Reviewer: 1  
  
Recommendation: Author Should Prepare A Major Revision For A Second Review  
  
Comments:  
[ Title and Abstract ]  
  
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."

Thanks. We have changed the title to “Challenge-Response Authentication using In-Air Handwriting Style Verification”  
  
[ Introduction ]  
  
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.

We have added explanations on a few terms, including ‘man-in-the-middle attacks’ and \*\*\*\*. In addition, we have added brief introduction to handwritings and explained why in-air writing is chosen. 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),…

[ Literature Review ]  
  
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?

Thanks for your suggestion, and we have added additional literatures on the topics of handwriting writer identification, which is close more close to our paper from the technique part. Handwriting verification, which based on content-dependent verification is also included.

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.

Thanks for your suggestion, we combined the background and literature part together.  
  
[ Figures and Tables ]  
  
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.

Printed text was added for easy read.  
  
Fig. 3 - if fewer examples can be provided, with clearer emphasis of key similarities and differences, it will be clearer for the reader.

We deleted one row out of five rows. We added more explanation on ‘p’ and ‘b’ part but leave ‘h’ and ‘n’ for readers’ reference.  
  
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.

Thank for your good suggestion, the major axis made the table clearer. And we also added printed text under each writing word.  
  
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.

We represented the features with symbols referred form low level, and eliminated the confusing illustrating numbers in histograms. We also grouped the different types of features by dashing boxing and left ‘vocabulary / codebook’ out.

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.

The default parameters were added to the caption of Table 2. To make the table clearer, we did not present all results while choosing the parameter. After we had chosen the default one, we varied one parameter and fixed others for comparison.  
  
  
[ Methodology ]  
  
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.

Thanks for this good question. The purpose of this paper is to discover the feasibility of using 3D writings and extract writing style from it for authentication. And the application scenario of this authentication is mainly for contactless environment, and also took privacy, security and public usage as concerns. We have considered about the comparison with 2D writings. However, we think the comparison is not reasonable as 1) the current application on 2D verification are mostly based on content-depended, 2) most current content-independent application included heavy training session for better verification rate which demands time and enrolling samples, thus not user-friendly for authentication, 3) statistically meaningful testing results should include large number of data. We have thousands in-air handwritings samples, which is not the same content from any existing 2D signature varication database, or 2D online text recognition database. In sum, we think the comparison is orange to apple, which cannot provide significant meaning in our research.

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.

We have added a brief introduction and also added a new sub-session in discussion part for the usability concerns while introducing Challenge-response (by content matching).

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.

Thanks for your suggestion. We think the “stroke segments” and “style-level” are clearer than component and sample-level. We have changed them in the paper.

We clarified the “sample” in introduction part for easy reading.  
  
[ Evaluation ]  
  
This is one of the weakest parts of the paper.  
  
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.

We explained why we did not compare with 2D writings in one of the above question. Copied from above: The purpose of this paper is to discover the feasibility of using 3D writings and extract writing style from it for authentication. And the application scenario of this authentication is mainly for contactless environment, and also took privacy, security and public usage as concerns. We have considered about the comparison with 2D writings. However, we think the comparison is not reasonable as 1) the current application on 2D verification are mostly based on content-depended, 2) most current content-independent application included heavy training session for better verification rate which demands time and enrolling samples, thus not user-friendly for authentication, 3) statistically meaningful testing results should include large number of data. We have thousands in-air handwritings samples, which is not the same content from any existing 2D signature varication database, or 2D online text recognition database. In sum, we think the experimental comparison is orange to apple, which cannot provide significant meaning in our research.

We added comparison on existing methods and on other verification system in related work.

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.

Thanks for your suggestion. We have added default parameter to the caption of the table to make the table easier to be understood. We added more experiment result with 16 as stroke segment length to the table to make the parameter chosen more reasonable.  
  
'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.

Maybe we have some misleading here. The “random” does not mean an attacker randomly comes from somewhere, but means the insider does not intend to learn or learned useful information from the victim.

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.  
Thanks for your suggestion, we re-ran the experiments as you suggested and have updated the results and the result does not affect the final conclusion of the paper.

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.

For this experiment. We first let one attacker attack more victims. However, after he learnt from four victims, he started feeling confused about the style he was learning and wrote some style. To keep the attacker energetic while learning, we decided to attack 4 victims. The 4 victims were randomly chosen. We could only either attack more victims by more but different attackers, which would lose the comparison base of attack data. Thus without loss of generosity, we decided to attack few victims from the same attackers. Based on the results, we can see one we could conclude that the “observing” may not help the attacker, or the learning may have helped a little.

The defined robot arm behavior seems arbitrary and fragile. Because co-occurences 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.'

Thanks for your suggestion. We also think the robot arm attacker is not reasonable here and our main purpose is to prevent human attackers. So we just deleted this part.   
  
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.

I am not sure if I understood your question well. The larger vocabulary size describes more variability. The text-independent method still depends on what the primitives extracted from the training text, but not depend on what the content of the testing text, as we assume the primitives are ok to represent the writing elements of new 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)?

In this paper we are focus on how to apply the in-air handwriting style for authentication purpose. For better security concern, we introduced the Challenge-response to the system. The content matching which is a similar work to handwriting recognition. We added this part in the future work as an extension of this project to have a comprehensive results of the whole system, which will be compared with the results in this paper.   
  
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.

Thanks for your suggestion. Extracting fragments of writing could be very interesting. We added to the future work part.  
  
  
3) Fig. 3  
« component index » : I assume you mean « primitive index» .

Thanks, we 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.

Thanks for your suggestion of using extra database for codebook training. For practical concerning, the ideal method is to use a separate database. However, our in-air writing database, which was collected by recruiting subjects, is not large enough to be separated for two databases for both generating effective codebook for lots of writing styles, and maintaining a statistically significant results on the authentication results. However, this is not as ideal as using distinct database to adapted for new users, but it is practical for a relatively small user group. To be more clear, we will address this in the discussion part.

5) The test impostors should not be known from the system (f.i. 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.

We defined the 'random insiders' for the purpose of maybe the attacker is from the group who has the key (the writing style already trained in the model). However, he may want to be authenticated as another user from the group, but this is not allowed. This is a possible case, so we also studied its security performance.

Sorry for the confusion, the codebook was built during the training session, using part of the training data, thus the actual impostor 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 allographic features, IEEE PAMI, 2007.). A distinct database could be used.

To avoid the confusion of training and testing data, we added one sentence for a briefly introduce before each experiment setting.

As method in previous question, just copied: *Thanks for your suggestion of using extra database for codebook training. For practical concerning, the ideal method is to use a separate database. However, our in-air writing database, which was collected by recruiting subjects, is not large enough to be separated for two databases for both generating effective codebook for lots of writing styles, and maintaining a statistically significant results on the authentication results. However, this is not as ideal as using distinct database to adapted for new users, but it is practical for a relatively small user group. To be more clear, we will address this in the discussion part.*

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?

The starting part of writing pieces has mixed pattern as normal strokes, and as the users mostly were recording one sentence at a time, the starting part is not as annoy as the transition part, between every two words.

Fig. 6 is for illustration purpose, so we only showed the transition segmentation part, and cut the first part for not disturbing.

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...

To avoid the confusion of training and testing data, we added one sentence for a briefly introduce before each experiment setting. And the length of the training/test samples were fixed after parameter chosen.  
  
9) what is the use of the overlap?

The overlap of different stroke segment (component) is for capturing more detail variations on handwritings. 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 ».

We have added total writing time for each column, the *L\_s* is number of frames and we ignored the additional Math notation here, as the texts are more directly self-explanation.  
  
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...).  
Sorry for the confusion, we were not using the directly binary output as the classification results, but apply a threshold according to the prediction probability generated by the SVM and then a separate threshold would be applied to output the final binary results. The ROC curve is a good baseline for threshold chosen. The EER is a standard measurement for security performance which based on threshold is chosen based on equal error rate during training session.  
  
12) ROC curve: make a figure with an y axis greater than 1 in order to better visualize the curve.

We showed ROC curves above the axis boxing therefore the curves were hidden by the boxing line. The (0% FPR, 100% TPR) is necessary for the readers to know how close the curves to the idea position.   
  
13) Please provide details about the « noisy » environment mentioned in line 45, page 12, and how you address it.  
sorry for the misleading, we have corrected it. We were talking about acoustic noises.

14) There is a confusion between « authentication » and « identification » : see f.i. Fig. 4 b)  
the output of SVM classification is not user identity but a binary reply (user or impostor)

Thanks for your pointing out. We have modified it to <authentication> and also a binary reply.  
  
15) Please provide illustrated samples of robot attack.

We have deleted this part, because the paper focuses on human attacks, and the robot attack is not very support to our work.  
  
16) theta\_yz appears in Fig. 9 but not in the list of features.

We have modified it. Thanks for pointing it out.  
  
 17) English has to be checked. Many typos. "authetication", 'have attract' etc...

Thanks, we went through the paper a few times again and tried to eliminate all typos.