Информационные процессы, Том 25, № 4, 2025, стр. 821–825.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

PHONE-SCAN: Набор изображений документов, оцифрованных и нормализованных с помощью смартфона

О. А. Славин, И. М. Янишевский

Федеральный исследовательский центр «Информатика и управление» Российской академии наук

Поступила в редакцию 01.10.2025 г. Принята 12.12.2025 г.

Аннотация—В работе описан датасет PHONE-SCAN для анализа русскоязычных документов формата A4, оцифрованных смартфонами в реальных условиях. Набор включает 451 образцов одиннадцати типов документов с различными уровнями шума и искажений, полученных разными пользователями и устройствами. К данным прилагаются результаты распознавания текста OCR Tesseract. Датасет может быть использован для разработки и тестирования алгоритмов классификации, нормализации и улучшения качества изображений документов.

КЛЮЧЕВЫЕ СЛОВА: датасет, изображение документа, документы A4, оцифровка изображений.

DOI: 10.53921/18195822_2025_25_4_821

1. ВВЕДЕНИЕ

Важнейшей частью машинного обучения и создания систем распознавания и обработки документов являются публичные наборы данных (датасеты). С одной стороны, имеется достаточно много опубликованных датасетов, содержащих изображения различных классов для различных задач, например:

1. для бинаризации: наборы данных платформы DIB [1];
2. для анализа мобильных фотографий документов, удостоверяющих личность: MIDV-500, MIDV-2019, MIDV-DM [2–4],
3. для типизации документов: Legacy Tobacco Document Library [5, 6],
4. для распознавания таблиц: TableBank [7].

В то же самое время, опубликованных датасетов изображений недостаточно, прежде всего, для анализа образов русскоязычных документов. Другим недостатком некоторых датасетов является недостаточное число зашумлённых образцов документов. Например, в датасете TableBank содержатся идеальные изображения.

2. ЦЕЛЬ СОЗДАНИЯ НОВОГО НАБОРА

При оцифровке с помощью камер мобильных устройств происходят искажения, вызванные аберрациями, бликами и отражениями внутри оптической системы. Также для камер характерны искажения изображений, такими как «цифровой шум». Еще один источник искажений – алгоритмы сжатия изображений, что справедливо и для сканов. Цифровой шум заметен на изображении в виде наложенной маски из пикселей случайного цвета и яркости. При съемке

камерой размещение документа может быть произвольным относительно плоскости сфокусированного изображения, что приводит к проективному искажению. Из-за изменения пропорций объектов в зависимости от ракурса съемки требуется предварительная проективная нормализация изображения.

Цель создания нового датасета состоит в оцифровке бумажных документов формата A4 неквалифицированными пользователями в условиях недостаточного освещения. В реальных потоках изображений документов иногда встречаются образы с сильными искажениями, такими как расфокусировка, нелинейные дисторсии и малая площадь образа документа в кадре. Это объясняется невнимательностью и усталостью пользователей, не придающих значения дальнейшей обработке оцифрованного документа.

Три пользователя использовали для оцифровки смартфоны Samsung S23 Ultra и Samsung S23. Съёмка проводилась в помещении при различных условиях освещения. При съёмке смартфоном использовалось приложение «Камера» в режиме «Сканировать». Этот режим позволяет полуавтоматически избавиться от проективных искажений и преобразовать образ документа к прямоугольному виду для дальнейшей обработки и распознавания текста. Такой способ позволяет получать качественные изображения документов, но не гарантирует отсутствия искажений.

3. СТРУКТУРА ДАТАСЕТА

В датасете содержатся образцы документов одиннадцати типов:

1. Договор аренды (t_1);
2. Акт приёма-передачи (t_2);
3. Заявление на отпуск (t_3);
4. Договор купли-продажи (t_4);
5. Авансовый отчёт (t_5);
6. Торговая накладная, ТОРГ-12 (t_6);
7. Налоговая справка, 2-НДФЛ (t_7);
8. Универсальный передаточный документ, УПД (t_8);
9. Доверенность (t_9);
10. Счёт оплаты (t_{10});
11. Устав организации (t_{11}).

Каждый из документов одного типа представлен пятью образцами. Набор из 11 документов был оцифрован 8 раз: один раз отсканирован и 6 раз оцифрован тремя пользователями с помощью двух смартфонов. Оцифровка смартфонами представлена шестью сеансами, оцифровка сканированием — одним сеансом. Также были отсканированы 11 документов без заполнения, составивших эталонный набор для возможного обучения.

Каждый документ представлен изображением с именем со следующей структурой:

<I>_<D>_<T>_<N>, где

1. I – тип оцифровки, одна цифра:
   (a) 0 (скан документа с заполнением);
   (b) 1 (фото документа с мобильного устройства с заполнением);
2. D – тип набора:
   (a) 0 – документ без заполнения, 1 – с заполнением (для сканов);
   (b) 1÷7 – номер сеанса оцифровки (для фото);
3. T – тип документа;

4. N – номер образца документа (от 1 до 5).

Каждое изображение было распознано с помощью OCR Tesseract 5.3.3 с параметрами `-l RUS -psm 1 -oem 1`.

Точнее, использовалось приложение, вызывающее API OCR Tesseract со следующими установками:

1. язык RUSSIAN;
2. Page segmentation modes – Automatic page segmentation with OSD;
3. OCR Engine modes – Neural nets LSTM engine only.

Приложение сохраняло результаты в текстовом виде в формате CSV в следующем виде:

W; X1; Y1; X2; Y1; X2; Y2; X1; Y2; NB; NP; NL; SW; NChars; SetChars.

Каждый символ представлен рамкой из двух точек числом альтернатив распознавания NAlts и самими альтернативами SetAlts. Каждая альтернатива состоит из кода символа в кодировке UTF-8 и оценки в диапазоне 0÷1.

Здесь:

1. W – слово в кодировке UTF-8;
2. N – число символов в слове;
3. X1;Y1; X2;Y1; X2;Y2; X1;Y2 – рамка слова;
4. NB, NP, NL – номер блока, номер параграфа в блоке и номер строки в параграфе;
5. SW – оценка слова в диапазоне 0÷1.

Указанные имена всех образов различаются, поэтому изображения и тестовые файлы могут быть сохранены в одной папке. В опубликованном датасете файлы сгруппированы по способу оцифровки и по типу документа в отдельных папках. Папки соответствуют наборам D_{00} (сканы документов без заполнения), D_{01} (сканы с заполнением), D_{11} – D_{17} (фото из 7 сеансов).

4. АПРОБАЦИЯ НАБОРА ДАННЫХ

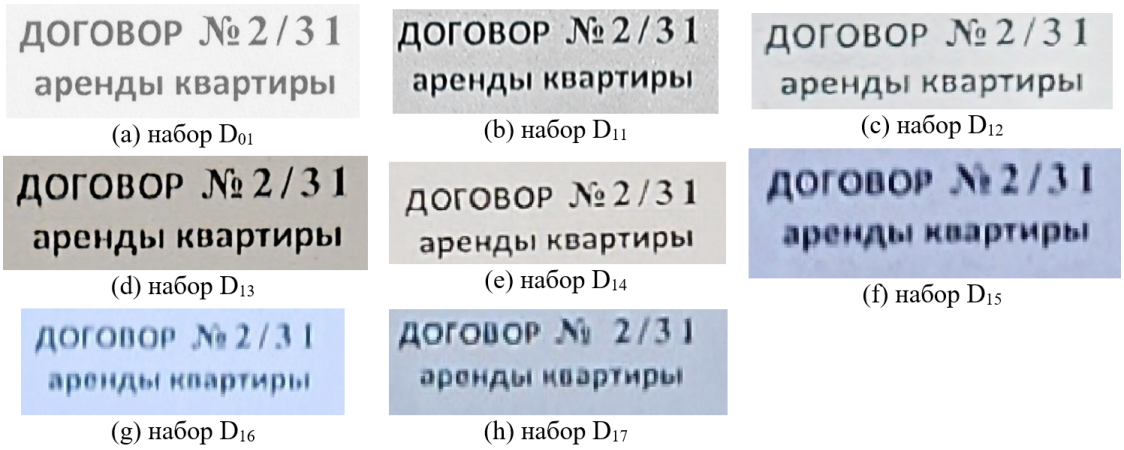
На рис. 1 приведены примеры фрагментов изображений документов из наборов D_{01} и D_{11} – D_{17}.

В табл. 1 приведены средние значения оценок распознавания отдельных слов, длиной не менее трех символов. Данные из табл. 1 иллюстрируют различия качества оцифровки в отдельных наборах. Оказалось, что набор D_{15} оцифрован с наибольшими искажениями.

Таблица 1. Средние значения оценок OCR Tesseract распознавания отдельных слов

D_{01}	D_{11}	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}
0,91	0,79	0,80	0,79	0,79	0,44	0,68	0,77

Предложенный датасет может быть использован как для анализа изображений, так и для анализа извлечённого текста. В работе [8] предложен алгоритм классификации типа изображений, апробированный на предложенном датасете. Алгоритм основа на модели «мешок слов» с частотами встречаемости и указанием принадлежности классам документов. Применяется сравнение слов с помощью модифицированного расстояния Левенштейна, описанного в [9]. Точность классификации варьируется от 1.0 (отсутствие ошибок) для наборов D_{01}, D_{11}, D_{12}, D_{13}, D_{14}, D_{15}, D_{16}, D_{17} до 0.4 в наихудшем случае для наборов D_{14}, D_{15}. В [8] отмечается,



(a) набор D_{01} (b) набор D_{11} (c) набор D_{12}

(d) набор D_{13} (e) набор D_{14} (f) набор D_{15}

(g) набор D_{16} (h) набор D_{17}

Рис. 1. Примеры фрагментов изображений различного качества из наборов D_{01} и $D_{11} - D_{17}$

что для документов типа Устав (t_{11}) модель «мешка слов» не позволяет классифицировать тип t_{11}, поскольку слово «устав» встречается в документах других типов. Для устранения этих ошибок предлагается применять структурную модель, использующую отношения между словами, и метрику сравнения слов с учётом соседних объектов.

5. CONCLUSION

В данной работе предлагается новый датасет PHONE-SCAN, включающий 451 изображение документов формата A4, напечатанных на русском языке. Изображения были оцифрованы сканированием и с помощью камер мобильных устройств. Последний способ предполагал нормализацию границ изображений с помощью приложения «Камера» для ОС Android. Кроме того, PHONE-SCAN содержит результаты распознавания изображений. Предлагаемый набор данных может быть полезен для исследований в области анализа русскоязычных документов формата A4, оцифрованных с помощью смартфонов.

Датасет опубликован в [10].

Представляет интерес дальнейшее пополнение предложенного датасета документами на других языках.

СПИСОК ЛИТЕРАТУРЫ

1. Document image binarization. URL: `https://dib.cin.ufpe.br`, 23-11-25
2. Arlazarov V., Bulatov K., Chernov T., Arlazarov V.L., MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream. Computer Optics, 2019, vol. 43, no. 5, pp. 818-824. `https://doi.org/10.18287/2412-6179-2019-43-5-818-824`.
3. Bulatov K., Matalov D., Arlazarov V.V., MIDV-2019: challenges of the modern mobile-based document OCR. 12th International Conference on Machine Vision (ICMV 2019); 11433: 114332N. `https://doi.org/10.1117/12.2558438`.
4. Chuiko A., Kunina I., Usilin S., Nikolaev D., Arlazarov V., MIDV-DM: A Document-Oriented Dataset for Image Manipulation Detection and Localization. Computer Optics, 2025, vol. 49, no. 6. `https://doi.org/10.18287/2412-6179-CO-1768`.
5. The Legacy Tobacco Document Library (LTDL), University of California, San Francisco, 2007. `http://legacy.library.ucsf.edu,23-11-25`.
6. Larson S., Lim G., Leach K., On Evaluation of Document Classification using RVL-CDIP. arXiv preprint. no. 2306.12550, 2023.

7. Li M., Cui L., Huang S., Wei F., Zhou M., Li. Z., TableBank: A Benchmark Dataset for Table Detection and Recognition. arXiv preprint. no. 1903.01949, 2020.

8. Slavin O., Concept controlling the quality of administrative document recognition. Studies in Systems, Decision and Control, Alla Kravets and Alexander A. Bolshakov (Eds): Cyber-Physical Systems: Engineering in Digital Era. Springer, Cham, 2025, vol. 624, pp. 243 – 253, https://doi.org/10.1007/978-3-032-02544-9_18.

9. Slavin O., Farsobina V., Myshev A.V., Analyzing the content of business documents recognized with a large number of errors using modified Levenshtein distance. Cyber-Physical Systems: Intelligent Models and Algorithms. Springer, Cham, 2022, vol. 417, pp. 267 – 279, https://doi.org/10.1007/978-3-030-95116-0.

10. PHONE-SCAN. https://github.com/slug183/phone-scan, 23-12-2025.

PHONE-SCAN: A set of document images digitized and normalized using a smart phone

O. A. Slavin, I. M. Janiszewski

This paper describes the PHONE-SCAN dataset for analyzing Russian-language A4 documents digitized by smartphones in real-world conditions. The dataset includes 451 samples of eleven document types with varying levels of noise and distortion, acquired by different users and devices. Tesseract OCR text recognition results are included. The dataset can be used to develop and test algorithms for classifying, normalizing, and enhancing document images.

KEYWORDS: control, information theory, algorithm.