

МЕТОДЫ ПРЕДОБРАБОТКИ ОЦИФРОВАННЫХ РУКОПИСНЫХ ДОКУМЕНТОВ

Т. М. Татарникова*, А. А. Шихотов

*Санкт-Петербургский государственный университет аэрокосмического приборостроения** *tm-tatarn@yandex.ru*

Аннотация. Решается задача автоматизации анализа рукописных документов. Показано, что при решении подобных задач используются искусственные нейронные сети, способные к распознаванию образов после обучения на исходном наборе данных. При этом качество распознавания новых образов большей частью зависит от этапа предобработки оцифрованных рукописных документов. Рассмотрена частная задача предобработки — удаление линий клетки с изображения листа тетради. Проанализированы четыре метода фильтрации изображения с использованием библиотеки OpenCV языка Python. Выполнено обучение нейронной сети сверточной архитектуры распознаванию рукописных символов. Продемонстрирована работа обученной нейронной сети на документах, предобработанных разными алгоритмами.

Ключевые слова: распознавание образов, рукописный документ, нейронные сети, шум, предобработка, качество распознавания

Ссылка для цитирования: Татарникова Т. М., Шихотов А. А. Методы предобработки оцифрованных рукописных документов // Изв. вузов. Приборостроение. 2025. Т. 68, № 6. С. 494–499. DOI: 10.17586/0021-3454-2 025-68-6-494-499.

METHODS OF PREPROCESSING OF DIGITIZED HANDWRITTEN DOCUMENTS

T. M. Tatarnikova*, A. A. Shihotov

*St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia** *tm-tatarn@yandex.ru*

Abstract. The problem of automating the analysis of handwritten documents is solved. It is shown that artificial neural networks capable of recognizing images after training on the original data set are used to solve such problems. At the same time, the quality of recognizing new images largely depends on the stage of pre-processing of digitized handwritten documents. A particular preprocessing problem is considered - removing cell lines from an image of a notebook sheet. Four methods of image filtering are analyzed using the OpenCV library of the Python language. A neural network of convolutional architecture is trained to recognize handwritten characters. The work of the trained neural network on documents preprocessed by different algorithms is demonstrated.

Keywords: image recognition, handwritten document, neural networks, noise, preprocessing, recognition quality

For citation: Tatarnikova T. M., Shihotov A. A. Methods of preprocessing of digitized handwritten documents. *Journal of Instrument Engineering*. 2025. Vol. 68, N 6. P. 494–499 (in Russian). DOI: 10.17586/0021-3454-2025-68-6-494-499.

Введение. Решение задачи автоматизированной обработки рукописных документов требуется, например, при оцифровке архивов, идентификации личности, сортировке почтовых отправлений [1–3]. Разнообразие стилей письма, наличие шумов и различной четкости изображений делают эту задачу нетривиальной [4].

В общем случае автоматизация обработки рукописных документов относится к методам распознавания образов [5], в которой этап предобработки является самым трудоемким — именно от него зависит качество решения задачи [6, 7].

Одной из частных проблем в обработке оцифрованных рукописных документов является необходимость удаления рисунка клетки с изображения тетрадного листа. Линии клетки могут мешать распознаванию текста, создавая визуальный шум и тем самым снижая точность извлеченной информации.

В статье рассматриваются несколько методов удаления линий клетки с изображения листа тетради с использованием библиотеки OpenCV языка Python и дальнейшим распознаванием рукописного текста обученной нейронной сетью [8]. Цель — выбрать метод предобработки, дающий наименьшую ошибку распознавания.

Применяемые методы предобработки. Первый метод использует алгоритм обнаружения границ Canny. В зависимости от требований к решаемой задаче возможно выделить большинство границ клетки или только те границы, которые имеют наибольшую яркость. Это становится возможным при изменении параметров порогового значения гистерезиса [9]. В результате применения данного алгоритма возможно на некоторых изображениях добиться „игнорирования“ клетки и выделения полезной информации, но вместе с клеткой удаляется немалая часть рукописного текста с изображения, из-за чего данный способ может быть эффективно использован только в случае четкого изображения и жирных линий рукописного текста.

Второй метод в дополнение к алгоритму Canny использует алгоритм преобразования Хафа для нахождения прямых линий клетки. Этот метод полезен при наличии четких контуров линий, но при его использовании могут быть потеряны и необходимые данные. После нахождения линий создается фильтр, который заменяет пиксели линии на смежные пиксели фона.

Преобразование Хафа — медленное, поскольку решения принимаются по каждому пикселу [10]. Существует более вычислительно быстрая версия данного алгоритма — вероятностное преобразование Хафа, при котором требуется только определенное количество пикселей для принятия решения. Это ускоряет вычисления, но за счет потери некоторой точности распознавания.

Стоит отметить, что преобразование Хафа применимо для отсканированных изображений или изображений, полученных фотосъемкой под прямым углом на идеально ровной поверхности. В эксперименте по предобработке оцифрованного рукописного документа были найдены лишь некоторые линии клетки, которые относятся к визуальному шуму.

Эксперимент по предобработке оцифрованного рукописного документа показал, что вероятностное преобразование Хафа может служить лишь дополнением к другим инструментам удаления шума из изображения. Также при изменении масштаба возможно захватывание большего числа клеток, но при этом обязательно будут захвачены линии у полезных данных на изображении, что приведет к потере качества изображения. По этим причинам далее данный метод не использовался.

Третий метод использует простейший алгоритм фильтрации изображения по цветовым характеристикам Color filter [5]. Так, возможно выделение необходимого диапазона цвета чернил, которыми был написан текст. Для упрощения задачи лучше конвертировать RGB-изображение в пространство HSV, где необходимо создать маску по цвету, насыщенности и яркости, а после применить ее к исходному изображению. В результате алгоритм фильтрации напрямую выделяет рукописный текст из изображения. Полученные данные имеют погрешности из-за узкого диапазона цветового спектра, которые устранить возможно, но при этом неотфильтрованные изображения клетки добавляют шум в изображение. Также имеется существенный недостаток алгоритма: его невозможно применить универсально ко всем данным: в одном случае насыщенность цвета линий клетки может совпадать с насыщенностью чернил, а в другом — цвет чернил будет отличаться от примера, на котором был создан фильтр.

Четвертый метод использует комбинацию алгоритмов. Примером может служить комбинация двух алгоритмов из библиотеки OpenCV. В первую очередь выполняется бинаризация изображения с использованием алгоритма Adaptive Threshold, а затем применяется морфологическая функция Morph Close для окончательного удаления шума на изображении [8, 9]. Как показывает эксперимент, комбинация алгоритмов позволяет лучше справиться с поставленной задачей: линии клетки практически удалены, оставшееся их изображение слабо влияет на качество распознавания. Попытки изменения регулирующих параметров приводят к ухудшению качества изображения, поэтому рекомендуется использовать методы с формированием маски клетки, которые будут исключать дальнейшее воздействие на полезное изображение во избежание ухудшения результата.

Анализ результатов предобработки. После анализа полученных результатов было замечено, что наивысшее качество данных в тестовом изображении обеспечил алгоритм нахождения границ Canny, однако изображения цифр (рис. 1) состоят из тонких линий, они не годились для дальнейшей обработки. Поэтому была применена морфологическая функция Morph Close, которая дала достаточно качественный результат.

Сравнивая три метода (использование алгоритма Canny в комбинации с Morph Close, алгоритма фильтрации по цветовым характеристикам Color filter в комбинации с Threshold, Adaptive Threshold в комбинации с Morph Close), можно подытожить, что все они могут быть использованы в предобработке. В частности, в результате работы первого метода на изображении остаются тонкие и прерывистые линии цифр, второго — толстые и немного смазанные линии цифр, третьего — тонкие и размытые линии цифр. Выбор метода зависит от тех данных, на которых производилось обучение нейронной сети. При достаточно качественной обучающей выборке разница в результате использования методов, как в данном примере, будет несущественной.



Рис. 1

Затем в результате применения функций Find Contours и Bounding Rect библиотеки OpenCV и простой фильтрации по размеру удалось извлечь образы из изображения прямоугольниками, как показано на рис. 1.

Далее полученные изображения цифр были преобразованы к размеру 28×28 пикселей в соответствии с обучающей выборкой для нейронной сети.

Оценка результатов распознавания образов. Для распознавания цифр была обучена сверточная нейронная сеть [11]. Известно, что нейронные сети данного типа привлекаются к решению подобных задач компьютерного зрения: выделяют характерные черты объектов на изображении и выполняют классификацию [12–14]. Также сверточные нейросети легко обучаются, что позволяет не прибегать к использованию больших объемов вычислительных ресурсов [15–17].

Архитектура нейронной сети приведена на рис. 2, функция ошибки — категориальная кроссэнтропия, оптимизатор — Adam. Качество модели оценивалось по метрике точности. Предобработка выполнялась на оцифрованном изображении, фото которого показано на рис. 3.

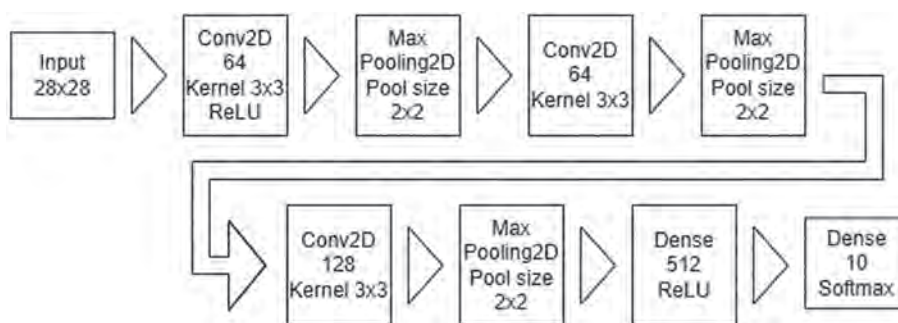


Рис. 2

Тестовое фото пропусклось через фильтры, результаты фильтрации тремя методами показаны на рис. 4 (*а* — Canny в комбинации с Morph Close; *б* — фильтр по цветовым характеристикам Color filter в комбинации с Threshold; *в* — Adaptive Threshold в комбинации с Morph Close). Первый и третий методы выделяются низким количеством шума, в отличие от фильтрации по цветовым характеристикам. Также можно отметить, что фильтр Canny в данном случае не нуждается в использовании преобразования Хафа, поскольку качество тестового изображения позволяет фильтру Canny самостоятельно отличить границы цифр от клетки.

Далее цифры помещались в прямоугольные рамки и их изображения передавались на вход нейронной сети. Результаты распознавания в виде метрик accuracy, precision, recall, F1 и F-beta приведены в таблице.

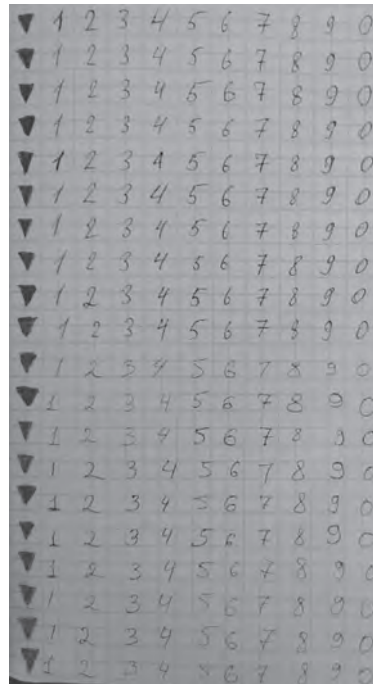


Рис. 3

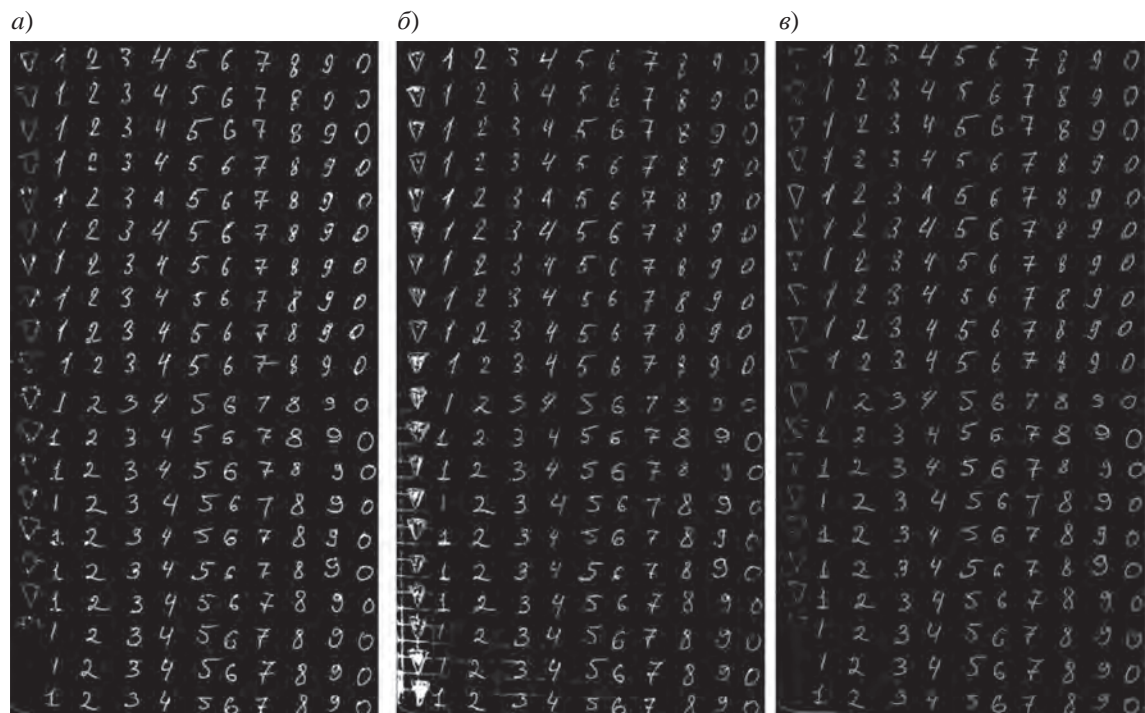


Рис. 4

Метрика	Алгоритм		
	Canny + Morph Close	Color filter + Threshold	Adaptive Threshold + Morph Close
Accuracy, %	86,98	80,86	81,59
Precision, %	88,19	83,85	84,16
Recall, %	86,81	81,94	82,47
F1, %	86,83	81,58	82,31
F-beta, %	87,48	82,59	83,22

Так, наилучшим для распознавания цифр на клетчатой бумаге методом по метрикам оказался Canny в комбинации с Morph Close. Однако в процессе эксперимента выяснилось, что он наиболее восприимчив к потере качества изображения. В то же время самый стабильный результат показал алгоритм фильтрации Adaptive Threshold в комбинации с Morph Close. Самым требовательным оказался Color filter, для которого необходимо было заново подбирать параметры алгоритма согласно тестовому изображению, но даже после настройки результат фильтрации нельзя назвать идеальным.

Заключение. Сравнивая подходы к обработке изображений, целью которой является сохранение качества рукописного текста и удаление шума, можно выделить несколько важных моментов:

— стоит учитывать, насколько можно пожертвовать качеством текста ради удаления значимого объема шума. Параметры каждого из рассмотренных алгоритмов можно регулировать, что позволяет балансировать между требуемым качеством рукописного текста и наличием шума;

— стоит выделять характерные черты рукописного текста в документе, такие как цвет и насыщенность букв, а также уровень освещенности документа. При обработке единичных образцов данные характеристики могут послужить созданию простейшего и эффективного фильтра;

— в некоторых случаях результат использования алгоритмов по отдельности может быть наилучшим, но целесообразно применять комбинацию нескольких подходов, которые в совокупности представляют многоэтапный процесс обработки, учитывающий множество особенностей оцифрованных рукописных документов.

СПИСОК ЛИТЕРАТУРЫ

1. Кривошеев А. В. Сопряжение мультиагентных и нейросетевых технологий в программной системе распознавания образов для роботизированного документооборота // Вестн. Рязанского государственного радиотехнического университета. 2024. № 87. С. 54–61. DOI: 10.21667/1995-4565-2024-87-54-61.
2. Ferrer M. A., Diaz M., Carmona-Duarte C., Morales A. A behavioral handwriting model for static and dynamic signature synthesis // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017. Vol. 39, N 6. P. 1041–1053.
3. Калько А. И. Система идентификации рукописных текстов как объект защиты и обработки массового объема документов // 9-я Междунар. науч.-практ. Конф. „BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня“. Минск, Республика Беларусь, 17–18 мая 2023 года. С. 276–284.
4. Кучуганов А. В., Лапинская Г. В. Распознавание рукописных текстов. Ижевск: Мир, 2006. 514 с.
5. Шапиро Л., Стокман Дж. Компьютерное зрение. М.: БИНОМ. Лаборатория знаний, 2015. 763 с.
6. Бреусова Е. И. К проблеме распознавания рукописного текста // Вестн. Челябинского государственного педагогического университета. 2017. № 1. С. 88–92.
7. Ахунджанов У. Ю., Старовойтов В. В. Предварительная обработка изображений рукописных подписей для последующего распознавания // Системный анализ. 2022. № 2. С. 4–9. DOI: 10.21122/2309-4923-2022-2-4-9.
8. OpenCV library. O'Reilly Media Inc, 2008. 6 p.
9. Горитов А. Н. Предварительная обработка изображений в системах технического зрения // Доклады ТУСУР. 2018. Т. 21, № 4-1. С. 53–58.
10. Семенов А. Н. Обнаружение радиолокационных целей с помощью преобразования Хафа // Наука и Образование: научное издание МГТУ им. Н.Э. Баумана. 2014. № 12. С. 619–632. DOI: 10.7463/1214.0738733.
11. Wu B., Wan A., Yue X. and Keutzer K. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud // 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, QLD, Australia, 2018. P. 1887–1893. DOI: 10.1109/ICRA.2018.8462926.
12. Скрыпников А. В., Денисенко В. В., Хитров Е. Г., Евтеева К. С., Савченко И. И. Распознавание рукописного текста с использованием нейронных сетей // Современные наукоемкие технологии. 2021. № 6-1. С. 91–95. DOI: <https://doi.org/10.17513/snt.38703>.
13. Makridakis S., Spiliotis E., Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward // PloS one. 2018. Vol. 13. P. 1–20.
14. Taherdoost H., Sahibuddin S., Jalaliyoon N. Exploratory Factor Analysis; Concepts and Theory // Advances in applied and pure mathematics. 2022. Vol. 27. P. 375–382.

15. He K., Zhang X., Ren S. and Sun J. Deep residual learning for image recognition // 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 27–30 June 2016. P. 770–778. DOI: 10.1109/CVPR33180.2016.
16. Татарникова Т. М., Бимбетов Ф., Богданов П. Ю. Выявление аномалий сетевого трафика методом глубокого обучения // Изв. СПбГЭТУ ЛЭТИ. 2021. № 4. С. 36–41.
17. Suprun A., Tatarnikova T., Sikarev I., Shmeleva A. Detection of malicious program for the android platform by deep training // Lecture Notes in Networks and Systems. 2022. Vol. 387. P. 31–38. DOI: 10.1007/978-3-030-93872-7_3.

СВЕДЕНИЯ ОБ АВТОРАХ

- Татьяна Михайловна Татарникова** — д-р техн. наук, профессор; Санкт-Петербургский государственный университет аэрокосмического приборостроения, кафедра прикладной информатики; E-mail: tm-tatarn@yandex.ru
- Алексей Алексеевич Шихотов** — магистр; Санкт-Петербургский государственный университет аэрокосмического приборостроения, кафедра прикладной информатики; E-mail: ashikhotov@bk.ru

Поступила в редакцию 21.02.25; одобрена после рецензирования 27.02.25; принята к публикации 25.04.25.

REFERENCES

1. Krivosheev A.V. *Vestnik of Ryazan State Radio Engineering University*, 2024, no. 87, pp. 54–61, DOI: 10.21667/1995-4565-2024-87-54-61. (in Russ.)
2. Ferrer M.A., Diaz M., Carmona-Duarte C., Morales A. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, no. 6(39), pp. 1041–1053.
3. Kalko A.I. *BIG DATA and Advanced Analytics. BIG DATA i analiz vysokogo urovnya* (BIG DATA and Advanced Analytics. BIG DATA and High-Level Analysis), Ninth International Scientific and Practical Conference, Minsk, Republic of Belarus, May 17–18, 2023, pp. 276–284. (in Russ.)
4. Kuchuganov A.V., Lapinskaya G.V. *Raspoznavaniye rukopisnykh tekstov* (Handwritten Text Recognition), Izhevsk, 2006, 514 p. (in Russ.)
5. Shapiro L., Stockman G. *Computer Vision*, Prentice Hall, 2001, 617 p.
6. Breusova E.I. *Bulletin of the Chelyabinsk State Pedagogical University*, 2017, no. 1, pp. 88–92. (in Russ.)
7. Akhundzhanov U.Yu., Starovoytov V.V. *System analysis*, 2022, no. 2, pp. 4–9, DOI: 10.21122/2309-4923-2022-2-4-9. (in Russ.)
8. *OpenCV library*, O'Reilly Media Inc, 2008, 6 p.
9. Goritov A.N. *Proceedings of the TUSUR University*, 2018, no. 4-1(21), pp. 53–58. (in Russ.)
10. Semenov A.N. *Science and Education: Scientific Publication of Bauman MSTU*, 2014, no. 12, pp. 619–632, DOI: 10.7463/1214.0738733. (in Russ.)
11. Wu B., Wan A., Yue X. and Keutzer K. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 1887–1893, DOI: 10.1109/ICRA.2018.8462926.
12. Skripnikov A.V., Denisenko V.V., Khitrov E.G., Evteeva K.S., Savchenko I.I. *Modern science-intensive technologies*, 2021, no. 6-1, pp. 91–95, DOI: https://doi.org/10.17513/snt.38703. (in Russ.)
13. Makridakis S., Spiliotis E., Assimakopoulos V. *PloS one*, 2018, vol. 13, pp. 1–20.
14. Taherdoost H., Sahibuddin S., Jalaliyoon N. *Advances in applied and pure mathematics*, 2022, vol. 27, pp. 375–382.
15. He K., Zhang X., Ren S. and Sun J. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27–30, 2016, pp. 770–778, DOI: 10.1109/CVPR33180.2016.
16. Tatarnikova T.M., Bimbetov F., Bogdanov P.Yu. *Izvestiya SPbGETU LETI*, 2021, no. 4, pp. 36–41. (in Russ.)
17. Suprun A., Tatarnikova T., Sikarev I., Shmeleva A. *Lecture Notes in Networks and Systems*, 2022, vol. 387, pp. 31–38, DOI: 10.1007/978-3-030-93872-7_3.

DATA ON AUTHORS

- Tatyana M. Tatarnikova** — Dr. Sci., Professor; St. Petersburg State University of Aerospace Instrumentation, Department of Applied Informatics; E-mail: tm-tatarn@yandex.ru
- Aleksey A. Shihotov** — M.Sc.; St. Petersburg State University of Aerospace Instrumentation, Department of Applied Informatics E-mail: ashikhotov@bk.ru

Received 21.02.25; approved after reviewing 27.02.25; accepted for publication 25.04.25.