Scribe Verification in Chinese manuscripts using Siamese, Triplet, and Vision Transformer Neural Networks

Dimitris-Chrysovalantis Liakopoulos[1], Yanbo Zhang[2], Chongsheng Zhang[2] and Constantine Kotropoulos[1]

[1]Department of Informatics, Aristotle University of Thessaloniki, Greece

[2]School of Computer and Information Engineering, Henan University, Kaifeng, China

The Eighteenth International Conference on Advances in Databases, Knowledge, and Data Applications

March 10, Valencia, Spain

Presenter's bio

Constantine Kotropoulos received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki. He is a Full Professor in the Department of Informatics at the Aristotle University of Thessaloniki. He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA, during the academic year 2008-2009. He also conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland, during the summer of 1993. He has co-authored 77 journal papers and 228 conference papers and contributed 9 chapters to edited books in his areas of expertise. He is co-editor of the book "Nonlinear Model-Based Image/Video Processing and Analysis" (John Wiley and Sons, 2001).
His research interests include forensics, audio, speech and language processing, signal processing, pattern recognition, multimedia information retrieval, biometrics, and forensics.
Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a fellow of IARIA, a senior member of the IEEE, and a member of EURASIP, IAPR, and the Technical Chamber of Greece. He was a Senior Area Editor of the IEEE Signal Processing Letters (SPL). He serves as Editor-in-Chief of SPL (2026-2028). He has been a member of the Editorial Board of the journals Advances in Multimedia, Computer Methods in Biomechanics & Biomedical Engineering: Imaging & Visualization, Artificial Intelligence Review, MDPI Imaging, MDPI Signals, and MDPI Methods and Protocols. Prof. Kotropoulos served as Track Chair for Signal Processing in the 6th Int. Symposium on Communications, Control, and Signal Processing, Athens, 2014; Program Co-Chair of the 4th Int. Workshop on Biometrics and Forensics, Limassol, Cyprus, 2016; Technical Program Chair of the XXV European Signal Processing Conf., Kos, Greece, 2017; Technical Program Chair of the 5th IEEE Global Conf. Signal and Information Processing, Montreal, Canada, 2017; General Chair of the 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop, Nafplio, Greece; Technical Program Chair of the 2023 IEEE International Conf. on Acoustics, Speech, and Signal Processing, Rhodes, Greece.

Outline

- **Goal:** Determine whether two manuscript fragments were written by the same scribe.
- **Traditional analysis** relies on expert paleographers.
- **Manual inspection** is time-consuming and subjective.
- **Automatic verification** supports objective, historical analysis, preservation, and authentication.

- **Scribe verification** is a similarity learning problem.
- Instead of classification, we **learn an embedding space**.
- **Same-scribe samples** are close in embedding space, while **different-scribe samples** are far apart.
- **Siamese and Triplet networks** are well-suited for such tasks.

- **Comparative evaluation** of **Siamese CNN**, **Triplet CNN**, and **Vision Transformer (ViT)** architectures.
- Unified PyTorch framework for **metric learning**.
- Evaluation on both ancient **(Tsinghua)** and modern **(MCCD Subset)** Chinese manuscripts.
- Standardized **fixed-pair evaluation** using **Receiver Operating Characteristic Curve (ROC)/Area under the ROC (AUC), Accuracy (ACC), False Acceptance Rate (FAR), and False Rejection Rate (FRR)**.

- The **Tsinghua Bamboo Slips** dataset consists of digitized images of ancient Chinese manuscripts discovered in 2008 at Tsinghua University.
- Each fragment corresponds to a section of bamboo slip text written in classical Chinese, attributed to specific scribes[1].

Figure 1: Representative examples from the Tsinghua Bamboo Slips dataset.

[1] Haiyang Wang et al. "Tsinghua Bamboo Slip Scribe Verification Using Siamese Networks". *Heritage Science* (2025)

- The **Multi-Attribute Chinese Calligraphy Character Dataset (MCCD)** is a large-scale modern collection of handwritten Chinese character images created by multiple professional calligraphers.
- To adapt this dataset for the scribe verification task, a curated subset was created by selecting only the Calligraphers folder and including the scribes with the largest number of samples[2].

Figure 2: Representative examples from the MCCD subset dataset.

[2] Yixin Zhao, Yuyi Zhang, and Lianwen Jin. *MCCD: A Multi-Attribute Chinese Calligraphy Character Dataset Annotated with Script Styles, Dynasties, and Calligraphers*. 2025. arXiv: 2507.06948 [cs.CV]. URL: https://arxiv.org/abs/2507.06948

- Each model is designed to learn an embedding representation of handwriting fragments such that samples from the same scribe are mapped close together in the embedding space.
- **Siamese CNN-based Models:** The first set of experiments employed convolutional Siamese architectures trained with the contrastive loss function.
- **Triplet Network:** The second set of experiments was based on the Triplet learning paradigm, a metric-learning approach designed to learn embeddings in which examples from the same class lie closer together than examples from different classes[3].
- **ViT Siamese Model:** To investigate the effectiveness of self-attention mechanisms in scribe verification, a ViT was implemented within the Siamese framework[4].

[3] Elad Hoffer and Nir Ailon. *Deep Metric Learning Using Triplet Network*. 2014. arXiv: 1412.6622 [cs.LG]. URL: https://arxiv.org/abs/1412.6622

[4] Alexey Dosovitskiy et al. *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. 2020. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929

- The contrastive loss function was used to train the Siamese and ViT models. Given a pair of images (x_1, x_2) with label $y \in \{0, 1\}$, where $y = 1$ indicates that both images belong to the same scribe and $y = 0$ indicates different scribes, the Euclidean distance between their embeddings is defined as $D(x_1, x_2) = \|f(x_1) - f(x_2)\|_2$.
- The loss is given by

$$L_{contrastive}(x_1, x_2) = \frac{1}{2}\, y\, D^2(x_1, x_2) + \frac{1}{2}(1 - y) \cdot \max\left(0, m - D(x_1, x_2)\right)^2 \qquad (1)$$

where m is a predefined margin that controls the separation between positive and negative pairs. The first term penalizes large distances for positive pairs (same scribe), while the second term penalizes small distances for negative pairs that lie within the margin.

- The triplet loss was applied to the Triplet Network, which processes three images simultaneously: an anchor x_a, a positive x_p (same scribe), and a negative x_n (different scribe). The goal is to enforce that the distance between the anchor and the positive sample is smaller than the distance between the anchor and the negative sample by at least a margin m.
- The loss is given by

$$L_{triplet}(x_a, x_p, x_n) = \max\left(0, D(x_a, x_p) - D(x_a, x_n) + m\right), \tag{2}$$

where $D(x_a, x_p)$ and $D(x_a, x_n)$ denote the Euclidean distances between the corresponding embeddings.

- Each image was first converted to a three-channel RGB format, regardless of its original grayscale input, to maintain compatibility with pretrained ImageNet backbones.
- All images were then resized to 224×224 pixels, following the configuration used in the Tsinghua Bamboo Slip experiments.
- Normalization was applied using the ImageNet mean and standard deviation:

$$\mu = (0.485, 0.456, 0.406), \quad \sigma = (0.229, 0.224, 0.225). \tag{3}$$

- Lightweight augmentations were applied randomly during training: 1) Horizontal flipping, 2) Random grayscale inversion (gray-flip) for minority classes, and 3) Random brightness and contrast adjustments

- Training samples were generated dynamically at runtime as follows.
- **Positive pairs:** Two images from the same scribe or calligrapher folder.
- **Negative pairs:** Two images from different scribes.
- **Triplets:** An anchor, a positive from the same scribe, and a negative from a different scribe.
- Positive and negative pairs were balanced using a **1:1 ratio** to prevent bias toward a particular class type.
- **Corrupted or unreadable image** files were automatically detected and replaced with a resampled, valid image from the same class.

- **Adam optimizer** used with an **initial learning rate** of 10^{-3} and no weight decay.
- A **batch size of 32** was employed, and each training session ran for **30 epochs**.
- A decision threshold τ was applied to classify whether two fragments x_1 and x_2 were written by the same or different scribes:

$$\text{prediction} = \begin{cases} \text{same scribe}, & D(x_1, x_2) < \tau, \\ \text{different scribes}, & D(x_1, x_2) \geq \tau. \end{cases} \tag{4}$$

The optimal threshold was determined by scanning possible values and selecting the one that maximized overall classification accuracy.

- To assess the discriminative power of the learned embeddings, several quantitative metrics were computed:
- **AUC**, summarizing ROC performance by a single scalar value.
- **ACC**, measuring the proportion of correctly classified pairs.
- **FAR**, quantifying the ratio of negative pairs incorrectly predicted as positive ones (i.e., same).
- **FRR**, quantifying the ratio of positive pairs incorrectly predicted as negative (i.e., different).

Experimental Results

Quantitative Performance Comparison

Model	**Dataset**	**AUC**	**ACC** (%)	**FAR** (%)	**FRR** (%)
MobileNetV3 (Siamese)	Tsinghua	0.897	82.27	13.28	22.19
MobileNetV3+ Custom (Siamese)	Tsinghua	**0.958**	**89.19**	**5.63**	**15.99**
ResNet18 (Siamese)	Tsinghua	0.921	84.65	16.56	14.15
ResNet34 (Siamese)	Tsinghua	0.912	82.56	19.5	15.37
ResNet50 (Siamese)	Tsinghua	0.884	80.4	21.5	17.69
VGG19 (Siamese)	Tsinghua	0.857	82.07	14.95	20.91
ViT-B/16 (Siamese)	Tsinghua	0.829	77	25.59	20.41
MobileNetV3+ Custom (Triplet)	Tsinghua	0.945	88.16	13.59	10.15
MobileNetV3 (Siamese)	MCCD	0.949	88.99	11.44	10.59
MobileNetV3+ Custom (Siamese)	MCCD	0.949	88.5	12.77	10.23
ResNet18 (Siamese)	MCCD	0.948	88.24	13.93	9.58
ResNet34 (Siamese)	**MCCD**	**0.952**	**89.06**	**14.08**	**7.79**
ResNet50 (Siamese)	MCCD	0.931	86.93	17.91	8.23
VGG19 (Siamese)	MCCD	0.9	82.93	25.05	9.10
ViT-B/16 (Siamese)	MCCD	0.89	82.06	24.41	11.47
MobileNetV3+ Custom (Triplet)	MCCD	0.931	87.63	16.49	8.66

Experimental Results

Best Performance Model Evaluation - MCCD dataset

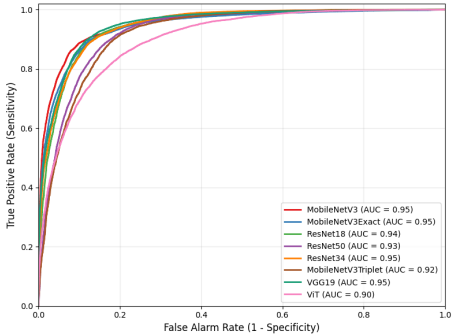


Figure 3: ROC curves of various models on the MCCD dataset[5]

.

[5] Thanks to my undergraduate student Jaho Islami.

Experimental Results

Best Performance Model Evaluation - MobileNetV3 Custom on Tsinghua dataset

- The **MobileNetV3+ Custom Siamese model** trained with **contrastive loss** achieved the best overall results, obtaining an **AUC of 0.958** and an **accuracy of 89.19%** in the Tsinghua Bamboo Slips Dataset.

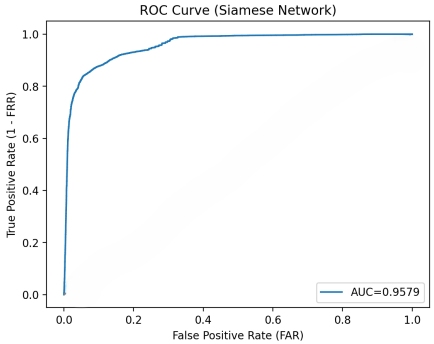


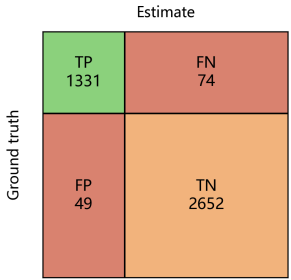
Figure 4: ROC curve of the MobileNetV3+ Custom Siamese model on the Tsinghua Bamboo Slips dataset.

Experimental Results

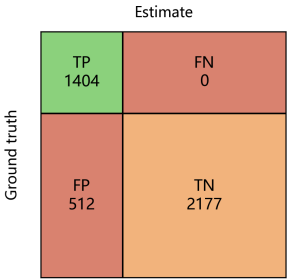
Best Performance Model Evaluation - MobileNetV3 Custom on Tsinghua dataset

- Apart from the ROC Curve, a confusion matrix is being exported for each class, visualizing True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) results.

Confusion Matrix (k vs rest) — Bao Xun



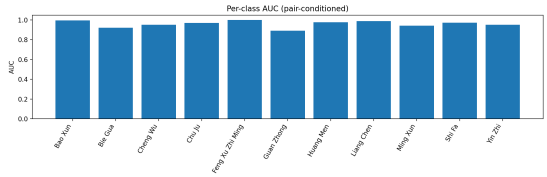
Confusion Matrix (k vs rest) — Bie Gua

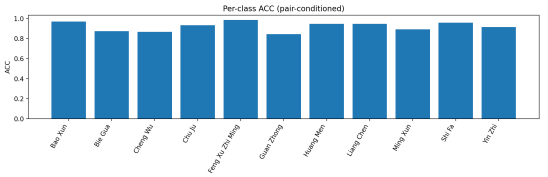


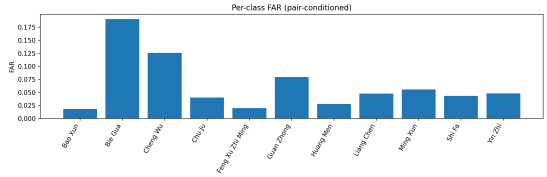
Experimental Results

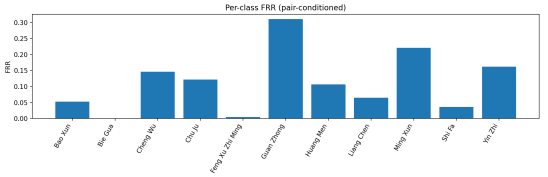
Best Performance Model Evaluation - MobileNetV3 Custom on Tsinghua dataset

- Additional graphs have been exported to evaluate per-class distribution regarding AUC, ACC, FAR, and FRR:









Conclusion and Future Work

Conclusion

1. The contrastive **Siamese networks have been highly effective** for this task, achieving robust performance across both datasets.
2. The **MobileNetV3+ Custom Siamese model** has achieved the top performance on the Tsinghua Bamboo Slips dataset despite challenging image degradation.
3. On the modern MCCD dataset, the **ResNet34 Siamese model** also achieved the highest overall accuracy and AUC, confirming the suitability of lightweight architectures for high-quality calligraphy data.

Conclusion and Future Work

Future Work

1. Exploring **hybrid CNN–Transformer architectures** that combine efficient local feature extraction with global attention mechanisms.
2. Incorporating hard positive and hard negative mining strategies to **improve triplet-based training efficiency**.
3. Expanding the dataset with **additional manuscript sources** to enable large-scale pretraining and improve generalization.
4. Investigating **few-shot** and **one-shot** learning approaches to scribe verification with minimal labeled data.
5. Including **multi-seed experiments** to quantify performance variability and further assess the robustness of metric learning methods that are sensitive to initialization and sampling strategies.

GitHub Code

The code for the proposed framework can be found at:

`https://github.com/dimiliakop/ScribeVerification`

or by scanning the QR code:

Figure 5: QR code to GitHub.

Thank You!

Thank you very much for your attention.
Q & A?
Email costas@csd.auth.gr