# Enhancement of scribal hands identification via self-supervised learning (Extended Abstract)

Lorenzo Lastilla[1]

[1]Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, Italy

**Abstract**

In this paper, the first successful application of the recent framework of self-supervised learning to the problem of handwriting identification for medieval and modern manuscripts is presented. To this end, a novel dataset consisting of both labeled and unlabeled manuscripts extracted from the Vatican Apostolic Library was produced. Moreover, this contribution shows that pretraining a convolutional neural network by leveraging large amounts of unlabeled manuscripts and fine-tuning this model to the task of interest significantly outperforms other baselines, including the common setup of initializing the network from general-domain features, or training the model from scratch, also in terms of generalization power. Overall, these results reveal the strong potential of self-supervised techniques in the field of digital paleography, where unlabeled data is nowadays available, while labeled data is scarcer.

**Keywords**

Self-Supervised Learning, Medieval and Modern Manuscripts, Handwriting Identification

## 1. Introduction and related works

In recent years, the long-standing issue of automatic handwriting identification (HI) – the subdivision of the texts into parts belonging to distinct scribes on the basis of the respective handwriting style – has been discussed both from a theoretical perspective [1, 2, 3] and an application point of view. Despite the increasing use of deep learning techniques to address this problem [4, 5, 6, 7], HI continues to be carried out with traditional methods by paleographers, due to the costs, time and expertise required for data labeling.

This work highlights the benefits of using a self-supervised learning (SSL) approach for the HI task on medieval and modern manuscripts (more precisely, a set of 24 digitized manuscripts selected from the Vatican Apostolic Library [8]), the vastness of which is considerable, even if most of the manuscript pages are not annotated with the information of the *copyist* who physically wrote them. SSL, indeed, is gradually taking hold to address the problem of learning good image representations from a few labeled examples while making best use of many unlabeled instances [9, 10], which would minimize the dependence on potentially costly corpora of manually annotated data [11], and makes this strategy ideal for the HI task. In particular, SSL methods try to solve a "pretext task" (which is not of genuine interest) to learn – from

*18th Italian Research Conference on Digital Libraries, Padova, Italy – 24-25 February 2022*

lorenzo.lastilla@uniroma1.it (L. Lastilla)

https://phd.uniroma1.it/web/LASTILLA-LORENZO_nP1612494_EN.aspx (L. Lastilla)

0000-0003-1099-6270 (L. Lastilla)

**Table 1**
Number of copyists and century of realization of the 24 selected manuscripts.

| Manuscript ID | Number of copyists | Century | Manuscript ID | Number of copyists | Century |
|---|---|---|---|---|---|
| Vat. lat. 12910 | 0 | XI | Vat. lat. 4939 | 0 | XII |
| Vat. lat. 2669 | 0 | XIII | Vat. lat. 4958 | 0 | XI |
| Vat. lat. 3313 | 0 | IX | Vat. lat. 4965 | 2 | IX |
| Vat. lat. 3317 | 0 | X | Vat. lat. 5775 | 0 | IX |
| Vat. lat. 378 | 3 | XI | Vat. lat. 579 | 0 | XI |
| Vat. lat. 3833 | 0 | XII | Vat. lat. 588 | 0 | XIV |
| Vat. lat. 3868 | 0 | IX | Vat. lat. 5951 | 3 | IX |
| Vat. lat. 42 | 0 | XII | Vat. lat. 620 | 0 | XII |
| Vat. lat. 4217 | 3 | XI | Vat. lat. 653 | 4 | XI |
| Vat. lat. 4220 | 8* | XVI | Vat. lat. 8487 | 2 | XI |
| Vat. lat. 4221 | 8* | XVI | Vat. lat. 907 | 2 | XIII |
| Vat. lat. 43 | 0 | IX | Vat. lat. 9882 | 0 | IX |

unlabeled data – representations that can be transferred to other tasks of actual interest (the "downstream tasks"), which often have only a few labeled instances [12, 13, 14, 15, 16, 17].

In practice, the proposed methodology consists of two main stages. First, all the pages contained in the manuscripts undergo the Online Bag-of-Visual-Words (OBoW) reconstruction-based SSL approach described in [15, 18]. Then, the available copyists are split into a background set and an evaluation set (whose samples are never seen during training nor validation), and a linear layer is trained on top of the frozen base encoder, with the aim of minimizing a triplet margin loss [19, 20]. The results obtained show that the visual representations learned in a self-supervised fashion outperform the ImageNet [21] ones with respect to the HI task, as well as the features learned after training the backbone model from scratch, also in terms of generalization power.

## 2. Case study

24 high-resolution digital manuscripts, included among the tables for Latin paleography exercises published by the Vatican Apostolic Library in 2004 [22], which collect very recognizable graphic types [23, 24], were selected from [8], obtaining a final corpus of 8745 pages and 27 scribes (identified in 9 manuscripts only – Vatt. latt. 4220 and 4221 share the same set of 8 copyists). The selected manuscripts, together with the number of available copyists and the century of realization, are recalled in Table 1.

Starting from the overall group of selected pages, two different datasets were created: as to the pretext task, the 8745 samples – organized in 24 classes – were randomly split into training, validation and test sets according to the ratio 0.8-0.15-0.05. For the handwriting identification task, instead, only the annotated pages were selected. The available copyists were split into an *evaluation set* (consisting of the 4 scribes from Vat. lat. 653, excluded from the training and validation stages of this task) and a *background set* (including the 23 remaining scribes).

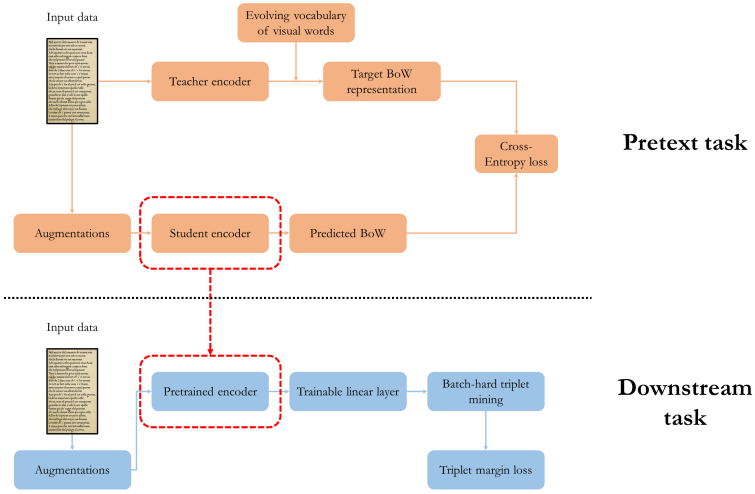



**Figure 1:** Schematic representation of both the self-supervised pretraining stage (pretext task) and the handwriting identification task (downstream task), conceived as a triplet margin loss minimization.

## 3. Methodology and results

The SSL strategy adopted in this work can be summarized as follows: a CNN-based feature extractor (or student network – a ResNet18-based encoder [25]) is trained to predict, with unlabeled data only, the BoW representation of a $380 \times 380$ random crop of a manuscript page given as input a set of perturbed crops of that page [18]. The BoW representation is generated by a momentum-updated teacher network, which receives as input the $380 \times 380$ random crop. The set of perturbed crops, instead, is obtained through several augmentations, including radiometric perturbations, Gaussian blur, random erasing, and mild geometric distortions.

Once self-supervised pretraining is completed, the frozen features of the student encoder are involved in the HI task, based on the minimization of a triplet margin loss, which can be seen as learning a distance function useful for discriminating instances belonging to different classes in the embedding space. At this stage, a linear layer only, added to backbone model, is trained to extract more powerful representations with respect to the task of interest. The triplets are generated through an online batch-hard triplet mining strategy, which is the optimal configuration for this kind of task [26, 27]. As to the pairwise distance function involved in the loss computation, the $L^2$ norm was chosen. The downstream task was carried out based on two mutually exclusive data augmentation schemes. Scheme **A)** is based on the extraction of a random $380 \times 380$ crop from the page and the application of a similar set of perturbations as the pretext task; scheme **B)**, instead, extends the same set of perturbations to the whole page: consequently, the encoder – which receives as input a page of arbitrary size – operates as a "fully convolutional" network [28]. In Figure 1, a schematic representation of both the self-supervised pretraining stage and the handwriting identification task is provided.

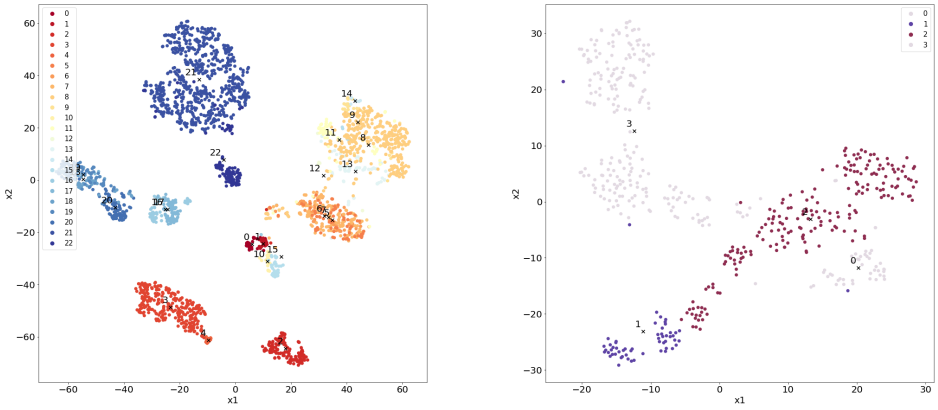
All the experiments were carried out using a Tesla V100 SXM2 32GB GPU, and involved a ResNet18-based architecture. The source code used for the experiments is available at

**Table 2**
Performance obtained for the HI task for the 6 tests, with respect to the MAP metric.

| Mode | Scheme | MAP [%] – Background set | MAP [%] – Evaluation set |
|---|---|---|---|
| OBoW pretraining | **B)** | **74.8*** | 72.0 |
| ImageNet pretraining | **B)** | 69.7 | 64.9 |
| Training from scratch | **B)** | 60.8 | 58.8 |
| OBoW pretraining | **A)** | 71.7 | **79.0*** |
| ImageNet pretraining | **A)** | 63.7 | 67.5 |
| Training from scratch | **A)** | 48.5 | 59.1 |

https://github.com/L9L4/HI-SSL. As to the BoW reconstruction task, the student encoder was trained for 100 epochs; the batch size was fixed to 64; finally, Stochastic Gradient Descent (SGD) was adopted, with learning rate set to 0.03 and progressively adjusted up to the final value of 0.00003 through a cosine scheduler with an initial warmup of 5 epochs.

As to the HI task, it was faced based on both the augmentation schemes **A)** and **B)**, considering 3 different configurations (and thus ending up with 6 tests in total): a linear layer was trained on top of the frozen backbone model pretrained with OBoW; a linear layer was trained on top of the ImageNet frozen features (also in this case, a ResNet18 backbone encoder – but pretrained on the ImageNet dataset – was used); a model initialized with random weights (but characterized by the same architecture as the other two cases) was fully trained from scratch directly on the downstream task. For all the 6 tests, the output dimension of the linear layer (embedding width) was fixed to 1024, while the margin $m$ of the triplet margin loss was set to 0.2 [20]. The model was trained for 100 epochs with SGD optimization: the learning rate, starting from 0.15, was increased up to 0.6 through a linear warmup for the first 10 epochs, and then decayed with a cosine annealing up to 0.0015. The batch size was set to 256 for the tests carried out under scheme **A)**, and to 32 for scheme **B)**. To quantitatively assess the performance of the single tests for the HI task, the Mean Average Precision (MAP) was computed. In Table 2, the results obtained for each test in terms of MAP are shown. It is immediately evident that the SSL-based approach is far more effective for the task of interest than the baselines, under both the **A)** and **B)** data augmentation schemes. This is true, indeed, both for the background scribes and for the evaluation ones, used to test the generalization capacity of the model, achieving a MAP of 74.8% for the background set, obtained under the **B)** scheme, and of 79.0% for the evaluation set – **A)** scheme. In Figures 2a and 2b, it is possible to visualize the 2D projection of the embeddings of the manuscript pages for the best results obtained among the 6 tests, both for the background and the evaluation set (together with the respective cluster centroids), after dimensionality reduction through the t-SNE technique [29]. Figures 2a and 2b are particularly helpful to appreciate the capability of the proposed framework of effectively performing the HI task, even for manuscripts excluded from training. This seems to suggest the possibility of extending the approach to any non-annotated manuscript, which could be partitioned through zero-shot learning. Figure 2a is also useful, however, to highlight the limitations of the proposed approach, showing the difficulties in properly clustering some scribes coming from very specific manuscripts (Vatt. latt. 4217 – scribes 5-7 in Figure 2a –, 4220 and 4221 – scribes 8-15), which constitute a particularly complex subset, since the copyists who realized them aimed for the maximum handwriting



(a) Best result for the background set. (b) Best result for the evaluation set.

**Figure 2:** Best results obtained among the 6 tests, both for the background and the evaluation set.

uniformity. Hence, the representations extracted from the model, generally sufficient for the other manuscripts, might have been unsuitable to grasp the set of discriminating elements valid for this subgroup.

## 4. Conclusion

In this paper, the task of automatic HI for ancient manuscripts was addressed in the face of the scarcity of large and annotated datasets, and the first empirical validation of the SSL framework in the medieval and modern manuscript domain was provided, assessing its capability to learn effective visual representations from a large amount of raw data and then to build a solid starting point for the task of interest, which can be performed based on just a few labeled samples, and with higher precision. The proposed approach was compared with (and outperformed, also from a generalization point of view) two common setups, namely the network initialization with general-domain (ImageNet) features, and training the full model from scratch. Regarding possible future developments, an explainability analysis of the methodology could be carried out [30]; moreover, the subset of scribes whose identification was most difficult could be analyzed in order to determine the necessary adjustments to the model to extract useful features even in complex cases like this; then, a broader experimental setup could be investigated; finally, a solution to tackle the problem of multigraphism could be included in the methodology [31].

## Acknowledgments

The author is very grateful to the Physics Department of Sapienza University for carrying out the experiments on the INFN NVIDIA DGX-1 server, and to Professors Serena Ammirati, Donatella Firmani, Nikos Komodakis, Paolo Merialdo, and Simone Scardapane for the supervision.

## References


[1] P. A. Stokes, Modeling Medieval Handwriting: A New Approach to Digital Palaeography, in: J. C. Meister (Ed.), Digital Humanities 2012, University of Hamburg, Hamburg, 2012, pp. 382–385.

[2] P. A. Stokes, Digital Approaches to Paleography and Book History: Some Challenges, Present and Future, Frontiers in Digital Humanities 2 (2015) 5. URL: https://www.frontiersin.org/article/10.3389/fdigh.2015.00005. doi:10.3389/fdigh.2015.00005.

[3] D. Stutzmann, C. Tensmeyer, V. Christlein, Writer identification and script classification: two tasks for a common understanding of cultural heritage, OpenX for Interdisciplinary Computational Manuscript Research (2018) 12–15.

[4] A. Abdalhaleem, B. K. Barakat, J. El-Sana, Case study: Fine writing style classification using siamese neural network, in: 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), IEEE, 2018, pp. 62–66.

[5] M. Kassis, J. Nassour, J. El-Sana, Writing Style Invariant Deep Learning Model for Historical Manuscripts Alignment, arXiv preprint arXiv:1806.03987 (2018).

[6] A. Pirrone, M. B. Aimar, N. Journet, Papy-S-Net: A Siamese Network to match papyrus fragments, in: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, 2019, pp. 78–83.

[7] N. Cilia, C. De Stefano, F. Fontanella, C. Marrocco, M. Molinara, A. Scotto Di Freca, An end-to-end deep learning system for medieval writer identification, Pattern Recognition Letters 129 (2020) 137–143. URL: https://www.sciencedirect.com/science/article/pii/S0167865519303460. doi:https://doi.org/10.1016/j.patrec.2019.11.025.

[8] Biblioteca Apostolica Vaticana, Website of the Biblioteca Apostolica Vaticana, https://www.vaticanlibrary.va/en/, 2021.

[9] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. E. Hinton, Big Self-Supervised Models are Strong Semi-Supervised Learners, Advances in Neural Information Processing Systems 33 (2020) 22243–22255.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, arXiv preprint arXiv:2006.07733 (2020).

[11] P. Bachman, R. D. Hjelm, W. Buchwalter, Learning Representations by Maximizing Mutual Information Across Views, Advances in Neural Information Processing Systems 32 (2019) 15535–15545.

[12] W. Falcon, K. Cho, A framework for contrastive self-supervised learning and designing a new approach, arXiv preprint arXiv:2009.00104 (2020).

[13] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[14] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[15] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, M. Cord, Learning representations by predicting bags of visual words, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6928–6938.

[16] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9912–9924. URL: https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.

[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, arXiv preprint arXiv:2103.03230 (2021).

[18] S. Gidaris, A. Bursuc, G. Puy, N. Komodakis, M. Cord, P. Pérez, Online bag-of-visual-words generation for unsupervised representation learning, arXiv preprint arXiv:2012.11552 (2020).

[19] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification., Journal of machine learning research 10 (2009).

[20] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.

[22] P. Cherubini, A. Pratesi, Paleografia latina. Tavole, volume 1, Scuola Vaticana di Paleografia, Diplomatica e Archivistica, 2004.

[23] P. Cherubini, A. Pratesi, Paleografia latina. L'avventura grafica del mondo occidentale, Littera Antiqua, 16, Scuola Vaticana di Paleografia, Diplomatica e Archivistica, Città del Vaticano, 2010.

[24] F. Coulson, R. Babcock, The Oxford Handbook of Latin Palaeography, Oxford University Press, USA, 2020.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[26] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).

[27] Olivier Moindrot, Triplet Loss and Online Triplet Mining in TensorFlow, https://omoindrot.github.io/triplet-loss, 2018.

[28] E. Shelhamer, J. Long, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 640–651. doi:10.1109/TPAMI.2016.2572683.

[29] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE., Journal of machine learning research 9 (2008).

[30] D. Basaj, W. Oleszkiewicz, I. Sieradzki, M. Górszczak, B. Rychalska, T. Trzcinski, B. Zieliński, Explaining Self-Supervised Image Representations with Visual Probing, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 592–598. URL: https://doi.org/10.24963/ijcai.2021/82. doi:10.24963/ijcai.2021/82, main Track.

[31] P. A. Stokes, Scribal Attribution across Multiple Scripts: A Digitally Aided Approach,

Speculum 92 (2017) S65–S85. URL: https://doi.org/10.1086/693968. doi:10.1086/693968.