MSDS: A Large-Scale Chinese Signature and Token Digit String Dataset for Handwriting Verification

Peirong Zhang
South China University of Technology
eeprzhang@mail.scut.edu.cn

Jiajia Jiang
South China University of Technology
eejiajia_jiang@mail.scut.edu.cn

Yuliang Liu
Huazhong University of Science and Technology
ylliu@hust.edu.cn

Lianwen Jin*
South China University of Technology
eelwjin@scut.edu.cn

1 Appendix

1.1 Motivation

For what purpose was the dataset created?
Although handwriting verification has been rapidly developed, the verification performances in specific scenarios are still unsatisfactory owing to the small sizes of datasets and the limited biometric mediums. For online Chinese signature verification, the largest existing dataset [2] has only 50 users. However, DeepSignDB [3], the largest online signature dataset in the Western language, possesses 1526 users and almost 70,000 samples. In addition, existing studies set foot on exploiting signatures for verification, but somehow ignored exploring new and more effective biometric mediums. To this end, we propose the large-scale Multimodal Signature and Digit String (MSDS) dataset, which contains two subsets: MSDS-ChS for Chinese signature and MSDS-TDS for Token Digit String (TDS). MSDS-ChS aims to boost Chinese signature verification and MSDS-TDS attempts to explore the effectiveness of Token Digit Strings, i.e. the actual phone numbers of users, in identity verification. MSDS-ChS is the largest publicly available Chinese signature dataset, which is at least eight times larger than the previous online ones [2, 1, 4]. MSDS-TDS is the first dataset that covers handwritten TDS for identity verification, which facilitates related research and brings long-term implications.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The MSDS dataset is created by the Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology.

1.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
The MSDS dataset possesses two modalities: the online time series modality and the offline image modality. The time sequences record the positional and temporal information produced in users' handwriting processes, including x, y coordinates, pressure, and time stamps, and are saved in separate text files. The static images are rendered from the time series and saved in the Portable Network Graphics (PNG) format. Therefore, the time series and images are uniquely corresponding.

How many instances are there in total (of each type, if appropriate)?
Our MSDS dataset consists of two subsets: MSDS-ChS and MSDS-TDS, which are contributed by the same 402 users, with 20 genuine samples and 20 skilled forgeries per user per subset. The data

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

is provided in two modalities. For each modality, there are $402 \times (20 + 20) = 16080$ samples per subset. Hence, there are in total $16080 \times 2 \times 2 = 64320$ instances in the MSDS dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
The MSDS dataset contains all possible instances.

What data does each instance consist of?
The MSDS dataset contains unprocessed time series and images.

Is there a label or target associated with each instance?
The only label for each instance is the user it belongs to.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
There is no relationship between individual instances.

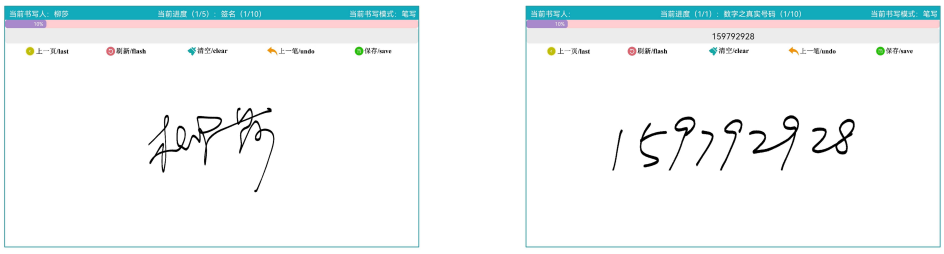
Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
The MSDS dataset is self-contained.

1.3 Collection Process

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How was the data associated with each instance acquired?
The data of MSDS was acquired with two types of Android tablets: HUAWEI MatePad BAH3-W59 and LENOVO TB-J706F, with three of each. Table 1 lists the specifications of these two types of devices. We specifically developed an Android app and the user interface is shown in Figure 1. Users directly performed handwriting on the tablets using specific styluses and the produced information was automatically recorded by the app.

Table 1: Specifications of the two devices.

Device	Input	Screen (Diagonal)	Resolution	Actual Writing Area
HUAWEI MatePad BAH3-W59	Stylus (Specific)	10.4 inches	2000×1000	2000×900
LENOVO TB-J706F	Stylus (Specific)	11.5 inches	2560×1600	2560×1180



(a) Signature (b) Token Digit String

Figure 1: The user interface of the data acquisition app.

1.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done?
After the data collection finished, we manually inspected the data to perform cleaning. If the content of data is wrong or out of format, the user was required to rewrite it. For instance, if the content of a skilled forgery does not match the corresponding genuine sample, we asked the user to rewrite the skilled forgery with the correct content.

1.5 Uses

Has the dataset been used for any tasks already?
This dataset has not been used in any previous work.

What (other) tasks could the dataset be used for?
The MSDS dataset is published for handwriting identity verification. Specifically, the MSDS-ChS subset could be exploited in online/offline Chinese signature verification, and the MSDS-TDS subset could be used in online/offline identity verification with Token Digit Strings. In addition, the MSDS dataset could be exploited in writer identification.

1.6 Ethics

Does the dataset contain personally identifiable information?
Yes, the signatures and Token Digit Strings in our dataset can be linked to the users.

Did data contributors provide their consent on the collection and use of the data? If they did, how?
Yes, all contributors gave their consent to the collection of the handwriting data. Before collecting the contributors' handwriting data, we signed a copyright agreement with each contributor, in which they agreed to grant us the license to use their handwriting, to use it for publication, and to use it for non-commercial academic research purposes. The copyright agreement in both Chinese and English is shown in Figure 2. Therefore, the contributors are well aware of how their data will be used.

Were data contributors compensated?
Yes, according to the payment agreement, all contributors were paid according to the amount of handwriting that they contributed.

Are there any potential negative societal impacts of this dataset? If yes, do you have any measures to prevent that?
To protect contributors' privacy, we carefully design the conditions of use of our dataset, which mainly include the following three rules. First, the MSDS dataset can only be used for non-commercial research purposes and we will verify the researcher's intentions regarding the dataset through the application process described on our official GitHub page. Second, we respectively shuffle the user order of signatures and TDS, resulting in an unassociated order between signatures and TDS. Third, we will issue a legal agreement that applicants are required to sign to prevent the illegal use of the data. If any applicant does not agree to these conditions of use, we will not provide he/she with the decompression code. Furthermore, whoever uses our dataset must comply with all use conditions; otherwise, we will revoke the authorization.

1.7 Distribution

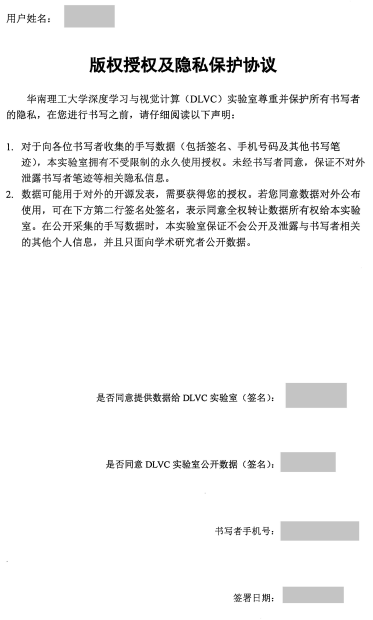
How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
The MSDS dataset is publicly available at https://github.com/HCIILAB/MSDS.

1.8 License

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
All authors bear all responsibility for the MSDS dataset in case of violation of rights, etc. The MSDS dataset should be used under Creative Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License for non-commercial research purposes.

1.9 Maintenance

Our dataset can be accessed through the aforementioned link. We will occasionally perform maintenance, such as providing data corrections and solving issues raised by developers.



用户姓名：

版权授权及隐私保护协议

华南理工大学深度学习与视觉计算（DLVC）实验室尊重并保护所有书写者的隐私，在您进行书写之前，请仔细阅读以下声明：

1. 对于向各位书写者收集的手写数据（包括签名、手机号码及其他书写笔迹），本实验室拥有不受限制的永久使用授权。未经书写者同意，保证不对外泄露书写者笔迹等相关隐私信息。
2. 数据可能用于对外的开源发表，需要获得您的授权，若您同意数据对外公布使用，可在下方第二行签名处签名，表示同意全权转让数据所有权给本实验室。在公开采集的手写数据时，本实验室保证不会公开及泄露与书写者相关的其他个人信息，并且只面向学术研究者公开数据。

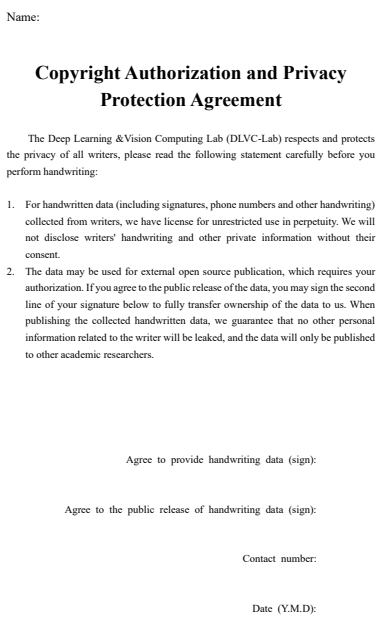
是否同意提供数据给 DLVC 实验室（签名）：

是否同意 DLVC 实验室公开数据（签名）：

书写者手机号：

签署日期：

(a) Copyright agreement in Chinese.



Name:

Copyright Authorization and Privacy Protection Agreement

The Deep Learning &Vision Computing Lab (DLVC-Lab) respects and protects the privacy of all writers, please read the following statement carefully before you perform handwriting:

1. For handwritten data (including signatures, phone numbers and other handwriting) collected from writers, we have license for unrestricted use in perpetuity. We will not disclose writers' handwriting and other private information without their consent.
2. The data may be used for external open source publication, which requires your authorization. If you agree to the public release of the data, you may sign the second line of your signature below to fully transfer ownership of the data to us. When publishing the collected handwritten data, we guarantee that no other personal information related to the writer will be leaked, and the data will only be published to other academic researchers.

Agree to provide handwriting data (sign):

Agree to the public release of handwriting data (sign):

Contact number:

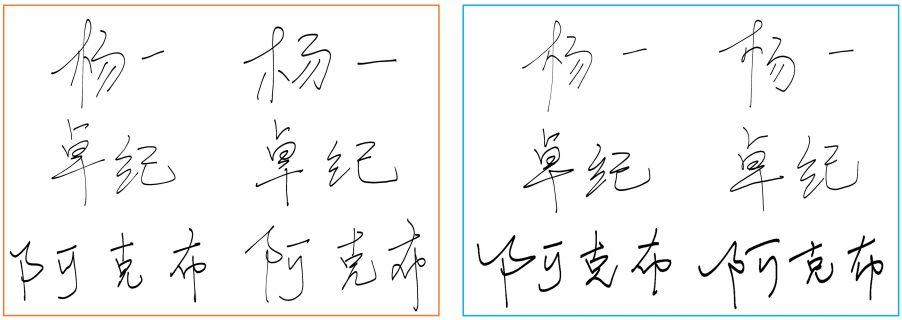
Date (Y.M.D):

(b) Copyright agreement in English.

Figure 2: The copyright agreement that signed with each writer before data collection.

1.10 Examples

MSDS contains two subsets, MSDS-ChS for handwritten Chinese signatures and MSDS-TDS for handwritten Token Digit Strings (TDS). We present several examples to view the quality of our datasets in Figure 3 and Figure 4.



genuine skilled forgeries

(a) Session 1

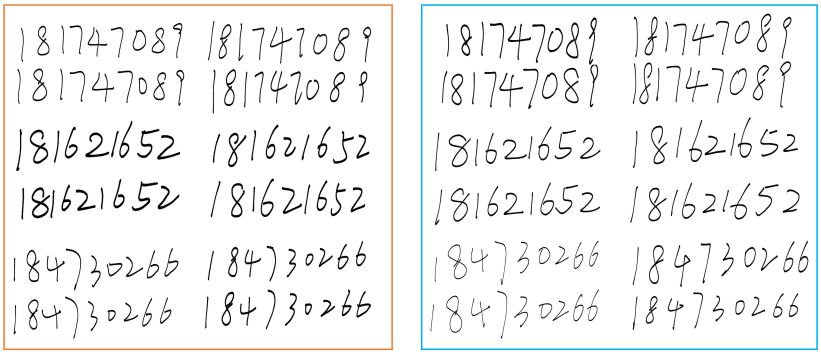
genuine skilled forgeries

(b) Session 2

Figure 3: Signature samples of two sessions. The left ones are genuine samples, while the right ones are skilled forgeries.

2 Case Study: Time Functions

In experiments, we extract 12 time functions for online time series using the x, y coordinates, and pressure as features and serve as input to the online handwriting verification systems. Table 2 respectively lists the details of 12 time functions.



genuine skilled forgeries (a) Session 1

genuine skilled forgeries (b) Session 2

Figure 4: TDS samples of two sessions. The left ones are genuine samples, while the right ones are skiiled forgeries.

Table 2: Time functions extracted from the dynamic time series.

No.	Features
1	First-order derivative of Coordinate $\dot{x}$
2	First-order derivative of Coordinate $\dot{y}$
3	Velocity magnitude $v = \sqrt{\dot{x}^2 + \dot{y}^2}$
4	Path-tangent angle $\theta = \arctan \frac{\dot{y}}{\dot{x}}$
5	Cosine of the path-tangent angle $\cos\theta$
6	Sine of the path-tangent angle $\sin\theta$
7	First-order derivative of velocity magnitude $\dot{v}$
8	First-order derivative of path-tangent angle $\dot{\theta}$
9	Log curvature radius $\rho = \log \frac{v}{\dot{\theta}}$
10	Velocity variation magnitude $\triangle v = v \cdot \dot{\theta}$
11	Acceleration $a = \sqrt{\dot{v}^2 + \triangle v^2}$
12	Pressure p

References

[1] M. Liwicki, M. I. Malik, C. E. v. d. Heuvel, X. Chen, C. Berger, R. Stoel, M. Blumenstein, and B. Found. Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011). In *11th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1480–1484, 2011.

[2] X. Lu, Y. Fang, W. Kang, Z. Wang, and D. D. Feng. SCUT-MMSIG: A Multimodal Online Signature Database. In *Chinese Conference on Biometric Recognition*, pages 729–738. Springer, 2017.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. DeepSign: Deep On-Line Signature Verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2): 229–239, 2021.

[4] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll. SVC2004: First International Signature Verification Competition. In *International Conference on Biometric Authentication*, pages 16–22. Springer, 2004.