and `how they write'. The first step will be discussed in the future work and this paper focused on `how they write'.*  
- 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 add some in section 5.4 classification.

*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 representatives of the user. 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 representatives of the writing style. We will examine appropriate length in Section 6.*

- 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?  
You are right. *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:  
  
- Page 1: Unclear: 'where a user obtained the secret' - what secret?

We clarify it by adding *(i.e., the response)*

- 'HandWritings' -> 'handwriting samples' (p. 1) ok  
- 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 delete the part.  
- page 3: "utilize [a] more sophisticated features" - remove 'a'. ok

- 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 add several ones.

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

- 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 omit this and the results section shows the trend.  
  
- 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:  
[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 ?   
Thanks for your suggestion. We added explanations on the first and last word.

*Exceptions are the first and the last words. MoCRA sets hotkeys for starting a recording and stopping a recording. Before starting, finger(s) are stay motionless and start writing after the press of the start hotkey. MoCRA introduces extra data frames before the intent writing. We simply include these frames in the first word. Then 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.*

*The possible extra frames are assigned to the last word for simplicity.*

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.

Sorry for the misleading. The subject did draw the point before lower part.  
  
[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 clarify it with more detailed description: *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.*

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

Thanks for your pointing out.

*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.*

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

Sorry for the misleading. We also collected 7 subjects’ data as observing attack, we now address this at the beginning of Section 6.

We changed the “…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.*

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

 [ROC curves]  
fig. 11 and 12 are not legible. You should enlarge the left-top corner.  
We zoomed in both figures into 0 to 0.2 on the x-axis and 0.8 to 1.0 on y-axis.

[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.  
Thanks for your suggestion, we removed the histogram to make it more clear.  
  
[Typos]  
page 8, line 12  : left end ? we delete the end.  
several times : relay attack instead of replay attack ok  
imposer : imposter ok