A Deep Learning Approach to Writer Identification Using Inertial Sensor Data of Air-Handwriting

Yanfang DING†, Nonmember  and  Yang XUE†a, Member

SUMMARY To the best of our knowledge, there are a few researches on air-handwriting character-level writer identification only employing acceleration and angular velocity data. In this paper, we propose a deep learning approach to writer identification only using inertial sensor data of air-handwriting. In particular, we separate different representations of degree of freedom (DoF) of air-handwriting to extract local dependency and inter-relationship in different CNNs separately. Experiments on a public dataset achieve an average good performance without any extra hand-designed feature extractions.

key words: writer identification, air-handwriting, acceleration, angular velocity, convolution neural network

1. Introduction

Owing to the increasing used mobile platforms (e.g., wearable devices or Virtual Reality) that lack of keyboard or camera to type passwords or capture visible data for user authentication, air-handwriting character/word-level writer identification systems have gained increasing attention [1]–[7]. Most of existing researches employ pen tip or finger tip movement (position data) for writer identification [1]–[6]. However, in order to obtain position data, more peripherals (e.g., cameras and infrared devices [8], [9]) are usually required. Such peripherals may be limited by light, background, etc. and the writer would have to hold his hand over the peripherals in the coverage area. In contrast to these peripherals, inertial sensor is not affected by environmental factors, and does not require the writer to write in a prescribed space, enabling true 3D air free writing. Besides, inertial sensor is less expensive. Therefore, the inertial sensor based model has a wide application scenario.

Most previous methods rely heavily on sophisticated hand-designed features for writer identification [2], [4]–[7]. It may introduce more parameters, which makes model complicated and not conducive to computing resource-limited devices. In addition, only a few publications [3], [7] used inertial sensor data for writer identification based on air-handwriting. [3] used position data and velocity, acceleration calculating from position data, while [7] used acceleration, angular velocity and Euler angles obtained by their glove. In other words, they were not just employing acceleration and angular velocity. One reason might be that acceleration and angular velocity contain more hidden information, which increases the difficulty of designing features manually.

In recent years, convolutional neural network (CNN) have improved great recognition performance for sequence data, especially inertial sensor signals [10]–[12]. CNN can well capture the two critical information containing in sequence data: local dependency and interrelationship. In addition, in contrast to recurrent neural network (RNN) [1] with memory limitations and weak parallelization within training, CNN has the advantages of high computing efficiency with highly optimized matrix multiplication code.

Our hand is very dexterous with more than 20 DoFs. Acceleration and angular velocity involve a total of six DoFs, and respectively represent two different representations of DoF: translational and rotational. If we simply concatenated all DoFs and sent them to CNN, the CNN would capture incorrect interrelationship because the different DoFs would affect each other. In the field of action recognition [13], [14], different type data cues were processed separately and they achieved good performance in their tasks.

Considering all factors mentioned above, we propose a deep learning structure, in which different representations of DoF of air-handwriting separately capture critical information. Specifically, the input data includes six DoFs of air-handwriting, three translational DoFs (acceleration) and three rotational DoFs (angular velocity). They respectively carry information about two different types of movement from different representation subspaces. Our CNN is present to reduce the influence among them. In addition, our CNN is beneficial to the joint of more different representations of DoF (e.g., orientation), and makes them not affect each other. More notably, our approach directly operates on raw data without any extra hand-designed feature extractions and it benefits computing resource-limited devices.

2. Our Proposed Approach

In the field of human activity recognition, Y. Chen proposed a structure of deep CNN based on acceleration, in which the size of the convolution kernel was modified to adapt the rectangular matrix data from acceleration sensor, yielding good performance without any feature extraction [10]. Our network inherits the advantages of Y. Chen’s [10] and is improved.

The pipeline of our approach is shown in Fig. 1. Firstly,
Fig. 1  The pipeline of our approach to writer identification based on inertial sensor data. Multi representations matrix $M$ is sliced vertically into $H$ submatrices $M_h$. For capture corresponding interrelationship, the $M_h$ are sent to different CNN structures, whose output $u_h$ are concatenated horizontally as $u$. Finally, the identification results are output after a full-connection layer and a softmax layer.

Fig. 2  Six degrees of freedom. Left: translational movement in three perpendicular axes (surge, sway, heave); center: rotational movement about perpendicular axes (roll, pitch, yaw); right: six degrees of freedom.

2.1 Data-Preprocessing

Given the lack of haptic or visual feedback in air-handwriting, it is difficult for the writer to write at a constant speed. Therefore, at a fixed sampling rate, varying length $N$ of a sample will be get. However, the CNN requires the input data with the same size. To solve this problem, we resample the original data by one-dimensional linear spline interpolation, which makes each sample has the same new length of $L$. We set a fixed $L = 256$.

2.2 Six Degrees of Freedom (6DoFs)

The DoF of a mechanical system is the number of independent parameters that define its configuration, which is an important term for analyzing and measuring properties of a mechanical system. As a mechanical system, our hand has more than 20 DoFs, but due to the interdependences between fingers, studies have shown that at least six DoFs are needed [15]. Half of these represent translational movement (surge, sway, heave), while the other half represent rotational movement (roll, pitch, yaw). The trajectories of the 6 DoFs are shown in Fig. 2.

The acceleration data from an accelerometer can represent the translational movement, while the angular velocity from a gyroscope sensor can represent the rotational movement. Therefore, acceleration and angular velocity are independent parameters from different representation subspaces. If we simply concatenated them and sent them to CNN, the CNN would capture incorrect interrelationship, since the different information from different representation subspaces would affect each other. Therefore, our approach separates them to reduce the impact of convolutional and pooling operations.

2.3 The General Structure of Our Approach

As shown in Fig. 1, the preprocessed input data is denoted by a matrix $M \in \mathbb{R}^{3H \times L}$, where $H \in \mathbb{N}^+$ is the number of representation subspaces of air-handwriting and $L \in \mathbb{N}^+$ is the length of preprocessed data. The multi representations matrix $M$ is sliced vertically into $H$ submatrices as $M_h \in \mathbb{R}^{3 \times L}$, where $1 \leq h \leq H, h \in \mathbb{N}^+$. Then each $M_h$ is sent to different CNN structures to extract corresponding vital information. The output of each CNN as $u_h \in \mathbb{R}^{1 \times 100}$ are concatenated horizontally as $u \in \mathbb{R}^{h \times 1 \times 100}$. After sending the concatenated vector $u$ to a full-connection layer and a softmax layer, the final identification result is output. Notes that the weights of different CNN structures are not shared and their structures could be different depending on corresponding representations of air-handwriting. Our approach is beneficial to the joint input of $H$ different inertial sensors data in different representation subspaces, and makes them not affect each
Fig. 3  The structure of our CNN. There are 3 convolutional layers followed by Max-pooling (MP) layer. Sigmoid activation function is used in the first convolutional and Max-pooling layer while ReLU is used in the others.

The structure of each CNN is shown in Fig. 3. There are 3 convolutional layers followed by Max-pooling (MP) layer. The size of the convolutional filters is $2 \times 15$ in the first layer and $1 \times 14$ in the subsequent layers. The convolution stride is set to 1. Max-pooling is carried out over a $1 \times 2$ window with the stride size of $1 \times 2$. Sigmoid activation function is used in the first convolutional and Max-pooling layer while ReLU is used in the others. Moreover, dropout (set to 0.5) is used in the last Max-pooling layer and the full-connection (FC) layer with size of 100. Finally, the output of the CNN denotes as $u_h \in \mathbb{R}^{1 \times 100}$.

3. Experiments

3.1 Dataset

Our approach are evaluated on the public dataset called 6DMG [16]. The dataset consists of air-handwriting digits and lowercase letters from 6 writers, uppercase letters from 25 writers, with a total number of 8571. The acceleration and angular velocity representations are employed in this work. The data was collected by a hybrid framework in [8], where the acceleration and angular velocity were recorded only by the embedded inertial sensor of the writing device in the hand of writers, and the sampling rate is 60Hz. Figure 4 shows the waveforms of inertial sensor data of digit ‘0’, lowercase letter ‘a’, and uppercase letter ‘A’ written by three different writers (J1, M1, Y2). The blue waveforms represent accelerations, and the yellow ones represent angular velocities.

3.2 Experimental Configurations

We set $H = 2$ and $M_1 \in \mathbb{R}^{3 \times L}$ represents acceleration data and $M_2 \in \mathbb{R}^{3 \times L}$ represents angular velocity data. The structure of CNN-1 and CNN-2 is the same\(^\dagger\), but we must emphasize that different CNN structure can be used for different representation submatrices. We perform our experiments on a 64bit PC with a single GTX Titan X GPU. We set up four groups of experiments based on text of air-handwriting: digits (Digits), lowercase letters (Lower), uppercase letters (Upper) and all characters (All). For all groups of experiments, the number of training epoch is set to 1000 and the training mini-batch size is fixed to 128. Because the size of the dataset is small, we apply 5-fold cross-validation method.

\(^\dagger\)See Sect. 2.3, but their weights are not shared.

3.3 Results

**Identification performance.** As shown in Table 1, our approach achieves an average accuracy of 98.59%, 99.84%, 99.22% and 99.38%, respectively on the experiment of Digits, Lower, Upper, and All. In addition, we evaluate average-based method to combine two branches like [13], [14], and the results show that training a joint FC layer on top of FC layers of two branches (referred to Sect. 2.3) perform better without over-fitting in our case. We argue that the reason is that the weights of combining two branches are not necessarily equal and the data-driven FC layer may better balance the relationship between the two branches. Notes that we do not apply SVM-based method to combine two branches like [13], because SVM-based method would introduce more parameters, and our FC-based method is good enough in our case.

For thorough comparison, the test error rates in training process are shown in Fig. 5, which help us to analyze the significance and verify the effectiveness of our approach. The identification results indicate that our approach with $H = 2$ can reduce the impact between two different representations from different representation subspaces owing to convolutional and pooling operations. We believe that the performance of our approach will be better than those without separating multi representation matrix when $H \geq 3$, which will be further studied in our future.

**Number of model parameters.** We investigate the relationship between the number of model parameters and $H$, as shown in Table 2\(^\ddagger\). When $H = 1$, the number of parameters

\(^\ddagger\)Assume that the size of the FC layer followed by softmax layer is $1 \times 25$. 

in CNN without representation matrix slice is equal to that of our approach. When $H \geq 2$, the number of parameters of our approach is $58082 \times (H - 1)$ less than that of that without separating multi representation matrix. This effect will become more and more significant with the increasing of $H$. A model with less parameters is conducive to computing resource-limited devices, which shows the practicality of our proposed approach.

**Training cost.** We evaluate the computing efficiency within training of our proposed approach. We perform our experiments on a 64bit PC with a single GTX Titan X GPU. All codes are written in Python and the deep learning framework is Caffe. As shown in Table 3, training 1000 epochs just costs about 19 minutes at most, and the time spent depends on the number of training samples on different experiments, which indicates that our proposed approach has high compute efficiency within training.

### 4. Conclusion

Most previous methods employ position data and rely heavily on hand-designed features for air-handwriting characters-level writer identification. In this paper, we only employ acceleration and angular velocity data without any artificial feature extraction to writer identification, and we propose an extensible CNN, in which different representations of DoF of air-handwriting separately capture critical information. Comparing with the Y. Chen’s Net without separation, we verify that our proposed approach can reduce the impact between different representations from different representation subspaces. Besides, we investigate and show that our approach has less parameters and high computing efficiency to practical applications with computing resource-limited devices.

Although we only employ two representation of DoF of air-handwriting for writer identification, we believe that our approach can perform well with the increasing of inertial sensors. In addition, the structure of different sub-CNN requires further investigations.

### References

1. X.-Y. Zhang, G.-S. Xie, C.-L. Liu, and Y. Bengio, “End-to-end online writer identification with recurrent neural network,” IEEE Trans. Human-Mach. Syst., vol.47, no.2, pp.285–292, 2017.
2. M.N. Abdi and M. Khemakhem, “A model-based approach to offline text-independent arabic writer identification and verification,” Pattern Recognit., vol.48, no.5, pp.1890–1903, 2015.
3. D. Lu, D. Huang, and A. Rai, “FMHash: Deep hashing of in-air-handwriting for user identification,” arXiv preprint arXiv:1806.03574, 2018.
4. J. Tian, Y. Cao, W. Xu, and S. Wang, “Challenge-response authentication using in-air-handwriting style verification,” IEEE Trans. Dependable and Secure Computing, no.1, p.1, 2017.
5. J. Malik, A. Elhayek, S. Ahmed, F. Shafait, M.I. Malik, and D. Stricker, “3DAirSig: A framework for enabling in-air signatures using a multi-modal depth sensor,” Sensors, vol.18, no.11, 3872, 2018.
6. D. Lu, D. Huang, Y. Deng, and A. Alshamrani, “Multifactor user authentication with in-air-handwriting and hand geometry,” 2018 Inter-

### Table 1

The average accuracy (%) of experiments with 5-fold cross-validation. Acceleration and angular velocity are abbreviated as Accel., Angu.

| Network       | Fusion method | Modality | Four groups of experiment based on text of air-handwriting |
|---------------|---------------|----------|------------------------------------------------------------|
|               |               |          | Digits (6Writers) | Lower (6Writers) | Upper (25Writers) | All (25Writers) |
| One-branch Net| —             | √        | 600 Samples      | 1470 Samples    | 6501 Samples      | 8571 Samples    |
| One-branch Net| —             | √        | 97.34 ± 1.36     | 99.06 ± 0.58    | 96.09 ± 0.70      | 95.62 ± 0.63    |
| One-branch Net| —             | √        | 99.06 ± 0.31     | 99.43 ± 0.85    | 99.43 ± 1.81      |                 |
| Two-branch Net| average       | √        | 98.59 ± 0.31     | 99.94 ± 0.31    | 99.94 ± 0.31      | 99.94 ± 0.31    |
| Two-branch Net| FC            |          | 99.99 ± 0.91     | 99.99 ± 0.91    | 99.99 ± 0.91      | 99.99 ± 0.91    |

### Table 2

Comparison of the number of model parameters.

| model                                | Number of Params. |
|--------------------------------------|-------------------|
| Y. Chen’s Net without matrix slice   | 181’746 × $H$ - 58082 |
| Our approach with matrix slice       | 123’664 × $H$     |

### Table 3

Average training time (ms) per epoch using our approach.

|               | Digits. | Lower. | Upper. | All.  |
|---------------|---------|--------|--------|-------|
| ms per epoch  |         |        |        |       |
|              | 90.08 ± 0.60 | 216.09 ± 1.21 | 877.20 ± 2.74 | 1149.09 ± 6.55 |
D. Lu, K. Xu, and D. Huang, “A data driven in-air-handwriting biometric authentication system,” 2017 IEEE International Joint Conference on Biometrics (IJCB), pp.531–537, IEEE, 2017.

M. Chen, G. AlRegib, and B.-H. Juang, “6DMG: A new 6D motion gesture database,” Proc. 3rd Multimedia Systems Conference, pp.83–88, ACM, 2012.

“Leap Motion controller,” www.leapmotion.com

Y. Chen and Y. Xue, “A deep learning approach to human activity recognition based on single accelerometer,” 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.1488–1492, IEEE, 2015.

A. Ignatov, “Real-time human activity recognition from accelerometer data using convolutional neural networks,” Applied Soft Computing, vol.62, pp.915–922, 2018.

Z. Zhuang and Y. Xue, “TS-ICNN: Time sequence-based interval convolutional neural networks for human action detection and recognition,” IEICE Trans. Inf. & Syst., vol.E101-D, no.10, pp.2534–2538, Oct. 2018.

K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in Neural Information Processing Systems, pp.568–576, 2014.

Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, and J. Yuan, “Multi-stream CNN: Learning representations based on human-related regions for action recognition,” Pattern Recognit., vol.79, pp.32–43, 2018.

A. Erol, G. Bebis, M. Nicolescu, R.D. Boyle, and X. Twombly, “Vision-based hand pose estimation: A review,” Computer Vision and Image Understanding, vol.108, no.1-2, pp.52–73, 2007.

M. Chen, G. AlRegib, and B.-H. Juang, “Air-writing recognition—part I: Modeling and recognition of characters, words, and connecting motions,” IEEE Trans. Human-Mach. Syst., vol.46, no.3, pp.403–413, 2016.