

# Offline Text-Independent Arabic and Chinese Writer Identification Using a Multi-Segmentation Codebook-Based Strategy

Mohamed Nidhal Abdi<sup>1,2</sup> <sup>a</sup> and Maher Khemakhem<sup>3</sup> <sup>b</sup>

<sup>1</sup>*Mir@cl Laboratory, FSEG, University of Sfax, Tunisia*

<sup>2</sup>*Institut Supérieur des Mathématiques Appliquées et de l'Informatique, ISMAI, University of Kairouan, Tunisia*

<sup>3</sup>*Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*

**Keywords:** Writer Identification, Grapheme Codebook, Segmentation, K-Medoid Clustering, Feature Combination.

**Abstract:** Many approaches rely on segmentation for offline text-independent writer identification. Segmentation schemes based on contours, junctions and projections are widely used and are very effective with Latin alphabet handwriting. However, these schemes seem to be less consistent in capturing writer individuality with Arabic and Chinese. As writing systems, the latter languages are morphologically different and are considered more complex than Latin alphabet languages. In this paper, four different segmentation techniques are tested for the identification of Arabic and Chinese writers. Then, these techniques are combined to increase the accuracy of identification. Experiments were realized on handwriting samples by 300 writers from Arabic IFN/ENIT dataset and 300 writers from Chinese HIT-MW dataset. An additional 300 writers from English/German CVL dataset were used as a control group. Taken separately, these segmentation techniques that gave good results with CVL (Top1% = 99.00%) were not as conclusive with IFN/ENIT and HIT-MW. Nevertheless, the use of different types of segmentation in combination proved to be highly efficient for Arabic and Chinese with Top1% = 96.33% and Top1% = 91.33%, respectively.


## 1 INTRODUCTION


Arabic and Chinese are considered complex scripts by learners and scholars alike (Guellil et al., 2021; Tahsildar, 2019). Although its writing system is 14 centuries old, the Arabic language changed little over the years and is still readable in its ancient form by contemporary readers. In both its handwritten and printed forms, it is a right-to-left semi-cursive script that exhibits unique morphological features. Writers are allowed to merge certain letters according to their handwriting style, simplify others and arbitrarily insert elongations between connected letters. A letter can have up to four different forms depending on its position in the word (Amara and Bouslama, 2003). Moreover, Arabic calligraphy is well known for being particularly permissive with letter shapes (Alshahrani, 2008).

On the other hand, Chinese is a 30-centuries-old logographic writing system where visually complex handwritten characters are composed using di-

rectional strokes (Wang et al., 1999). Most words consist of two or more characters from approximately 50000 available ones, hence the lack of the grapheme-phoneme correspondence characterizing alphabet-based languages (Taylor and Taylor, 2014). Chinese characters are in majority (80%) phonetic-logographic characters, formed of combinations and recombination of *radical* components for meaning, and *phonetic* components for pronunciation. The remaining character categories are pictographs, ideographs, denotation of events, figurative extension of meaning and phonetic loans (Chen, 1996).

Up until recently, the automatic identification of Arabic and Chinese writers was not addressed as extensively as with the English language. In the context of offline text-independent identification, this paper aims at assessing the extent to which segmentation schemes that proved efficient for Latin alphabet languages work for Arabic and Chinese. Thus, four different segmentation techniques are studied. Experiments are conducted with Arabic IFN/ENIT, Chinese HIT-MW and compared to the English/German CVL dataset results.

<sup>a</sup>  <https://orcid.org/0009-0002-8783-1642>

<sup>b</sup>  <https://orcid.org/0000-0002-1287-1634>

In the remainder of this paper, section 2 surveys the recent literature of Arabic and Chinese writer identification. Section 3 summarizes the approach proposed. In sections 4 and 5, segmentation techniques are presented and codebook generation is explained. Experimental results are detailed in section 6 and the conclusions are drawn in the last section.

## 2 RELATED WORKS

A review of writer identification for Arabic is hereby presented. Al-Ma'adeed et al. combined multi-scale edge-hinge and grapheme features for Arabic writer identification in (Al-Ma'adeed et al., 2008). Djeddi and Souici-Meslati applied artificial immune recognition system (AIRS) in (Djeddi and Souici-Meslati, 2011) with grey level co-occurrence matrices (GLCM). kNN, SVM and naïve Bayes were also tested from classification. For historical writer identification, Fecker et al. (Fecker et al., 2014) introduced multiple features, namely histograms of oriented gradients (HOG), oriented basic image (OBI) features and scale-invariant feature transform (SIFT). Feature combination was achieved using averaging and voting schemes. Synthetic grapheme codebooks were proposed for Arabic in (Abdi and Khemakhem, 2015) with a model-based segmentation-free approach. Bagged discrete cosine transform (BDCT) descriptors were utilized in an approach by Khan et al. (Khan et al., 2017) that was trained and tested on IFN/ENIT and AHTID/MW datasets. The textural features of local binary patterns (LBP), local ternary patterns (LTP) and local phase quantization were extracted in (Chahi et al., 2019). As for classification, Chahi et al. opted for a 1-nearest neighbor classifier. Semma et al. presented an approach based on deep learning in (Semma et al., 2021), where features of interest are determined with accelerated segment test (FAST) key points and Harris corner detector. Rasoulzadeh and BabaAli (Rasoulzadeh and BabaAli, 2022) tested a neural network architecture inspired by the vector of locally aggregated descriptors (VLAD) on Arabic KHATT dataset among other datasets. Finally, Ahmed et al. explained concavity/convexity of contours (CON<sup>3</sup>) and contour point curve angle (CPCA) codebooks for Arabic writer identification in (Ahmed et al., 2023).

As for Chinese writer identification, He et al. proposed two-dimensional wavelet textural features and hidden Markov tree (HMT) for similarity classification (He et al., 2008). Tan et al. explained in (Tan et al., 2011) sixteen morphological handwriting features extracted from characters bounding and TBLR

quadrilateral boxes. Four more features also used adaptative similarity adjustment. In a bag-of-features approach (Hu et al., 2014), Hu et al. compared SIFT descriptors in the form of improved Fisher kernels (IFK) and locality-constrained linear coding (LLC) to hard voting (HV) and vector quantization (VQ). Yang et al. employed deep convolutional neural network (DCNN) for writer identification (Yang et al., 2015), trained on an augmented dataset. Another deep learning approach that relies on deep multi-stream CNN is explained in (Xing and Qiao, 2016). Xiong and Lu used contour-directional features (CDF) and character pair similarity measurement (CPSM) in (Xiong and Lu, 2017). A global identification scheme based on edge co-occurrence features (ECF) was enhanced in (Xiong et al., 2019) with displacement field-based similarity (DFS) to confirm characters' authorship. Local phase quantization (LPQ) features were fused with PCA-reduced deep learning features in another approach (Xu et al., 2021). Finally, Semma et al. (Semma et al., 2022) tested features learned from DCNN and encoded with VLAD on multiple languages including Chinese.

## 3 PROPOSED APPROACH

Our proposed approach encompasses a training phase and a testing phase. In the first phase, handwriting is segmented into graphemes that are clustered with the K-medoid algorithm to form a grapheme codebook. Codebooks serve to encode handwriting samples into grapheme histograms prior to similarity comparisons. In the second phase, classification is performed using the City-Block metric, where unknown samples are matched against samples of established authorship. Finally, writer normalized distances obtained with individual codebooks are combined using a log-based weighting formula to enhance identification outcome. Results of Arabic IFN/ENIT and Chinese HIT-MW datasets are compared and interpreted in relation to those of English/German CVL dataset.

## 4 SEGMENTATION

Handwriting is binarized using Otsu's thresholding algorithm (Otsu, 1979). To yield separate graphemes, segmentation operates by disrupting ink continuity at specific morphological features. Four basic segmentation techniques are considered, where handwriting is segmented at the level of (a) contours, (b) morphological points of interest (MPI), (c) vertical projection maxima and (d) filled lower halves minima.

### 4.1 Contours

The external layout of handwriting shapes are separated and taken as segmentation output. Hence, disruption of ink continuity is realized at full connected-component contours. As illustrated in Figure 1, resulting graphemes may consist in Arabic cursive words or piece of words (PAW), Chinese logographs or parts of them, and Roman letters or group of letters.

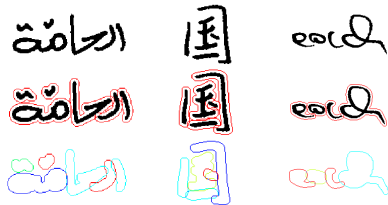


Figure 1: Contour-level segmentation of Arabic, Chinese and English.

### 4.2 Morphological Points of Interest

Branch points, cross-points and corner-points are the MPI marking segmentation disruption points (Abdi et al., 2009). They are detected in thinned handwriting skeletons. The segmentation product is the connected components left after the deletion of MPI pixels along with their 8-connected neighbors and the removal of residual small components. Figure 2 shows the result consisting of detached Arabic segments and ligatures, pen strokes from Chinese logographs and Roman letter fragments.



Figure 2: MPI-level segmentation of Arabic, Chinese and English.

### 4.3 Vertical Projection Maxima

Handwriting is vertically projected and its curve smoothed with a moving average of 8 pixels. Local projection maxima are detected with a  $\delta = 2$  threshold. These maxima, excluding the first and the last ones, are used to divide the script in a vertical manner (Figure 3). Morphologically speaking, the segmentation lines obtained tend to coincide with the height peak of certain characters in Arabic, logographs' largest vertical lines in Chinese and Roman letter halves.

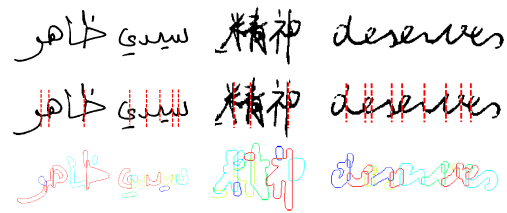


Figure 3: Segmentation of Arabic, Chinese and English based on vertical projection maxima.

### 4.4 Filled Lower Halves Minima

Handwriting shapes are processed as connected components. First, enclosed regions in the shape are filled. Then, using the position of its largest horizontal projection line, the upper half of its bounding box is filled so that only the protruding structures of the lower half remain intact. After vertical projection, curve smoothing occurs with a window of 8 pixels. Finally, the local minima x-coordinates, except for the first and last ones, are detected with a  $\delta = 1$  threshold and retained for vertical segmentation of the original handwriting. An example of segmentation results is depicted in Figure 4.

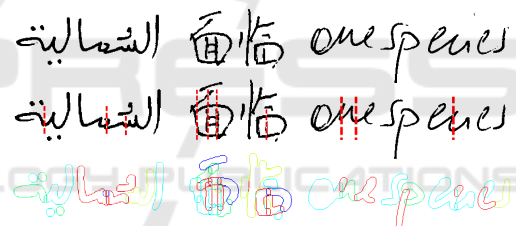


Figure 4: Segmentation of Arabic, Chinese and English based on filled lower halves minima.

Each previously described segmentation technique concludes with a post-processing step, which starts with a morphological dilation using a structuring element of 5 pixels and ends with extracting the contours of the resulting shapes as the final graphemes to be retained.

## 5 CODEBOOK GENERATION

Codebooks are created in two steps: graphemes are first encoded into feature vectors (FV), then processed with a clustering algorithm.

Feature encoding consists in subsampling grapheme shapes to 50 evenly spaced points and encoding them into 100-dimensional FVs of polar coordinates. An angular coordinate of a point is taken in  $[0, 90^\circ]$  with origin (0,0) being the upper-left corner of the grapheme's bounding box, and its distance

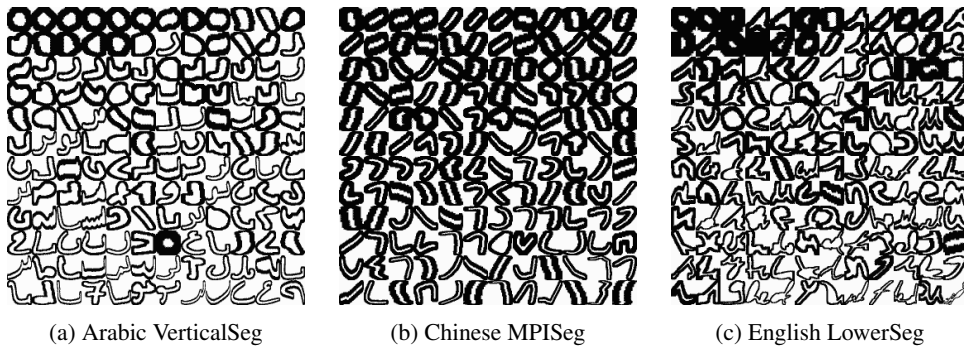


Figure 5: Best performing segmentation-issued codebooks for (a) IFN/ENIT, (b) HIT-MW and (c) CVL datasets.

coordinate is noted in pixels. FVs are standardized feature-wise to mean zero and standard deviation one according to the training dataset’s global values.

Clustering of encoded graphemes is realized on the training dataset with the K-medoid algorithm (Kaur et al., 2014). The latter takes graphemes as cluster centers (medoids) instead of graphemes’ mean values (centroids) and is more resilient to noise and outliers than K-means. For initial medoid attribution, the K-means++ algorithm is retained (Arthur and Vassilvitskii, 2007). Average medoid dissimilarity is iteratively minimized until convergence. Therefore, four codebooks of 144 graphemes are computed for Arabic, Chinese and English for each segmentation technique explained earlier.

## 6 EXPERIMENTAL RESULTS

Experimentations are realized on handwriting samples by 300 writers from Arabic IFN/ENIT (Pechwitz et al., 2002) and 300 writers from Chinese HIT-MW (Su et al., 2007). An additional 300 writers from English/German CVL (Kleber et al., 2013) are retained as a control group for results comparison (Figure 6).

Handwriting data is divided into equal training and testing parts per writer. To classify a testing sample, the training samples are ranked according to their similarity yielding one (Top1) or  $N$  (TopN) most probable matches. Distances are computed with the City-Block metric as follows:

$$d(u, v) = \sum_{i=1}^n |u_i - v_i| \tag{1}$$

where  $u$  is the testing sample,  $v$  is the training sample and  $n$  is the feature vector dimensionality (100).

### 6.1 Writer Identification

Writer identification results of the codebooks associated to the different segmentation techniques are re-

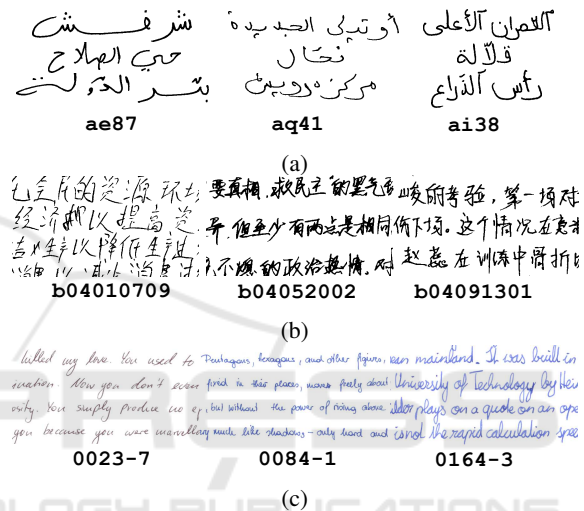


Figure 6: Handwriting samples from (a) IFN/ENIT, (b) HIT-MW and (c) CVL datasets.

ported for IFN/ENIT, HIT-MW and CVL. In Table 1, writer identification accuracy is presented in the form of Top1 and Top5 percentages (Top1% and Top5%). ContourSeg, MPISeg, VerticalSeg and LowerSeg are the codebooks based on contours, MPI, vertical projection maxima and filled lower halves minima segmentations, respectively.

All four segmentation techniques worked well for CVL with a Top1% ranging from 98.00% to 99.00%. Lower performances were obtained for Arabic and Chinese with IFN/ENIT giving a Top1% in the 58.33-70.67% range and HIT-MW a Top1% in the 45.00-59.00% range. The best results were achieved with VerticalSeg for Arabic (Top1% = 70.67%) using the codebook shown in Figure 5a, and with MPISeg for Chinese (Top1% = 59.00%) using the codebook in Figure 5b. On the other hand, LowerSeg did not perform well for Arabic (Top1% = 58.33%) and Chinese (Top1% = 45.00%), despite being the best performing codebook with CVL (Top1% = 99.00%) as illustrated in Figure 5c.

Table 1: Writer identification results.

Codebook $\nabla$	IFN/ENIT		HIT-MW		CVL (control)	
	Top1%	Top5%	Top1%	Top5%	Top1%	Top5%
ContourSeg	59.33	79.67	49.67	73.67	98.33	99.00
MPISeg	62.33	86.33	59.00	80.00	98.67	99.33
VerticalSeg	70.67	88.67	47.67	71.67	98.00	99.67
LowerSeg	58.33	82.00	45.00	68.00	99.00	100.00

Table 2: Writer identification combination (IFN/ENIT).

Codebook combination $\nabla$	Top1%	Top5%
VerticalSeg & MPISeg	95.67	98.67
VerticalSeg & MPISeg & ContourSeg	95.67	98.67
VerticalSeg & MPISeg & ContourSeg & LowerSeg	96.33	98.00

Table 3: Writer identification combination (HIT-MW).

Codebook combination $\nabla$	Top1%	Top5%
MPISeg & ContourSeg	89.67	97.00
MPISeg & ContourSeg & VerticalSeg	90.33	97.67
MPISeg & ContourSeg & VerticalSeg & LowerSeg	91.33	98.33

For morphological reasons, the codebooks that succeeded in capturing inter-writer variability of CVL writers were not as biometrically significant with IFN/ENIT and HIT-MW. To address this issue, a multi-segmentation approach to Arabic and Chinese is proposed.

## 6.2 Codebooks Combination

After classification, the training/testing distance matrix obtained with each codebook is min-max normalized in  $[0, 150] \cap \mathbb{N}$ , using the global values of the training dataset inter-sample distance matrix. Matrices are ordered by their Top1% in a descending order. The following reciprocal log weighting formula is used to merge matrices element-wise:

$$\tilde{d} = \frac{1}{n} \sum_{i=1}^n d_i \left( \frac{1/\log(1+i/10)}{\sum_{i=1}^n 1/\log(1+i/10)} \right) \quad (2)$$

where  $\tilde{d}$  is the merged distance,  $n$  is the codebooks count (up to 4) and  $d_i$  the training/testing distance according to codebook number  $i$ . Table 2 and Table 3 present the combination results for IFN/ENIT and HIT-MW, respectively.

Codebook combination substantially increased identification rates for IFN/ENIT and HIT-MW. For Arabic, the combination of the two best performing codebooks, i.e., VerticalSeg and MPISeg, increased Top1% by 25 percentage points. For Chinese, Top1% gained 30.67 percentage points by combining MPISeg and ContourSeg. The best Top1% values are obtained by combining all codebooks, and are 96.33% and 91.33% for IFN/ENIT and HIT-MW, respectively.

## 7 CONCLUSIONS

This paper addressed the efficiency of segmentation for Arabic and Chinese writer identification in comparison to Latin alphabet handwriting. Segmentation schemes based on contours, MPI, vertical projection maxima and filled lower halves minima were considered. Then, the issued graphemes were sub-sampled and their polar coordinates encoded into feature vectors. Codebooks were generated using the K-medoid clustering algorithm with experimentations performed on Arabic IFN/ENIT, Chinese HIT-MW and English/German CVL datasets. Taken individually, the studied segmentation techniques did not perform as well for Arabic (Top1% = 70.67%) and Chinese (Top1% = 59.00%) as they did for English/German (Top1% = 99.00%). However, combining codebook results at the level of normalized training/testing distance matrices substantially enhanced writer identification, with Top1% = 96.00% for Arabic and Top1% = 91.33% for Chinese. Further investigations are being conducted to assess the impact of other aspects such as preprocessing and clustering on Arabic and Chinese writer identification.

## REFERENCES

- Abdi, M. N. and Khemakhem, M. (2015). A model-based approach to offline text-independent arabic writer identification and verification. *Pattern Recognition*, 48(5):1890–1903.
- Abdi, M. N., Khemakhem, M., and Ben-Abdallah, H.

- (2009). A novel approach for off-line arabic writer identification based on stroke feature combination. In *2009 24th International Symposium on Computer and Information Sciences*, pages 597–600. IEEE.
- Ahmed, B. Q., Hassan, Y. F., and Elsayed, A. S. (2023). Offline text-independent writer identification using a codebook with structural features. *Plos one*, 18(4):e0284680.
- Al-Ma'adeed, S., Al-Kurbi, A.-A., Al-Muslih, A., Al-Qahtani, R., and Kubisi, H. A. (2008). Writer identification of arabic handwriting documents using grapheme features. In *2008 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE.
- Alshahrani, A. A. (2008). Arabic script and the rise of arabic calligraphy. *Online Submission*.
- Amara, N. E. B. and Bouzlama, F. (2003). Classification of arabic script using multiple sources of information: State of the art and perspectives. *Document Analysis and Recognition*, 5:195–212.
- Arthur, D. and Vassilvitskii, S. (2007). K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.
- Chahi, A., Ruichek, Y., Touahni, R., et al. (2019). An effective and conceptually simple feature representation for off-line text-independent writer identification. *Expert Systems with Applications*, 123:357–376.
- Chen, M. J. (1996). An overview of the characteristics of the chinese writing system. *Asia Pacific Journal of Speech, Language and Hearing*, 1(1):43–54.
- Djeddi, C. and Souici-Meslati, L. (2011). Artificial immune recognition system for arabic writer identification. In *International Symposium on Innovations in Information and Communications Technology*. IEEE.
- Fecker, D., Asit, A., Märgner, V., El-Sana, J., and Fingscheidt, T. (2014). Writer identification for historical arabic documents. pages 3050–3055, Stockholm, Sweden. IEEE.
- Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., and Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.
- He, Z., You, X., and Tang, Y. Y. (2008). Writer identification of chinese handwriting documents using hidden Markov tree model. *Pattern Recognition*, 41(4):1295–1307.
- Hu, Y., Yang, W., and Chen, Y. (2014). Bag of features approach for offline text-independent chinese writer identification. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2609–2613. IEEE.
- Kaur, N. K., Kaur, U., and Singh, D. (2014). K-Medoid clustering algorithm-a review. *Int. J. Comput. Appl. Technol*, 1(1):42–45.
- Khan, F. A., Tahir, M. A., Khelifi, F., Bouridane, A., and Almotaryi, R. (2017). Robust off-line text independent writer identification using bagged discrete cosine transform features. *Expert Systems with Applications*, 71:404–415.
- Kleber, F., Fiel, S., Diem, M., and Sablatnig, R. (2013). CVL-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th international conference on document analysis and recognition*, pages 560–564. IEEE.
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66.
- Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., Amiri, H., et al. (2002). IFN/ENIT-database of hand-written arabic words. In *Proc. of CIFED*, volume 2, pages 127–136. Citeseer.
- Rasoulzadeh, S. and BabaAli, B. (2022). Writer identification and writer retrieval based on NetVLAD with re-ranking. *IET Biometrics*, 11(1):10–22.
- Semma, A., Hannad, Y., Siddiqi, I., Djeddi, C., and El Kettani, M. E. Y. (2021). Writer identification using deep learning with fast keypoints and harris corner detector. *Expert Systems with Applications*, 184:115473.
- Semma, A., Hannad, Y., Siddiqi, I., Lazrak, S., and El Kettani, M. E. Y. (2022). Feature learning and encoding for multi-script writer identification. *International Journal on Document Analysis and Recognition (IJ-DAR)*, 25(2):79–93.
- Su, T., Zhang, T., and Guan, D. (2007). Corpus-based HIT-MW database for offline recognition of general-purpose chinese handwritten text. *International Journal of Document Analysis and Recognition (IJ-DAR)*, 10:27–38.
- Tahsildar, M. N. (2019). Chinese language complexities among international students in china. *Education Quarterly Reviews*, 2(1):67–76.
- Tan, J., Lai, J.-H., Wang, C.-D., and Feng, M.-S. (2011). A stroke shape and structure based approach for off-line chinese handwriting identification. *International Journal of Intelligent Systems and Applications*, 3(2):1.
- Taylor, M. M. and Taylor, I. (2014). Writing and literacy in chinese, korean and japanese. *Writing and Literacy in Chinese, Korean and Japanese*, pages 1–506.
- Wang, J., Chen, H.-C., Radach, R., and Inhoff, A. (1999). *Reading Chinese script: A cognitive analysis*. Psychology Press.
- Xing, L. and Qiao, Y. (2016). Deepwriter: A multi-stream deep CNN for text-independent writer identification. In *2016 15th international conference on frontiers in handwriting recognition (ICFHR)*, pages 584–589. IEEE.
- Xiong, Y.-J., Liu, L., Lyu, S., Wang, P. S., and Lu, Y. (2019). Improving text-independent chinese writer identification with the aid of character pairs. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(02):1953001.
- Xiong, Y.-J. and Lu, Y. (2017). Chinese writer identification using contour-directional feature and character pair similarity measurement. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 119–124. IEEE.
- Xu, Y., Chen, Y., Cao, Y., and Zhao, Y. (2021). A deep learning method for chinese writer identification with

feature fusion. In *Journal of Physics: Conference Series*, volume 1883, page 012142. IOP Publishing.

Yang, W., Jin, L., and Liu, M. (2015). Chinese character-level writer identification using path signature feature, DropStroke and deep CNN. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 546–550. IEEE.

