Handwritten documents author verification based on the Siamese network

Vitaliy Kiselev[1], Dmitry Kropotov[2], Nataliia Pronina[3]

[1] TSU, 36 Lenin Avenue, Tomsk, 634050, Russia - kv-uliss@mail.ru
[2] HSE University, 11 Pokrovsky Bulvar, Moscow, 109028, Russia - dmitry.kropotov@gmail.com
[3] M. V. Lomonosov MSU, GSP-1, 1 Leninskiye Gory, Moscow, 119991, Russia - natalka-pronina@mail.ru

Keywords: Siamese Neural Networks, Author identification and verification, Manuscript, Handwritten text.

Abstract

The paper presents a method for verifying the handwriting of a certain author in a corpus of handwritten documents based on a small number of examples, and proposes an algorithm for data preprocessing.
The use of Siamese neural network is proposed to compare and analyze unique characteristics of handwriting, writing style. This way of training allows to obtain powerful discriminative image features, embeddings, on the basis of which it is possible to make a qualitative classification of the author.
The proposed approach was applied to the task of verification of possible autographs of Zhukovsky among manuscripts of unknown authors. The approach was also applied to the classification task on a fully labeled IAM dataset.

1. Introduction

Deep neural networks have long demonstrated high performance on image classification tasks. The most striking example - in 2015, the ResNet (He et al., 2015) neural network with 152 layers showed fewer errors in the ImageNet competition compared to manual partitioning. However, the more parameters a model has, the higher the risk of overtraining and the greater the need for more data. This problem is most acute in the task of verifying the handwriting of a certain author due to the large number of classes and the small number of representatives of each class.

In works on the identification of the author of a handwritten text, the task is often reduced to the analysis of minimal units of writing, graphemes (Bensefia et al., 2002, Bensefia et al., 2005, Koch et al., 2015). Taking into account the specificity of the provided photos, their low quality, it will not be possible to single out graphemes. But this is not required: it is not necessary to recognize all graphemes of the handwritten text to establish authorship. It is hypothesized that the author can be recognized even in the case when it is impossible to read the written text. In ordinary life this is done intuitively by the general style of writing: by the slant and size of the text, by unique strokes, by spaces between lines, by the mutual arrangement of lines, etc. Therefore, in this paper, handwriting style will be considered as an image pattern, and identification will be made by a fragment with several lines.

The relevance of the task is conditioned by the fact that the machine search of texts with handwriting of a certain author in large databases of raster images of handwritten documents, on the one hand, considerably expands the possibilities of researchers, in particular, literary critics or historians, to identify texts of a certain person in archival collections; on the other hand, it provides archival workers with a convenient tool for automatic classification: the register of handwriting correspondences identified by the program can serve as a basis for the description of a manuscript.

2. Siamese neural networks

The main data problem, data scarcity, is supposed to be solved by using a one-shot learning algorithm — *Siamese neural network*. Siamese networks were first introduced in the early 1990s by Bromley and LeCun to solve the (Bromley et al., 1993b) signature verification problem. It is a type of deep learning neural network that utilizes two or more identical subnets with the same architecture. They also use the same parameters for training.

Siamese networks are particularly useful in the case of classification with a large number of classes and with a small number of objects of each class. In such cases, there are not enough examples of each class to train a deep convolutional neural network. In addition, if new classes were added, the network architecture would have to be changed and retrained. Instead, Siamese networks are trained on the task of binary classification of pairs of objects: whether the objects belong to the same class or not. This makes them particularly useful in tasks that require flexibility and efficiency with a limited data set.

This type of networks allows to obtain feature vectors, embeddings, of two objects reflecting their semantic similarity or difference. Examples of applications for Siamese networks are: signature verification (Bromley et al., 1993a), face recognition (Solomon et al., 2023), sentence paraphrase identification (Yin and Schütze, 2015).

The class labels are obtained after training a simple classifier on embeddings, in this paper a two-layer full-link neural network will be used. A similar method for character classification was used in (Koch et al., 2015).

The two most popular loss functions for training Siamese neural networks are– contrastive loss function (Chopra et al., 2005, Hadsell et al., 2006) and triplet loss function (Schroff et al., 2015).

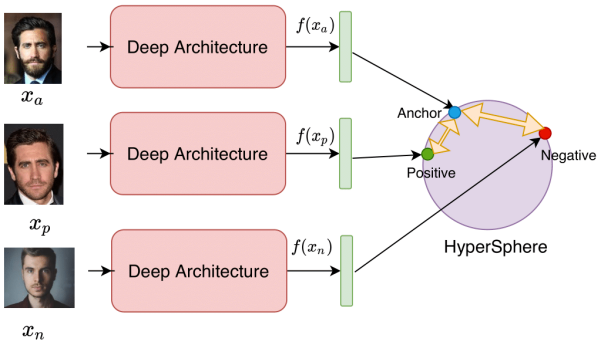

Figure 1. Training the siamese network with triplet loss (source: https://pyimagesearch.com/2023/03/06/triplet-loss-with-keras-and-tensorflow/)

2.1 Contrastive loss function

Contrastive loss uses a pair of objects, they can be of both positive and negative class.

$$L(x,y) = (1 - Z(x,y))\ d^2(x,y) + \\ +\ Z(x,y)\ max(0,\ margin - d^2(x,y)) \quad (1)$$

$$Z(x,y) = \begin{cases} 0 \mid x,y \text{ from the same class} \\ 1 \mid x,y \text{ from different classes} \end{cases} \quad (2)$$

$$d(x,y) = ||x - y||_p \quad (3)$$

Objects of the same class are penalized to minimize the distance between them, objects from different classes are penalized if the distance is less than the $margin$.

2.2 Triplet loss function

An improvement of contrastive loss is triplet loss. Unlike contrastive loss, it uses three objects: an object of the considered class (anchor), with which the comparison will be made, and two other objects: one belonging to the same class (positive) and one belonging to the opposite class (negative).

$$L(a,p,n) = max\{d(a,p) - d(a,n) + margin,\ 0\} \quad (4)$$

The function tends to bring objects of the same class closer and increase the distance between objects of different classes. Also, the function does not penalize if the required ratio of distances between three objects has already been reached. $margin$ — a pre-defined parameter indicating how much distance difference to penalize for.

When training a model with triplet loss, fewer samples are required for convergence because the network is updated simultaneously using both similar and dissimilar samples. Therefore, this loss function will be used in this paper.

3. Author verification task

The manuscripts of V.A. Zhukovsky, the great Russian poet, are kept in many archives in Russia and abroad. Meanwhile, archival inventories and catalogs rarely reach exhaustive completeness in terms of revealing the content of individual storage units. Convolutes, collections of manuscript documents described as a whole, but including individual texts, including those by different authors, represent a big "dark zone" here. The digitization of manuscripts, which is carried out in these archives with varying degrees of intensity, makes it possible to apply the developed software methods of attribution to the analysis of their content in the foreseeable future. A more complete description of the context of this problem is given in the article (Kiselev V.S. et al., 2023).

3.1 Problem statement

In technical terms, the conditions of the task are as follows: a small collection of Zhukovsky’s handwritings, his *autographs* (fig. 2) of different periods of his life — early, mature and late, of different degrees of neatness — perfect, ordinary, sloppy. There are 25 images in total.

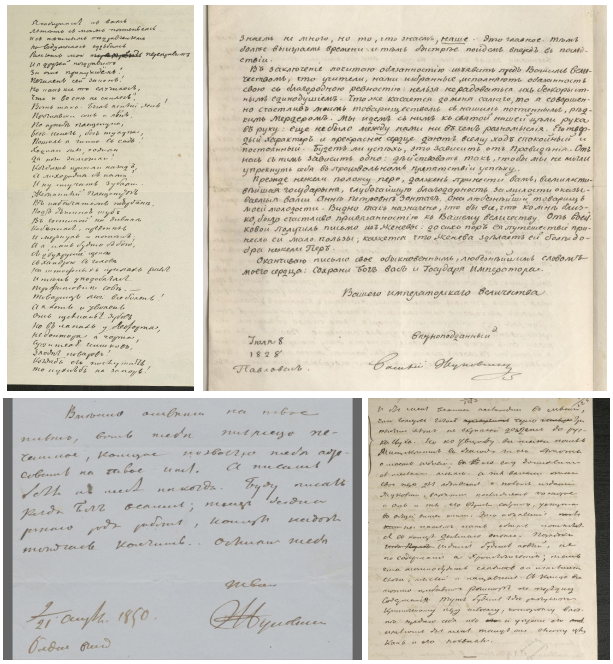

Figure 2. Autographs

Also given is a set of images of a selection of manuscript documents of unknown authors, *convolut* (fig. 3), 222 images in total. It is required to identify the documents in the collection of images that are likely to belong to Zhukovsky’s hand.

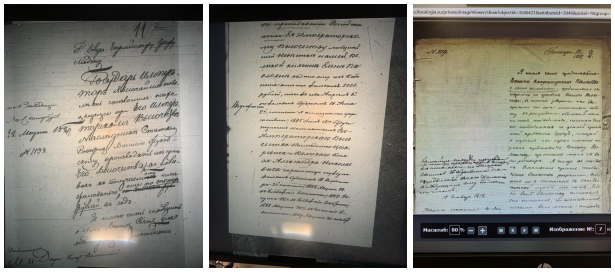

Figure 3. Convolutes

Formal problem formulation: for each convolute it is required to determine the probability that this document is an autograph of Zhukovsky. Then, according to the threshold probability value, the user will be able to select the top documents for manual expert verification.

Features of the provided data:

1. Lack of labeled data. Only 25 objects have an author.
2. Autographs and convolutes have different backgrounds.
3. Handwritings have different scales.
4. Convolutes are digitized much worse than autographs.
5. The images from the convolutes are distorted in the form of moiré patterns because they were taken from a screen, and glare is possible.

3.2 Data preprocessing

Enhance image enlargement with a photo enhancer (Picsart). To prevent overtraining to the background, we binarize the images using the DocEnTr (Souibgui et al., 2022) transformer. Figure 4 shows an example of binarizing autographs and images from convolutes.

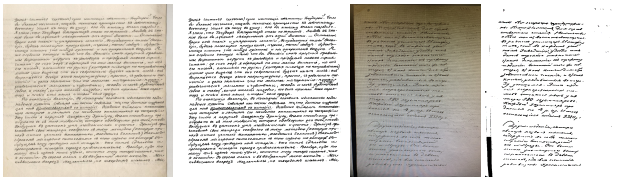

Figure 4. Binarization

Convolutes are taken at approximately the same scale and the size of handwriting does not differ much, which cannot be said about autographs. Therefore, for all autographs we will manually resize them so that a fragment of 300 pixels in height will contain 7-8 lines.

3.3 Training and validation sample

It is assumed that there are no more than 2% of Zhukovsky's autographs among the convolutes. Therefore, we take all convolutes as the *negative* class and Zhukovsky's autographs — as the *positive* class.

To prevent data leakage to the validation sample, each photograph was cut in half. If the height of the photograph was greater than the width, the top part was assigned to the training sample and the bottom part — to the validation sample. If the width of the photo is greater than the height, the left side — to the training sample and the right side — to the validation sample.

To increase the number of positive and negative objects and balance the classes in the samples, augmentation was performed by randomly cutting out a 300 x 300 pixel fragment and randomly transforming the perspective with a distortion rate of 0.3. To prevent white images from entering the samples, augmentation was performed so that the proportion of white pixels was no more than 95%. The number of positive instances increased 40 times, the number of negative instances — 5 times. Total: 1000 positive, 1110 negative — in the training sample, the same number in the validation sample.

3.4 Probability generation

A general method for obtaining the probability of belonging to a positive class:

1. Training the Siamese network and obtaining image embeddings.
2. Training embeddings classifier (loss function - cross entropy).
3. Obtaining the model prediction for the negative class.
4. Applying the softmax function, obtaining a number from a probability simplex.
5. Calibrating the probability.
6. Selecting the top of negative class objects with the highest probability. Hereinafter we will call such objects *«suspicious»* images.

3.5 Probability calibration

Suppose that a binary classifier produces some estimate of whether an object belongs to a positive class. Even if the estimate belongs to a probabilistic simplex, it may poorly estimate the real probability that the object belongs to a positive class. As a consequence, the response of the classifier is poorly interpretable.

Let us call a classifier *well calibrated* if for classifier a, object x of class y satisfies:

$$a(x) \approx P(y(x) = 1) \tag{5}$$

That is, if the classifier gave a group of objects a score of 0.9, then in the case of a good calibration, this group will contain about 90% of positive class objects. In reality, the algorithm is unlikely to give an identical score to a large group of objects, so all objects are divided into groups according to the probability value, *bins*.

To evaluate the calibration of the classifier, a *calibration curve* is constructed: the average predicted probability in each bin is plotted on the x-axis, while the y-axis shows the fraction of objects in each bin whose class is positive. Thus, for a well-calibrated classifier, the calibration curve is a straight line.

3.6 Histogram Binning

We apply one of the simplest and most universal calibration methods — the nonparametric Histogram Binning (Zadrozny and Elkan, 2001, Guo et al., 2017) method. It solves the problem of optimization by θ parameters:

$$\theta = \arg\min_{\theta} \sum_{m=1}^{M} \sum_{i=1}^{n} [a_i \in B_m](y_i - \theta_m)^2 \tag{6}$$

where a_i — classifier score
y_i — true class of i-object
$B_1, \ldots, B_M$ — bins

Analytical solution of the problem: θ_m corresponds to the average value of the positive class probability estimates falling into B_m. Usually bins of equal width are used in the method. After solving the problem, if a grade falls into the i-th bin, it is replaced by the corresponding value of θ_i.

Cons of the method:

1. There is a hyperparameter – the number of bins.
2. The probability transformation is not continuous.
3. If bins of equal width are used, some bins may contain a small number of objects.

3.7 Solution method

Let us apply the idea of training Siamese network on binary images. Let’s take the pre-trained ResNet18. The last layer has a dimensionality of 1000 on the output, so the Siamese network will learn 1000-dimensional image embedding. We will train the last two layers, 513000 trained parameters. We will build a dataset of 8000 image triples of the training sample, and a dataset of 2000 triples of the validation sample: as anchor and positive we will take random elements of the positive class, as negative— random element of the negative class.

The graphs 5 show the triplet loss and accuracy on training. Accuracy is considered as the fraction of triplets for which the Euclidean distance between the embeddings of anchor and positive is smaller than that between anchor and negative. The training accuracy is 99%, and the validation accuracy is 90.75%.

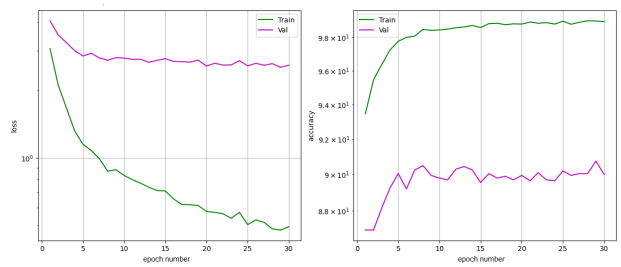

Figure 5. Triplet loss and accuracy

We train a classifier on the obtained embeddings: a two-layer fully connected neural network with 513,538 parameters. The training accuracy is 100%, and the validation accuracy is 97.63% (Fig. 6).

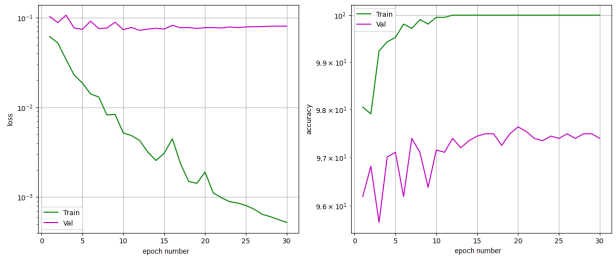

Figure 6. Loss and accuracy

The error matrices (Fig. 7) show that the model has no errors on the training sample, while on validation — 3% errors.

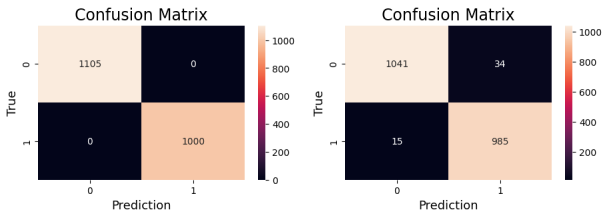

Figure 7. Confusion matrices on training and validation samples

Let us plot the calibration curves with 20 bins. The 8 plot shows that the curve is far from a straight line on the validation sample. After calibrating the probabilities using the Histogram Binning method, the classifier became well calibrated.

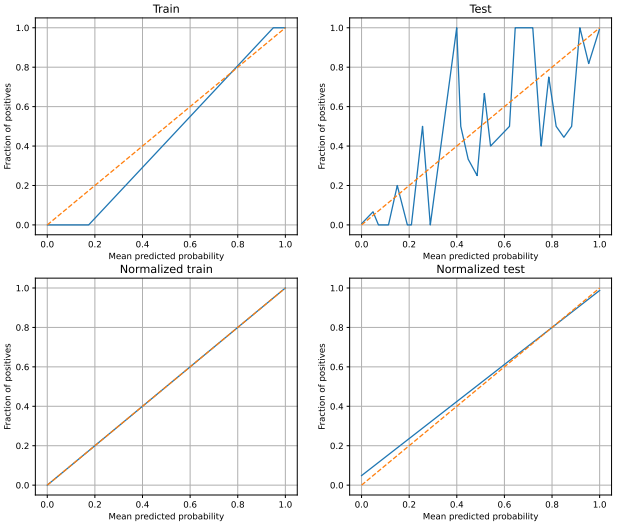

Figure 8. Histogram Binning

Figure 8 shows the probability distributions within the positive and negative class before and after calibration. The distributions turned out to be strongly skewed to the edges, there are pronounced outliers. Before calibration, the classifier suffered from «overconfidence» of prediction.

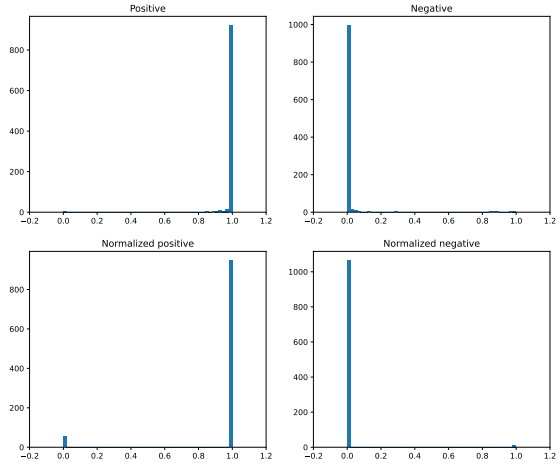

Figure 9. Probability distributions

«Suspicious» images also turned out to be of good quality (Figure 10).

Figure 10. Top images with maximum probability

A similar experiment was conducted for embeddings of the Siamese network of lower dimensionality, 128. The quality was slightly worse: 90.1% - accuracy of the Siamese network, 96.3% - accuracy of embeddings classification, the same images were classified as suspicious.

3.8 Result

The constructed model takes a fragment of an image as input, but we need to obtain the probability for the whole image. Let’s apply a simple idea: cut 10 random fragments of 300 x 300 pixels from the image, and for each fragment we get a prediction of the model. We will consider the maximum of 10 predictions as the final probability of the image.

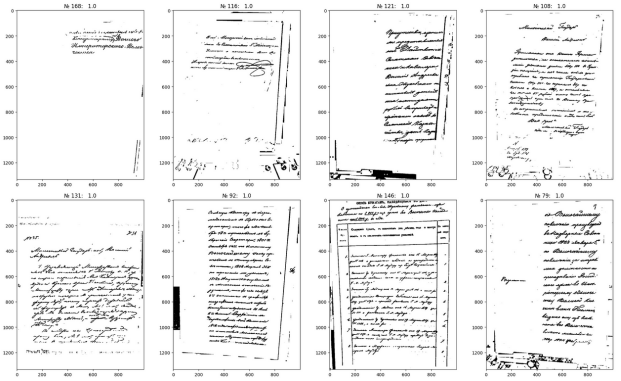

Figure 11. «Suspicious» images

4. Authors identification task

Since there is too little accurate labeling of the authors’ handwriting in the previous task, we cannot objectively assess the quality of the result.

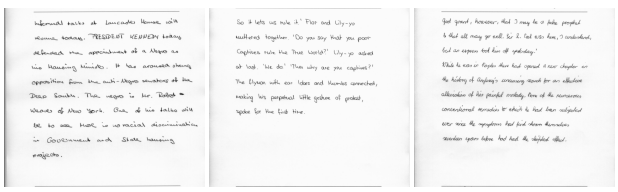

Figure 12. IAM dataset

Let us test the proposed approach on a similar handwriting classification task on the IAM (U. Marti, 2002) corpus. It consists of handwritten English sentences based on the Lancaster-Oslo/Bergen (Stig et al., 1978) corpus. IAM contains 1539 mappings with 657 authors. Due to its public availability, flexible structure, and large number of writers, IAM data is widely used for identification, verification of Latin writers, and handwriting recognition.

The experiment will be conducted on a shortened IAM, as there is one example for the vast majority of authors. We will select only those authors with at least five objects. Thus, the shortened IAM contains 397 pictures with 39 authors. Let’s crop the upper and lower part of the image with printed text and author’s signature so that they do not affect the prediction. Examples of handwriting images of different authors are shown in Figure 12.

A Siamese network with embedding dimension of 1000 was trained on a set of 1000 triplets and achieved an accuracy of 100% on validation (Figure 13). The accuracy of the two-layer classifier on 39 classes is — 98.75% (Fig. 14).

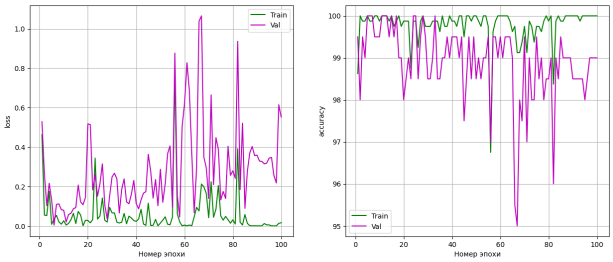

Figure 13. Triplet loss and accuracy

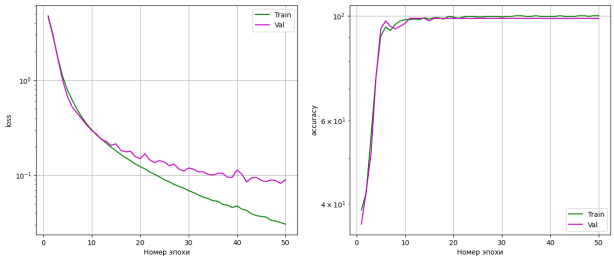

Figure 14. Loss and accuracy

This high accuracy compared to the result on the convolutes is explained by the better quality of the IAM dataset: it does not have the problem of different background, writing scale and text location.

5. Conclusion

This paper proposes an approach to verify the handwriting of a certain author in a corpus of historical documents based on a small number of samples. Experiments confirming its effectiveness were performed. In the future, the proposed method can be improved by switching from classification of an image fragment with several lines to classification of a single line of handwritten text, which will help to significantly increase the sample and eliminate the problem of different handwriting scale.

6. Acknowledgements

This work was supported by the Russian Science Foundation, project no. 22-68-00066.

References

Bensefia, A., Nosary, A., Paquet, T., Heutte, L., 2002. Writer identification by writer's invariants. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, IEEE, 274–279.

Bensefia, A., Paquet, T., Heutte, L., 2005. A writer identification and verification system. *Pattern Recognition Letters*, 26(13), 2080–2092.

Bromley, J., Bentz, J., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Sackinger, E., Shah, R., 1993a. Signature Verification using a "Siamese" Time Delay Neural Network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7, 25.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1993b. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, IEEE, 539–546.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q., 2017. On calibration of modern neural networks. D. Precup, Y. W. Teh (eds), *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 70, PMLR, 1321–1330.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 2, IEEE, 1735–1742.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Kiselev V.S., Lebedeva O.B., T. E. et al., 2023. The problem of machine identification of texts with the handwriting of a specific author as part of large databases of raster images of handwritten documents (based on the experience of identifying letters from Vasily Zhukovsky in office convolutes of the Russian State Historical Archive). *Imagologiya i komparativistika – Imagology and Comparative Studies*, 20, 247–262.

Koch, G., Zemel, R., Salakhutdinov, R. et al., 2015. Siamese neural networks for one-shot image recognition. *ICML deep learning workshop*, 2, Lille.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE.

Solomon, E., Woubie, A., Emiru, E. S., 2023. Deep Learning Based Face Recognition Method using Siamese Network. *arXiv preprint arXiv:2312.14001*.

Souibgui, M., Biswas, S., Jemni, S., Kessentini, Y., Fornes, A., Llados, J., Pal, U., 2022. Docentr: An end-to-end document image enhancement transformer. *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 1699–1705.

Stig, J., Leech, G. N., Goodluck, H., 1978. Manual of information to accompany the lancaster-oslo: Bergen corpus of british english, for use with digital computers. *(No Title)*.

U. Marti, H. B., 2002. The iam-database: An english sentence database for off-line handwriting recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5, IEEE, 39 – 46.

Yin, W., Schütze, H., 2015. Convolutional neural network for paraphrase identification. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 901–911.

Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *Icml*, 1, 609–616.