УДК 004.891.3

ПРИМЕНЕНИЕ МЕТОДОВ КОМПЬЮТЕРНОГО ЗРЕНИЯ К РАСПОЗНАВАНИЮ СТАРОТАТАРСКОГО ТЕКСТА

И. А. Валишин[0009-0006-6891-031X]

Институт информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, ул. Кремлевская, 35, г. Казань, 420008

iskander1998@list.ru

Аннотация

Разработан инструмент, распознающий строки, слова и арабские символы с отсканированного изображения. Рассмотрены возможности и перспективы применения инструмента в исследовательской деятельности. Приведены результаты экспериментов по проверке работоспособности инструмента на примере старотатарских оцифрованных произведений.

Ключевые слова: YOLO, распознавание арабских символов, нейронные сети, компьютерное зрение.

ВВЕДЕНИЕ

В последние годы отечественные библиотеки и архивы переводят свой фонд в цифровое представление. Многие исторические документы (книги, чертежи, карты, рукописи и т. п.) в силу ветхости или уникальности доступны только ограниченному кругу специалистов. Их перевод в цифровое представление обеспечивает возможность доступа к ним широкому кругу читателей [1].

Ключевое множество общественно значимых, древних данных остается за бортом развития человеческого общества. Для расшифровки подобной информации используется «ручной труд» узкопрофиллированных специалистов. Это касается и старотатарского языка, ведь зачастую процесс перевода старинных рукописей сопровождается существенными временными и человеческими затратами. Подобные исследовательские проекты становятся высоко бюджетными и тяжело исполнимыми в надлежащие сроки. Автоматизация подобных процессов позволила бы сэкономить время и ресурсы. Также появилась бы возможность помочь

специалистам в области переводов древних рукописей. Модель автоматизации может заменить большую часть работы и значительно облегчить процесс расшифровки данных.

Предметом представленного исследования являются древние рукописные и печатные тексты на старотатарском языке. Объект исследования: применение методов компьютерного зрения в задачах распознавания старотатарских текстов.

Целью проведенного исследования является разработка модели, способной с помощью методов компьютерного зрения распознать старотатарский текст. Для достижения указанной цели были поставлены следующие задачи:

1) Провести анализ научной литературы в предметной области;
2) Обнаружить или подготовить необходимый набор данных;
3) Разработать модель распознавания текста;
4) Разработать графический интерфейс модели.

1. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для достижения цели исследования был произведен поиск аналогичных решений с помощью таких систем поиска научных статей, как Google scholar, «КиберЛенинка», eLibrary.ru и ряда других. Желаемым результатом поиска было обнаружение существующих решений, которые могли бы помочь достичь поставленную цель или максимально приблизиться к ее достижению. Поиск был декомпозирован на следующие подпункты:

1) Поиск моделей и решений в области распознавания текста любого языка по типу OCP (оптическое распознавание символов);
2) Поиск моделей и решений в области распознавания старотатарского текста;
3) Поиск моделей и решений в области распознавания текстов на основе арабской письменности.

Был также осуществлен поиск информации об инструментах и моделях, которые могли бы распознавать тексты на различных языках с изображения и перевести его в цифровое поле. Результаты поиска оказались удовлетворительными по причине достаточно глубокой изученности данной проблемы. Одним из самых популярных инструментов в области OCR является Tesseract, который берет свое начало с середины 1980-х годов. Эта программа разрабатывалась компанией

Hewlett-Packard, а с 2006 года разработка финансировалась компанией Google [2]. Названный инструмент дает возможность пользователям работать в различных операционных системах и хорошо справляется с распознаванием и переводом в цифровое пространство печатных и письменных текстов на изображениях [3]. Но в основном подобного рода модели натренированы на английском языке и не имеют графического интерфейса, как в случае с Tesseract, что является причиной их меньшей доступности для обычных пользователей. Исследователи также отмечают программы Microsoft Onenote, Abbyy Finereader и др. [4] [5]. Названные инструменты являются коммерческими и одними из самых популярных решений, которые представлены на рынке OCR. Анализ научной литературы, в которой описаны модели OCR, показал, что старотатарский язык отсутствует во всех решениях в качестве языка для распознавания. Присутствие арабского языка в подобных инструментах не играет существенной роли, поскольку арабица является основой письменности старотатарского языка, но слова являются татарскими.

Поиск информации об оптическом распознавании старотатарского языка результатов не дал – авторами эта проблематика не была затронута. Поэтому было решено проанализировать научную литературу об оптическом распознавании арабоязычных текстов и арабской графики. В результате было обнаружено значительное количество научных статей, которые представляют разработки в области распознавания арабоязычного текста, слов и букв (см., например, [17–20]). Имеются также публикации, в которых представлен сравнительный анализ существующих решений в рассматриваемой предметной области. В частности, установлено, что модели, в которых использовались нейронные сети (LSTM, deep CNN, CNN-RNN и др.), показали высокий уровень метрики Accuracy (в процентном соотношении от 76.3% до 99.3%), что позволяет решать множество разнообразных задач в области распознавания арабоязычных текстов. В названных моделях использовались такие датасеты, как IFN/ENIT, HACDB, APTI, HMBD, Hijja data set (рис. 1). В ходе поиска оптимального датасета для решения поставленной задачи был осуществлен поиск названных выше наборов данных. Было также установлено, что имеющиеся датасеты разделились на рукописные и печатные. При этом арабоязычные датасеты рукописного текста представлены в виде множества изображений различных арабских букв и слов с особенностями почерка авторов.

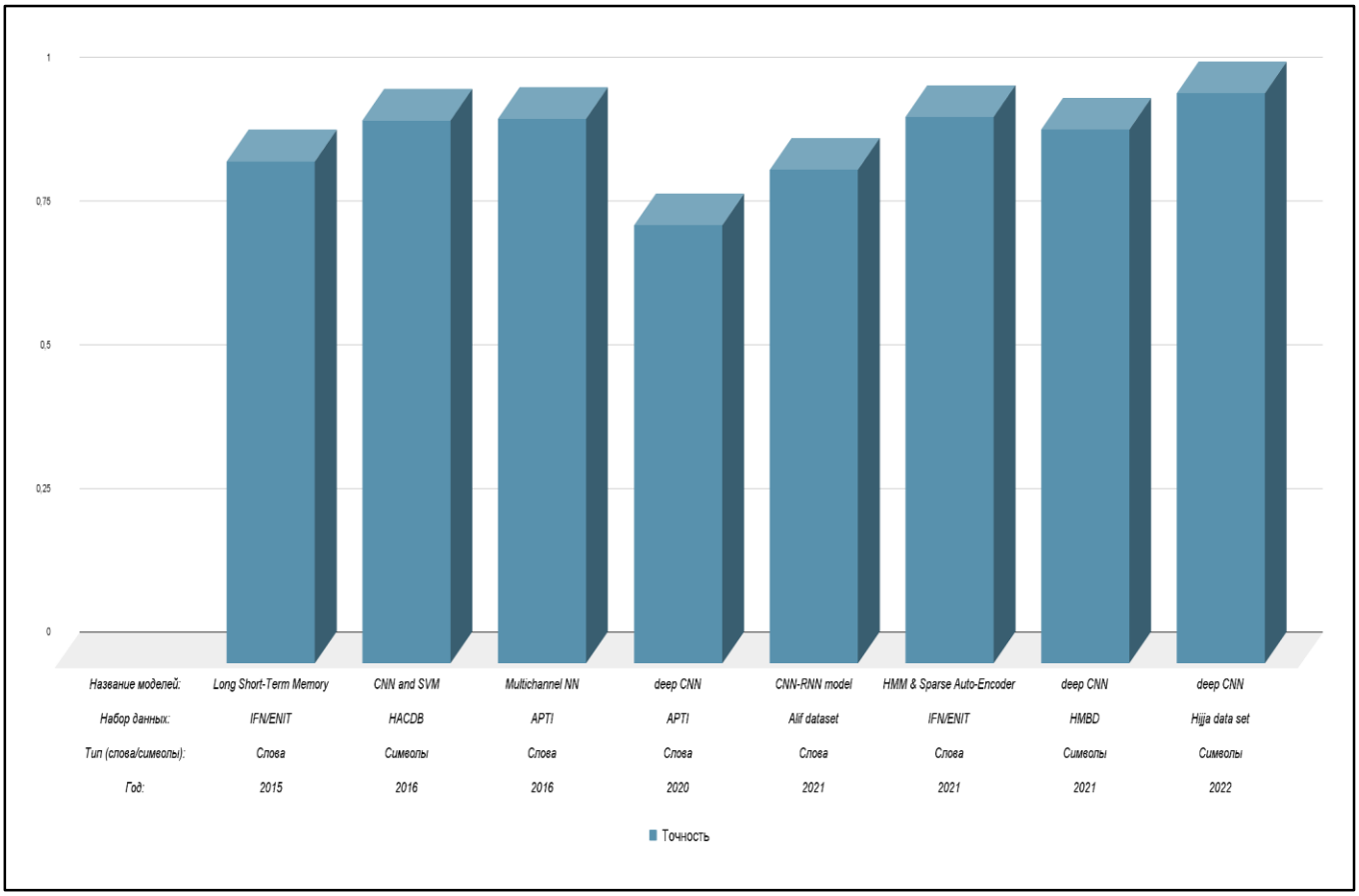

Рис. 1. Оценка качества моделей

IFN/ENIT датасет рукописного арабского текста представлен на официальном сайте, но только в демоверсии. Он содержит 2200 изображений из Тунисских городов с качеством изображения в 300 dpi. Включает в себя 26450 арабских слов и 212211 букв [6]. Полная версия набора данных предоставляется платно для коммерческих целей и бесплатно — для учебных.

Набор данных HACDB состоит из 6600 рукописных арабских букв. В этом датасете учитываются особенности арабского языка, в частности, то, что буква может быть в 4-х положениях: начальной, средней, конечной и изолированной. Этот факт делает датасет полезным для исследователей в области распознавания арабоязычных текстов, потому что буквы в нем представлены во всех возможных их проявлениях в письме [7].

Набор данных Hijja dataset включает коллекцию рукописных арабоязычных букв, которые были собраны у школьников в возрасте от 7 до 12 лет [8]. Данные

были собраны в Рияде, Саудовской Аравии в январе – апреле 2019 года. Датасет включает 47434 арабских символа разного почерка от 591 участника.

Самым упоминаемым датасетом в научных работах, которые касаются области распознавания арабоязычных текстов, является набор данных APTI. Он включает 45313600 изображений. Каждое изображение – одно арабское слово. Этот датасет был искусственно сгенерирован из 113284 слов с 10 типами шрифта и размером шрифта от 6 до 24 pt.

Перечисленные наборы данных позволяют успешно решать задачи распознавания арабоязычных текстов как в печатном, так и в рукописном видах.

Для датасета Hijja совместно с набором AHCD была разработана модель на основе CNN [9]. Модель достигает точности распознавания рукописных арабских символов 97% для Hijja и 88% для AHCD. Использование сверточных нейронных сетей демонстрирует существенные результаты в области распознавания арабских символов. Модели, которые будут представлены в следующих разделах настоящей статьи, в большинстве случаев используют CNN.

Система распознавания арабских символов AHCR-DLS показала высокую точность распознавания на тренировочных датасетах HMB1 и HMB2 от 94,9% до 97,3% [10].

Модели на основе сверточных сетей также предлагают новые подходы в классификации арабских символов с оптимизированной функцией активации ReLU. Подходы, предложенные в [11], показали точность модели 97,8% с использованием новой методики классификации.

Модель UnCNN распознавания изолированных арабоязычных символов показала конкурентоспособную точность в сравнении с приведенными выше моделями. Для оценки точности модели использовались датасеты IFHCDB, AHCD, AIA9K и HACDB [12].

Помимо обычных моделей в научной литературе также представлены и новые подходы в решении задач распознавания арабоязычных текстов, например, подход, который состоит из 4 этапов и включает state of the art модели для решения задачи. На первом этапе в YOLOv4 с помощью CNN обучается распознавание арабских печатных символов. Второй метод включает обработку перекрывающихся ограничивающих рамок, чтобы обеспечить выбор наиболее точной рамки

для каждого символа. Третий метод использует библиотеку Hunspell для проверки правописания слов и исправления ошибок. Четвертый метод использует расстояние редактирования для сравнения слов с ошибками в написании OCR с предложениями Hunspell и выбора ближайшего правильного слова. Предложенная система PAOCR достигла впечатляющей точности 82,4% для набора данных, состоящего из печатных арабских символов.

Таким образом, выше описаны датасеты арабоязычных символов (как печатных, так и письменных) для решения задачи распознавания арабских символов. Отмечены актуальные модели на основе CNN, их метрики и датасеты, которые они использовали для обучения. Указаны новые подходы в решении задачи распознавании арабских символов, что существенно помогает в распознавании старотатарского текста, который состоит из арабских символов. Выделены также коммерческие OCR, которые могли бы быть полезны для решения поставленных задач.

2. ПОДГОТОВКА НАБОРА ДАННЫХ

Решение задачи распознавания старотатарского текста требует обнаружения наиболее подходящего датасета для работы с символами и словами. В предыдущем разделе были рассмотрены наборы данных для распознавания арабских символов: IFN/ENIT, HACDB, APTI, Hijja. Эти наборы данных позволяют решить задачу распознавания арабских символов, но не затрагивают задачу распознавания слов и строк.

Для достижения поставленной цели был осуществлен выбор в пользу методики локализации символа – она предполагает локализацию объекта до тех пор, пока он не будет точно определен в качестве конечного результата поиска. В нашем исследовании предметом является старотатарский оцифрованный текст. Если рассматривать подобный текст в качестве примера, то, исходя из названной методики, нужно определить конечный результат поиска системы. Результатом будет являться арабский символ для последующей возможности собрать готовое слово из символов. Для локализации и повышения точности распознавания символа на оцифрованном изображении высокого качества необходима локализация в виде разделения текста на строки, а после – на слова. Таким образом повы-

сится точность распознавания символа, и он не затеряется среди других распознанных символов. Для решения такой задачи возникает необходимость создания набора данных для трех моделей распознавания. Первая модель будет разделять текст на строки, вторая – на слова, а третья – на символы.

В первом случае (разделения на строки) выбор был сделан в пользу старотатарского текста. Оцифрованное изображение со старотатарским текстом было размечено с одним основным классом «lines». Для работы с разметкой данных использовался сервис Roboflow, который позволяет аннотировать изображения в браузере, а затем экспортировать их для обучения модели. Было размечено 26 изображений, в них 569 аннотаций (строк) (рис. 2). Средний размер изображений составил 1466x2381 пикселей. В качестве встроенного препроцессинга изображений в сервисе Roboflow было использовано: Auto-Orient: Applied, Resize: Fit within 736x736, Grayscale: Applied, Auto-Adjust Contrast: Using Contrast Stretching. Аугментация позволила увеличить размер датасета до 37 изображений. Для аугментации использовались: Flip: Horizontal, Vertical Grayscale: Apply to 15% of images, Saturation: Between -25% и +25%, Blur: Up to 2px. Набор данных был разделен на 33 тренировочных, 3 валидационных и 1 тестовых изображений.

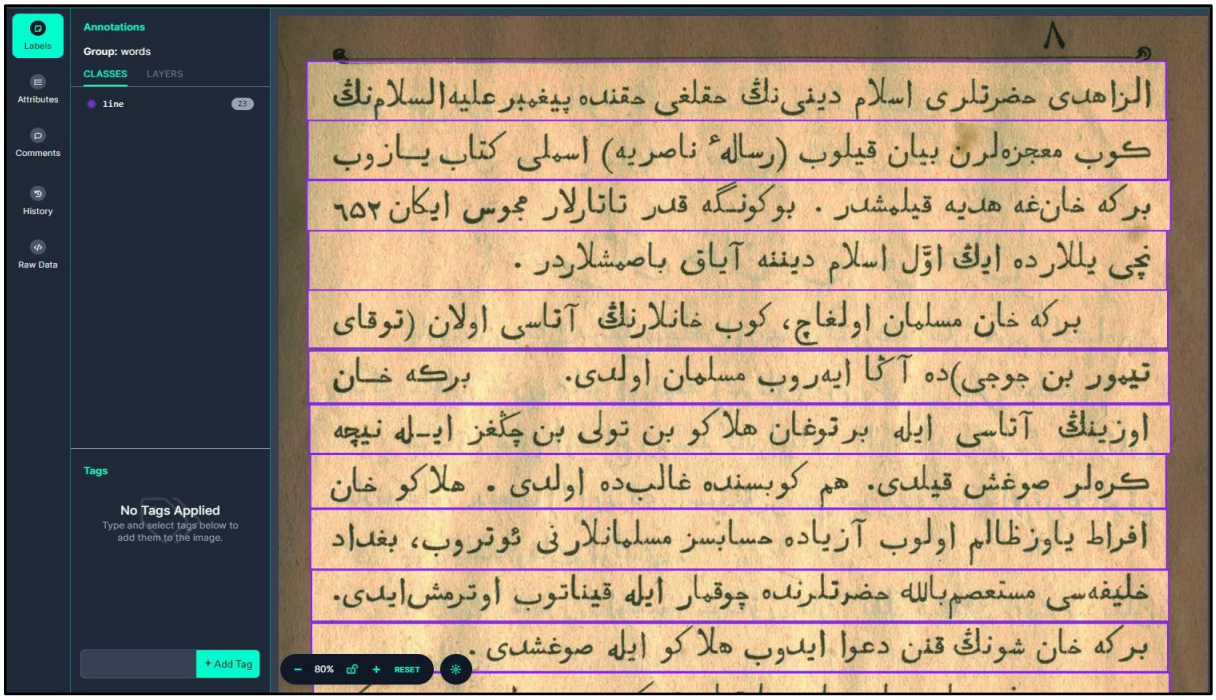

Рис. 2. Пример аннотирования изображения для модели распознавания строк в сервисе Roboflow

Следующей задачей после распознавания строк является распознавание слов внутри выделенных ранее строк. Для сбора датасета использовались вырезанные изображения строк из старотатарского оцифрованного текста. Они стали основой набора данных для обучения модели, которая может распознать старотатарское слово и выделить его в отдельное изображение. Для разметки набора данных со старотатарскими словами использовался сервис Roboflow. Единственным классом набора являлось «word». Во время разметки данных в одном изображении (строке) размечалось от 2 до 5 слов. Было размечено 100 изображений, в них 788 аннотаций (слов). Средний размер изображений составил 1458x108 пикселей. В ходе препроцессинга использовались: Auto-Orient: Applied, Resize: Fit within 640x640, Grayscale: Applied, Auto-Adjust Contrast: Using Contrast Stretching. Аугментация позволила увеличить размер датасета до 220 изображений. Для аугментации использовались: Outputs per training example: 3, Grayscale: Apply to 15% of images, Exposure: Between -13% and +13%, Blur: Up to 1.9px. Набор данных был разделен на 180 тренировочных, 25 валидационных и 15 тестовых изображений (рис. 3).

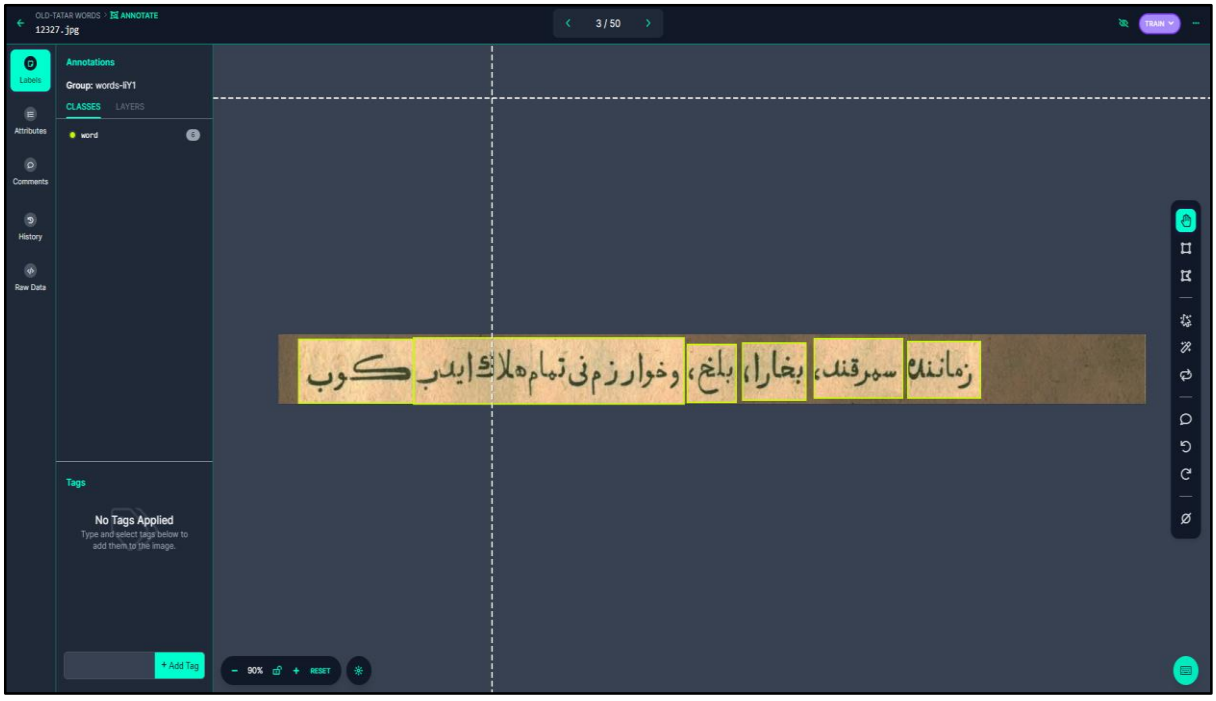

Рис. 3. Пример аннотирования изображения для модели распознавания старотатарских слов в сервисе Roboflow

Большая часть датасетов, перечисленных выше, находится в ограниченном доступе. В ходе поисков наборов данных, находящихся в открытом доступе, был обнаружен датасет арабоязычных рукописных символов HMBD v1 [13]. Этот набор данных включает 115 категорий (четыре вида написания одного символа), 54115 уникальных изображений с белым фоном и черными буквами [14]. Высота и ширина каждого изображения в датасете составляют 300 пикселей (рис. 4). В 2021 году с использованием этого датасета была создана модель deep CNN, которая показала 92,8% точности распознавания арабских рукописных символов. В наборе данных было 115 классов арабских символов, на каждый символ приходилось в среднем 4 класса для одного символа: начальное написание, между буквами, конечное написание, изолированный символ.

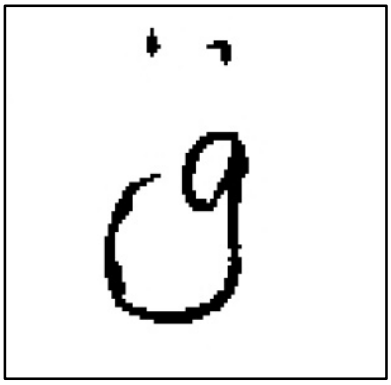

Рис. 4. Пример изображения из датасета HMBDv1 (класс Qaf_Isolated)

В старотатарском языке буквы арабские, а слова татарские. Этот факт позволил использовать датасет HMBD v1 для создания собственного набора данных, который подойдет под условия распознавания старотатарского языка, где главное условие – наличие арабских символов. Для разметки набора данных с арабскими рукописными символами был использован сервис Roboflow. Множества классов из датасета были объединены в локальные классы для одного символа. Например, классы: Kaf_End, Kaf_Isolated, Kaf_Middle, Kaf_Start, Khaa_End, Khaa_Isolated, Khaa_Start, Qaf_End, Qaf_Isolated, Qaf_Middle, Qaf_Start были объединены в один класс под названием «К». Изображения из этих классов были собраны в один класс «К» в количестве около 300 изображений и впоследствии распределены между тренировочной, валидационной и тестовой выборками в процентном соотношении 70%, 20%, 10% соответственно. Приведенное ранее решение применялось и к остальным классам. Это позволило сократить количество

классов до 22, что в 5,2 раза меньше оригинального набора данных. Также это позволило оптимизировать датасет для последующего обучения, количество изображений сократилось до 6400. Всего в созданном датасете для модели распознавания арабских (старотатарских) символов было создано 22 класса (рис. 5): A, Wau, Iy, N, K, R, L, D, B, T, M, S, Ha, G, Sh, La, Z, F, Ta, h, Zzh. Часть из них обозначает звуки или буквы из татарского языка для более точного воссоздания слова. В каждом классе набора данных было собрано от 325 до 945 изображений. Для улучшения выборки датасета было добавлено и размечено 600 изображений со староататарскими словами. Это позволило увеличить размер датасета до 7000 изображений с размеченными арабскими символами. Аугментация позволила увеличить размер датасета до 11476 изображений. В ходе препроцессинга использовались:Auto-Orient: Applied Resize: Stretch to 320x320, Grayscale: Applied, Auto-Adjust Contrast: Using Contrast Stretching. Для аугментации использовались: Flip: Horizontal Grayscale: Apply to 15% of images, Brightness: Between -22% and +22%, Blur: Up to 4.5px. Набор данных был разделен на 8952 тренировочных, 1599 валидационных и 925 тестовых изображений.

В результате сбора данных было создано три датасета, каждый из которых определял будущую модель: модель для распознавания строк на отсканированном изображении, модель для распознавания старотатарских слов в распознанных ранее строках и модель распознавания арабских символов в распознанных ранее словах. В сумме на три датасета были размечены более 7000 изображений и 24 класса.

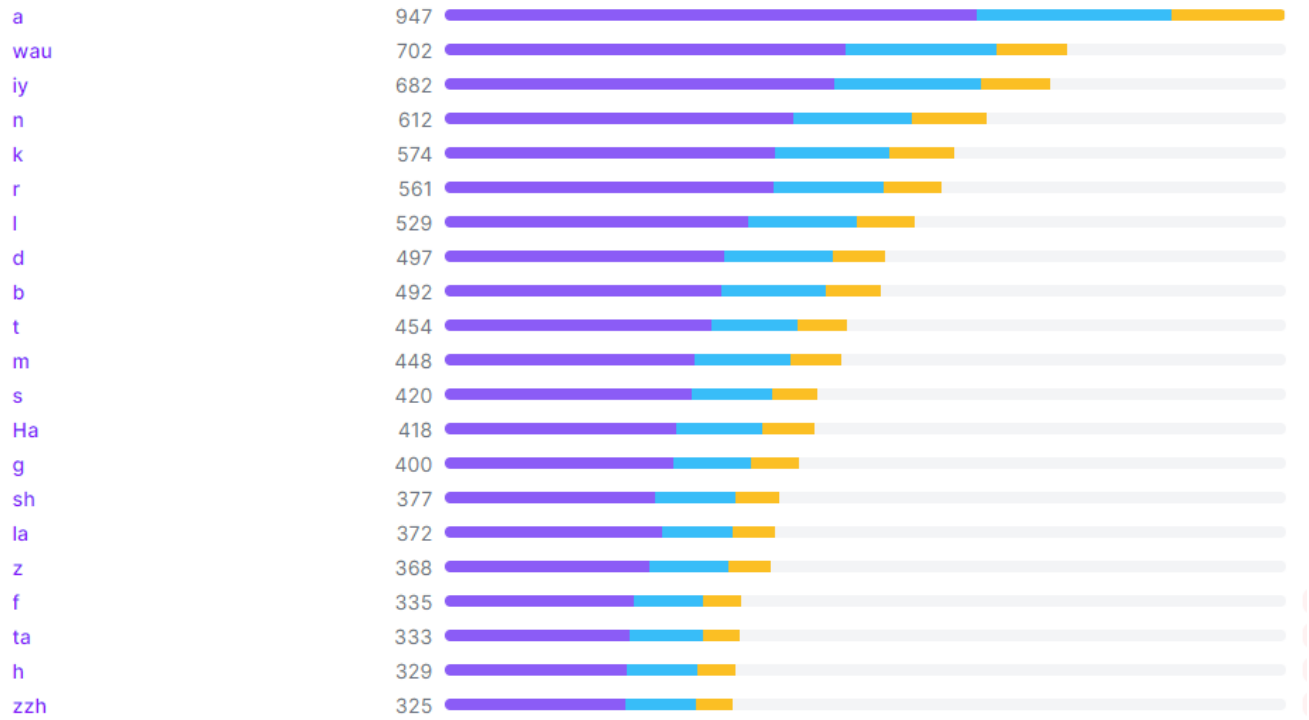

Рис. 5. Вкладка Health Check со всеми классами датасета по распознаванию арабских символов в сервисе Roboflow

РАЗРАБОТКА МОДЕЛИ РАСПОЗНАВАНИЯ ТЕКСТА

Исходя из сведений, приведенных в предыдущих двух разделах статьи, был разработан подход, который позволил оптимально решить поставленную задачу. Отсканированное изображение будет передаваться в систему распознавания старотатарского текста. Система будет делить отсканированное изображение со старотатарским текстом на строки. Далее система будет находить в строках старотатарские слова. В заключение старотатарское слово будет распознано по буквам таким образом, чтобы пользователь смог его самостоятельно определить (рис. 6).

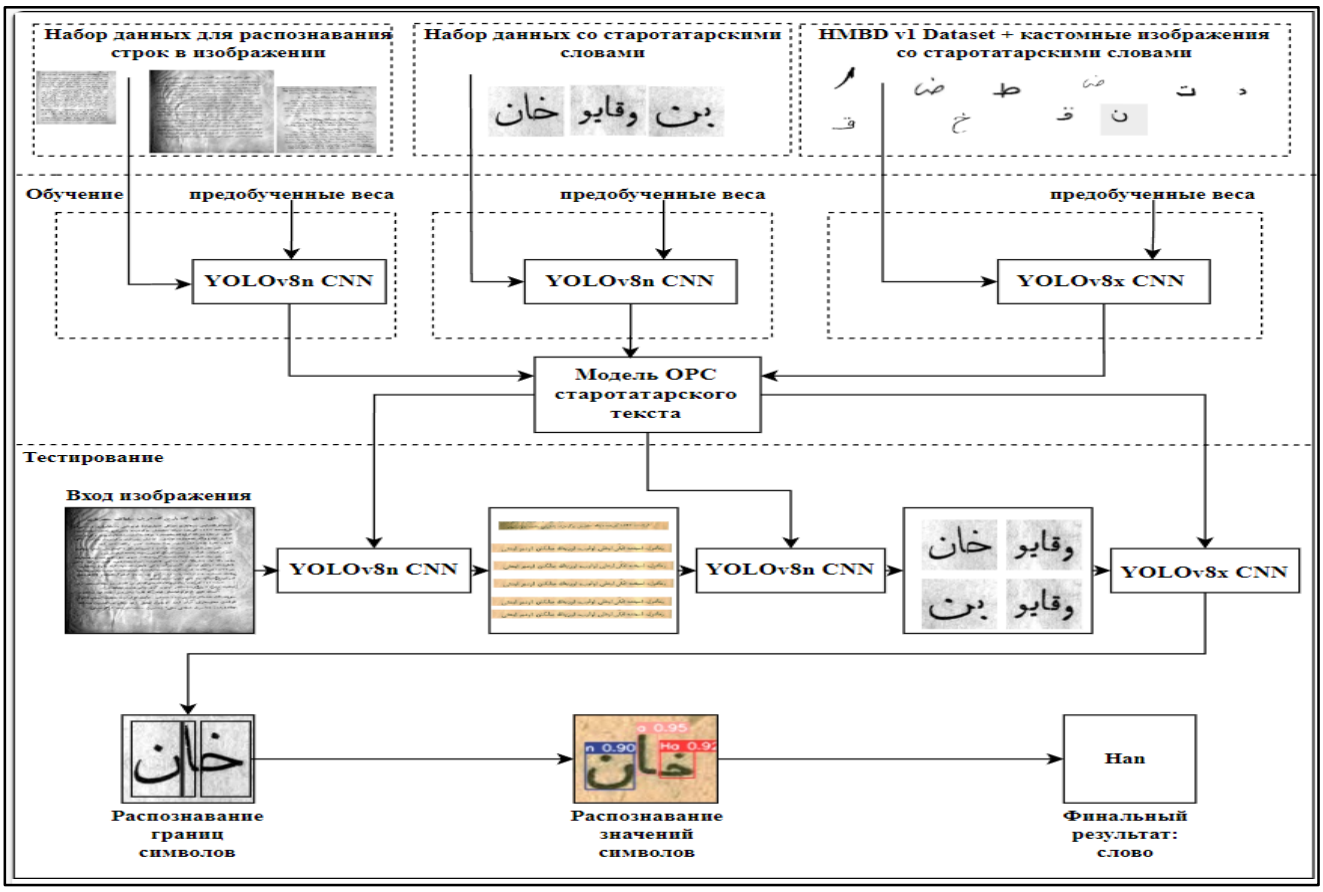

Рис. 6. Схема разработанного подхода для распознавания старотатарского текста

Всего система включает три модели и графический функциональный интерфейс: модель по распознаванию строк на основе YOLOv8n(3.2м параметров), модель по распознаванию старотатарских слов на основе YOLOv8n (3.2м параметров), модель по распознаванию арабских символов на основе YOLOv8x (68.2м параметров), графический интерфейс на основе PyQt5.

YOLO (You Only Look Once) CNN является state of the art моделью в области детекции объектов [15]. По сравнению с другими предложенными CNN моделями в распознавании арабских символов, YOLO также обладает повышенным FPS, если на вход модели подаётся видео формат. В дальнейших исследованиях возможна реализация подобного решения.

Для работы по распознаванию строк и старотатарских слов достаточно было использовать модель YOLOv8n. Эта модель обладает меньшим количеством параметров по сравнению с YOLOv8x, но больше подходит для решаемой задачи, потому что в датасетах слов и строк представлен лишь один класс.

Это позволило сократить время обучения нейронной сети и достигнуть необходимых результатов распознавания строк и слов в строках. Для модели распознавания арабских символов, напротив, использовалась модель YOLOv8x по причине того, что в наборе данных было свыше 11 тысяч изображений и 22 класса. Для большого количества классов необходима и высокая точность распознавания каждого класса. Наибольшее количество параметров и более точное распознавание объектов среди других моделей YOLOv8 позволило достичь необходимых результатов в задаче распознавания арабских символов (рис. 7).

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

Рис. 7. Сравнительная таблица моделей YOLOv8

Модель распознавания строк обучалась на кастомном датасете, который был описан во втором разделе статьи. За основу был взят YOLOv8n с одним классом «line» (рис. 8). Обучение происходило на 30 эпохах и позволило достигнуть результатов: Precision: 0.89, Recall: 0.88, mAP50: 0.96, mAP50-95: 0.66. (Рис. 9). Для обучения модели использовались видеокарты NVIDIA GeForce GTX 1060 6GB и NVIDIA GeForce GT 1080 8GB. Приведенные метрики в дальнейшем также использовались для оставшихся двух моделей распознавания старотатарских слов и арабских символов.

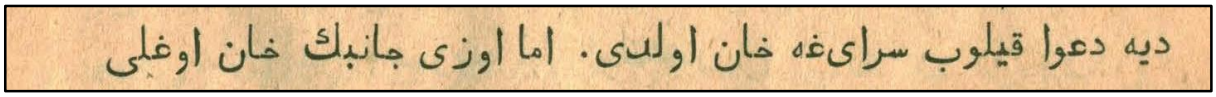

Рис. 8. Результат работы модели по распознаванию строк

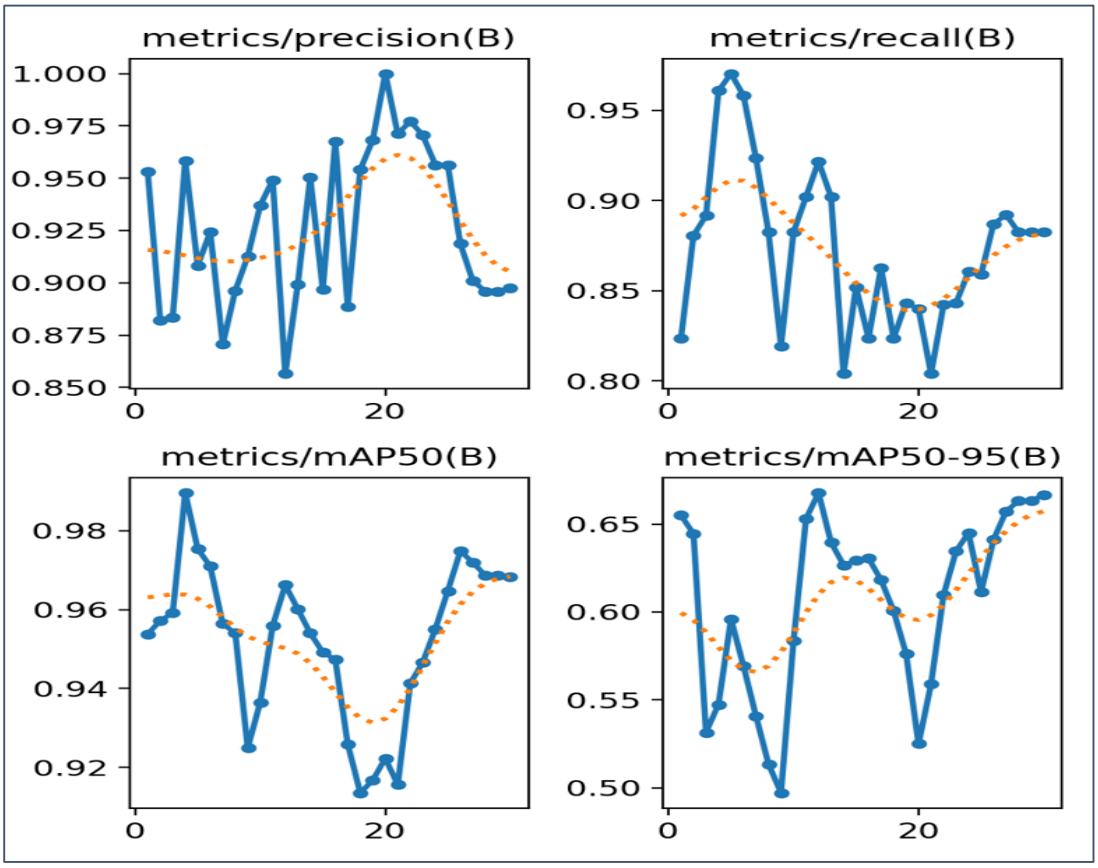

Рис. 9. Метрики, полученные в результате обучения модели распознавания строк

Модель распознавания старотатарских слов обучалась на кастомном датасете, который включал результаты работы предыдущей модели. На изображениях строк размечались старотатарские слова (рис. 10). За основу был взят YOLOv8n с одним классом «word». Обучение происходило на 50 эпохах и позволило достигнуть результатов: Precision: 0.95, Recall: 0.94, mAP50: 0.98, mAP50-95: 0.65 (рис. 11). Для обучения модели использовались видеокарты: NVIDIA GeForce GTX 1060 6GB и NVIDIA GeForce GT 1080 8GB.

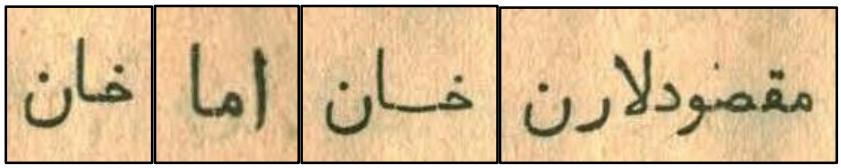

Рис. 10. Результат работы модели по распознаванию старотатарских слов из строк

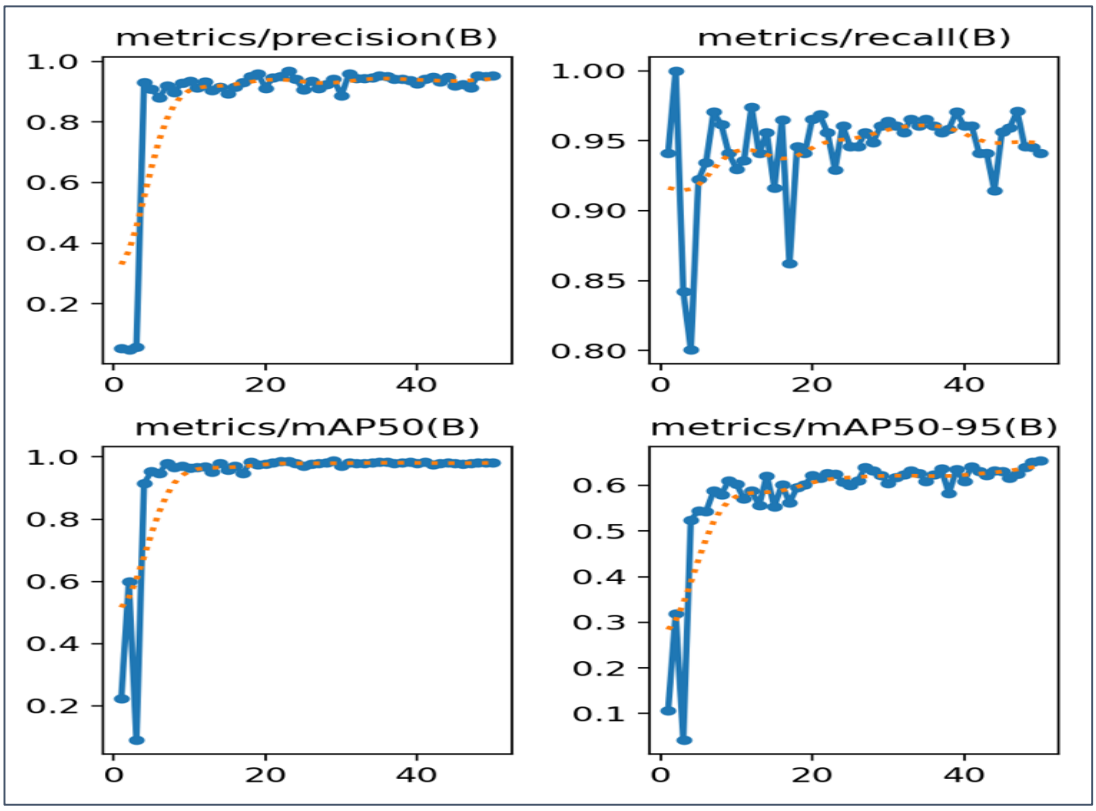

Рис. 11. Метрики, полученные в результате обучения модели распознавания старотатарских слов

Модель распознавания арабских символов обучалась на кастомном датасете изображений (результатов) из предыдущих моделей совместно с датасетом арабских рукописных символов HMBD v1. На изображениях старотатарских слов размечались арабские символы (рис. 12). За основу был взят YOLOv8x с 22 классами: A, Wau, Iy, N, K, R, L, D, B, T, M, S, Ha, G, Sh, La, Z, F, Ta, h, Zzh. Обучение происходило на 40 эпохах и позволило достигнуть результатов: Precision: 0.94, Recall: 0.93, mAP50: 0.95, mAP50-95: 0.67 (рис. 13). Для обучения модели использовались видеокарты NVIDIA GeForce GTX 1060 6GB и NVIDIA GeForce GT 1080 8GB.

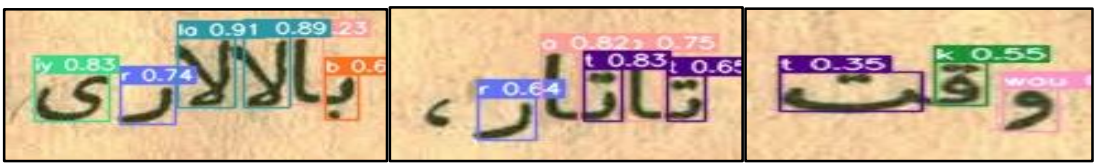

Рис. 12. Результат работы модели по распознаванию арабских символов в старотатарском слове. Слова: «Балалары», «Татар», «Вакыт».

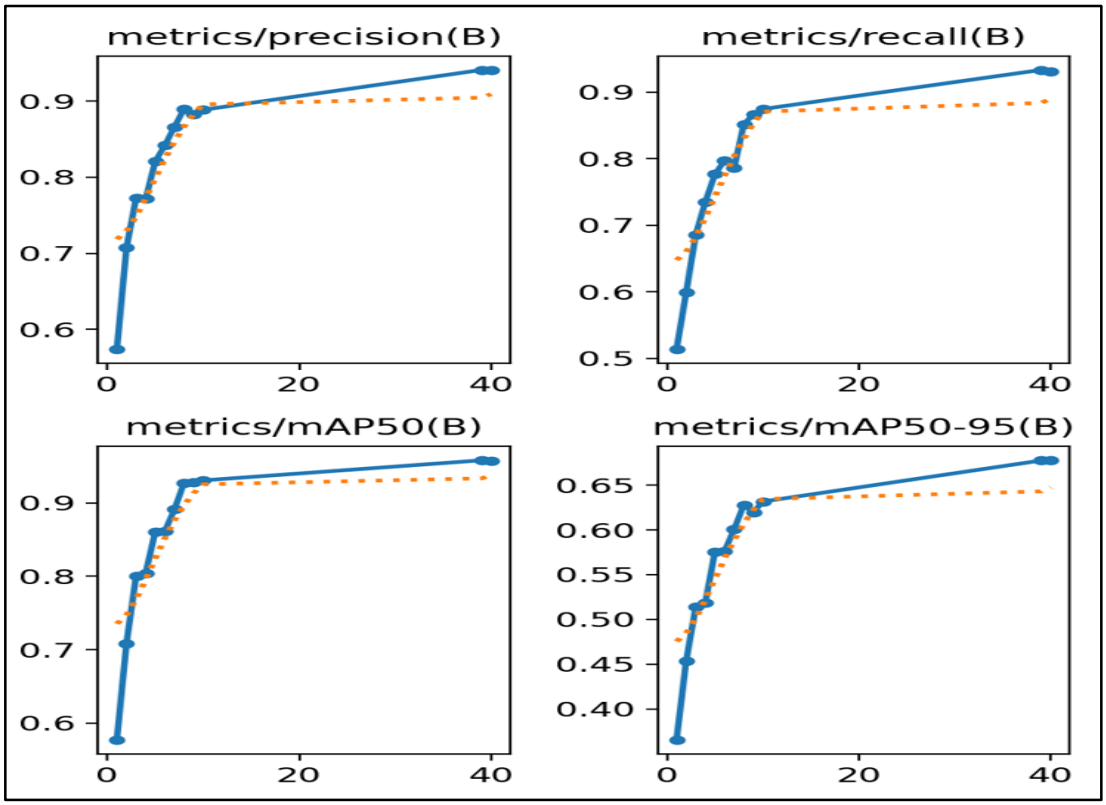

Рис. 13. Метрики, полученные в результате обучения модели распознавания арабских символов

В сумме модели выделяли строку, слово и арабский символ из отсканированного текста для последующих операций с ним. Результатом работы модели является старотатарское слово, буквы которого расшифрованы для удобного прочтения на латинице.

Итак, описаны разработанные подходы в решении задачи распознавания старотатарского текста, а также ход обучения моделей: распознавания строк, старотатарских слов и арабских символов. Представлены результаты работы моделей, гиперпараметры и результаты обучения в виде метрик.

ТЕСТИРОВАНИЕ СИСТЕМЫ

Чтобы определить точность работы системы и моделей, необходимо проверить ее на экземплярах, приближенных к реалиям. За основу тестирования системы были взяты отсканированные страницы со старотатарским текстом из произведения «Татар ханлары» (تاتار خانلاري) Мухаммадьяра Султанова, Типография братьев Каримовых, Казань, 1911 г. (рис. 14).

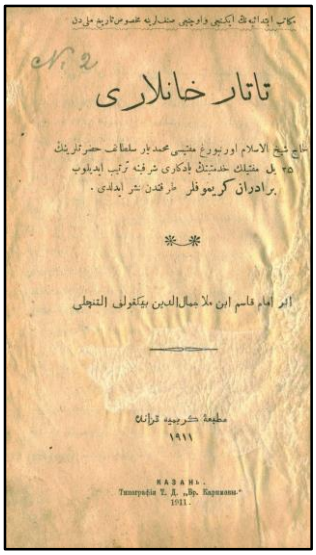

Рис. 14. Титульная страница произведения Мухаммальяра Султанова «Татар ханлары» (تاتار خانلاري) на старотатарском языке, Типография братьев Каримовых, Казань, 1911 г.

Сначала система реализует модель распознавания и разделения строк. Строки необходимо разделять для дальнейших операций с распознанными изображениями в моделях распознавания слов и символов. Для тестирования модели распознавания строк было выбрано 3 примера из представленного выше произведения.

Для модели распознавания слов рассмотрено несколько примеров с использованием старотатарского текста. Изображения, которые подаются на вход модели, имеют размерность в среднем 1450 x 100 px.

Из отсканированного выше изображения на вход модели распознавания слов были поданы отрезки изображений в виде строк из текста. Модель смогла определить 95 слов из 135 (рис. 15). Это могло произойти по причинам распознавания не всех строк, также слишком близкого расстояния между словами. Такой результат возможен при установке conf (порог доверия) на уровень 65%. При снижении порога доверия результат становится значительнее в плане результативности.

Рис. 15. Пример результата работы модели распознавания старотатарских слов

Далее на вход модели распознавания арабских символов подаются изображения слов, чтобы можно было разбить их на буквы и сделать, по возможности, транслитерацию этих слов. Из первых 10 слов примера из начала главы верно было распознано 57 арабских символов из 65 (рис. 16).

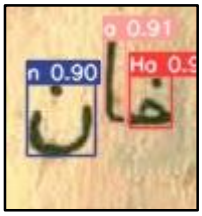

Рис. 16. Пример результата работы модели распознавания арабских символов. Распознанное слово «Хан»

При тестировании второго рандомного изображения из произведения «Татар ханлары» результатом стало 64 распознанных арабских символа из 66 на первые 10 распознанных слов.

Для более объективных результатов на вход системе подавались изображения из различных старотатарских оцифрованных произведений. Использовались в качестве примеров произведения: Мухаммадьяр Султанов «Татар ханлары» (تاتار خانلاري), Типография братьев Каримовых, Казань, 1911 г. 48 б. [21]; Шариф Камал. «Акчарлаклар»(آقچارلاقلر), Оренбург: "Вакыт" матбагасы, 1915. 88 б. [22]; «Аек бул»(آيق بول) Типо-литография Императорского университета, Казань, 1907 г. 16 б. [23]. Для тестирования системе подавалась одна рандомная страница из приведенных выше произведений. В произведении «Татар ханлары» было распознано 14 из 17 строк (82.3 %), 95 слов из 135 (70.3 %), 614 символов из 644 (95.3 %), средняя точность по всем распознанным элементам в системе:

(82.3+70.3+95.3)/3=82.6%.

В произведении «Аек бул» было распознано 6 из 9 строк (66.6 %), 16 слов из 23 (69.5 %), 50 символов из 52 (96.1 %), средняя точность по всем распознанным элементам в системе: (82.3+70.3+95.3)/3=77.4%.

В произведении «Акчарлаклар» было распознано 19 из 25 строк (76 %), 145 слов из 174 (83.3 %), 854 символов из 893 (95.6 %), средняя точность по всем распознанным элементам в системе: (82.3+70.3+95.3)/3=84.9%.

Результаты тестирования представлены в таблице 1.

Таблица 1. Результаты тестирования системы на основе выборки изображений из трех старотатарских оцифрованных произведений.

Произведение на старотатарском:	**"Татар ханлары" 1911**	**"Аек бул" 1907**	**"Акчарлаклар" 1915**
Распознанные строки	14 / 17	6 / 9	19 / 25
Распознанные слова	95 / 135	16 / 23	145 / 174
Распознанные символы	616 / 644	50 / 52	854 / 893
Средняя точность распознавания элементов	82.6 %	77.4 %	84.9 %

Результаты тестирования системы позволяют сделать вывод, что она умеет работать со старотатарским текстом: разделять изображение на строки, разделять строки на слова, слова на арабские буквы. Модель является показательной и не ставит перед собой цели получить высокую точность, но при этом может являться демонстрацией дальнейшего потенциала разработанного программного обеспечения.

РАЗРАБОТКА ГРАФИЧЕСКОГО ИНТЕРФЕЙСА

Для комфортной работы с моделями был создан графический интерфейс на основе PyQT5. Были разработаны функциональные требования для работы с моделями распознавания старотатарского текста. Перед разработкой графического интерфейса стоял выбор между Tkinter и PyQT. Выбор был сделан в пользу PyQT по причине большей производительности и более продвинутых возможностей для работы с изображениями по сравнению с Tkinter.

Графический интерфейс дает возможность пользователю осуществлять:

- Выбор изображения;
- Старт процесса распознавания выбранного изображения;
- Приближение изображения;
- Отдаление изображения;
- Подсчёт количества распознанных строк (рис. 17);
- Подсчёт количества распознанных слов (рис. 17);

- Подсчёт количества распознанных символов (рис. 17);
- Отображение изображения распознанных строк;
- Отображение координат распознанных строк;
- Отображение изображений распознанных слов;
- Отображение кириллической транслитерации распознанных слов;

Графический интерфейс состоит из 4 окон:

- Окно с изображением распознанных строк (рис. 18);
- Окно с прокруткой изображений распознанных слов (рис. 19);
- Окно с транслитерацией на кириллице распознанных слов (рис. 20);
- Окно с координатами распознанных строк (рис. 21).

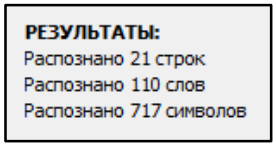

Рис. 17. Вывод результатов распознавания строк, слов и символов

Рис. 18. Окно с изображением распознанных строк

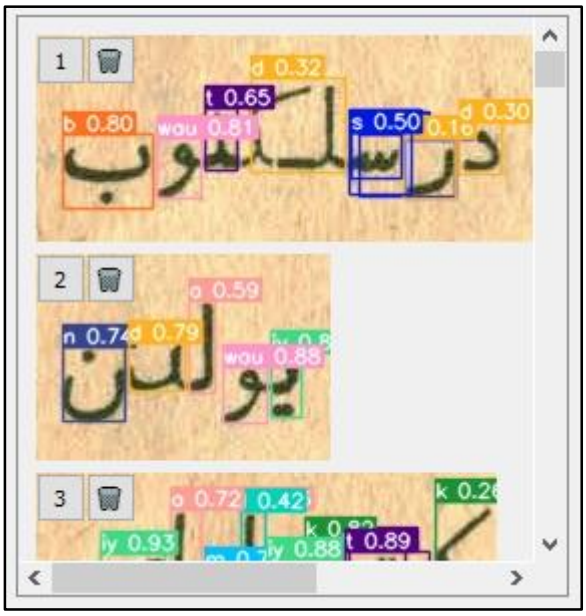

Рис. 19. Окно с прокруткой изображений распознанных старотатарских слов

Рис. 20. Окно с транслитерацией на кириллице распознанных старотатарских слов

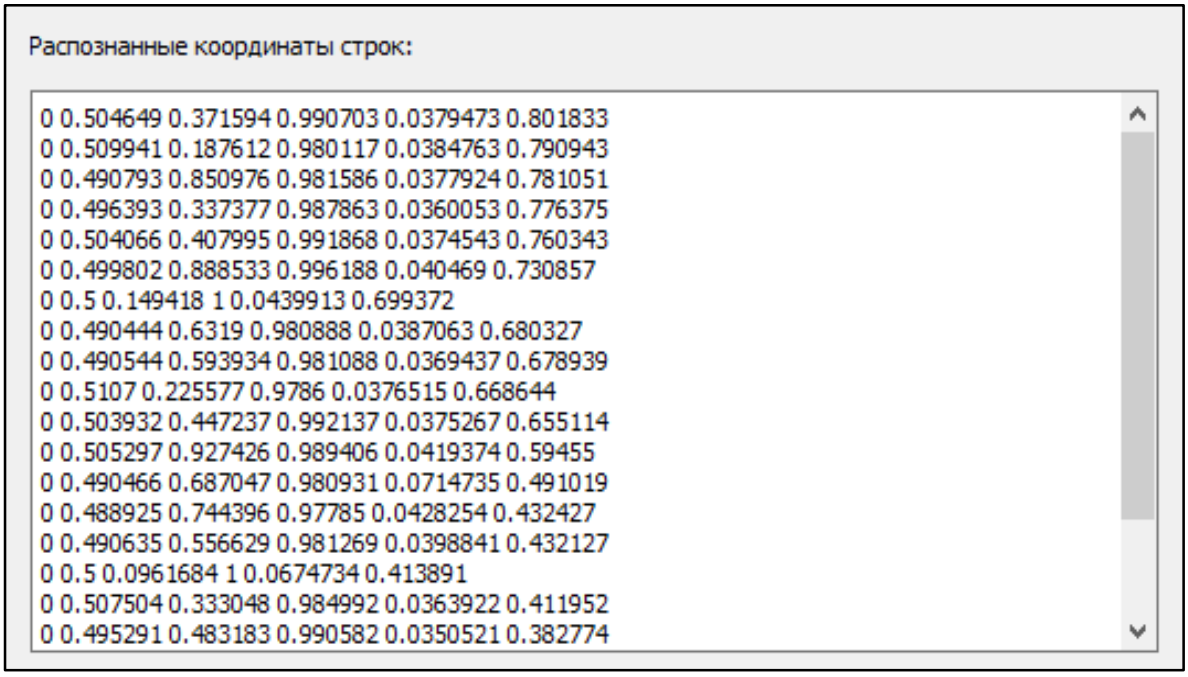

Рис. 21. Окно с координатами распознанных строк

Для удобства работы с функционалом приложения созданы кнопки:

- Выбор необходимого изображения для сканирования (рис. 22);
- Сканирование изображения (запуск процесса распознавания) (рис. 23);
- Приближение распознанного изображения (рис. 24);
- Отдаление распознанного изображения (рис. 25).

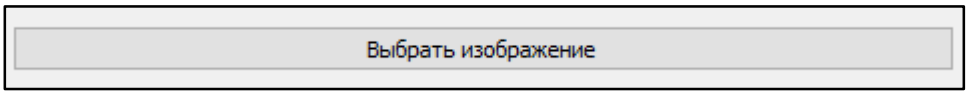

Рис. 22. Кнопка выбора необходимого изображения для сканирования

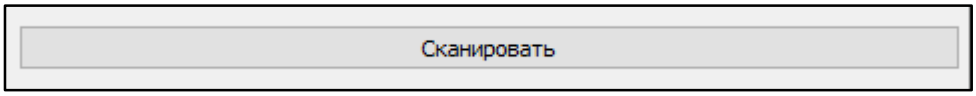

Рис. 23. Кнопка сканирования (запуска процесса распознавания текста)

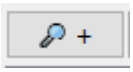

Рис. 24. Кнопка приближения распознанного изображения

Рис. 25. Кнопка отдаления распознанного изображения

Чтобы новые пользователи смогли ориентироваться в приложении, была создана инструкция, которая отображается в окне справа при первом запуске программы (рис. 26).

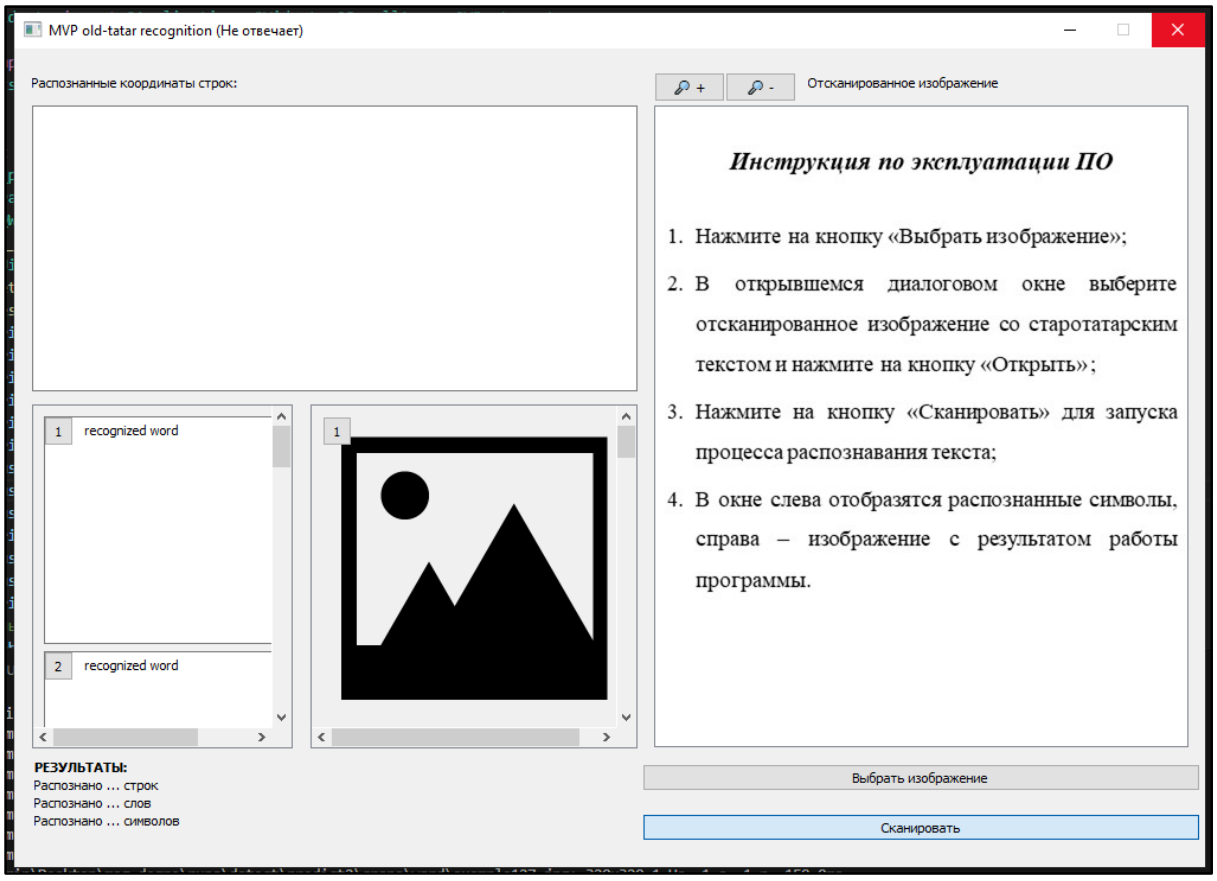

Рис. 26. Инструкция при начальном запуске программы

После выбора изображения и запуска процесса распознавания путем нажатия на кнопку «Сканировать» происходит процесс работы моделей распознавания строк, слов и символов. Отрабатываются функции транслитерации распознанных слов, подсчёта строк, подсчёта слов и подсчёта символов.

В пользовательском интерфейсе появляются результаты работы программы (рис. 27).

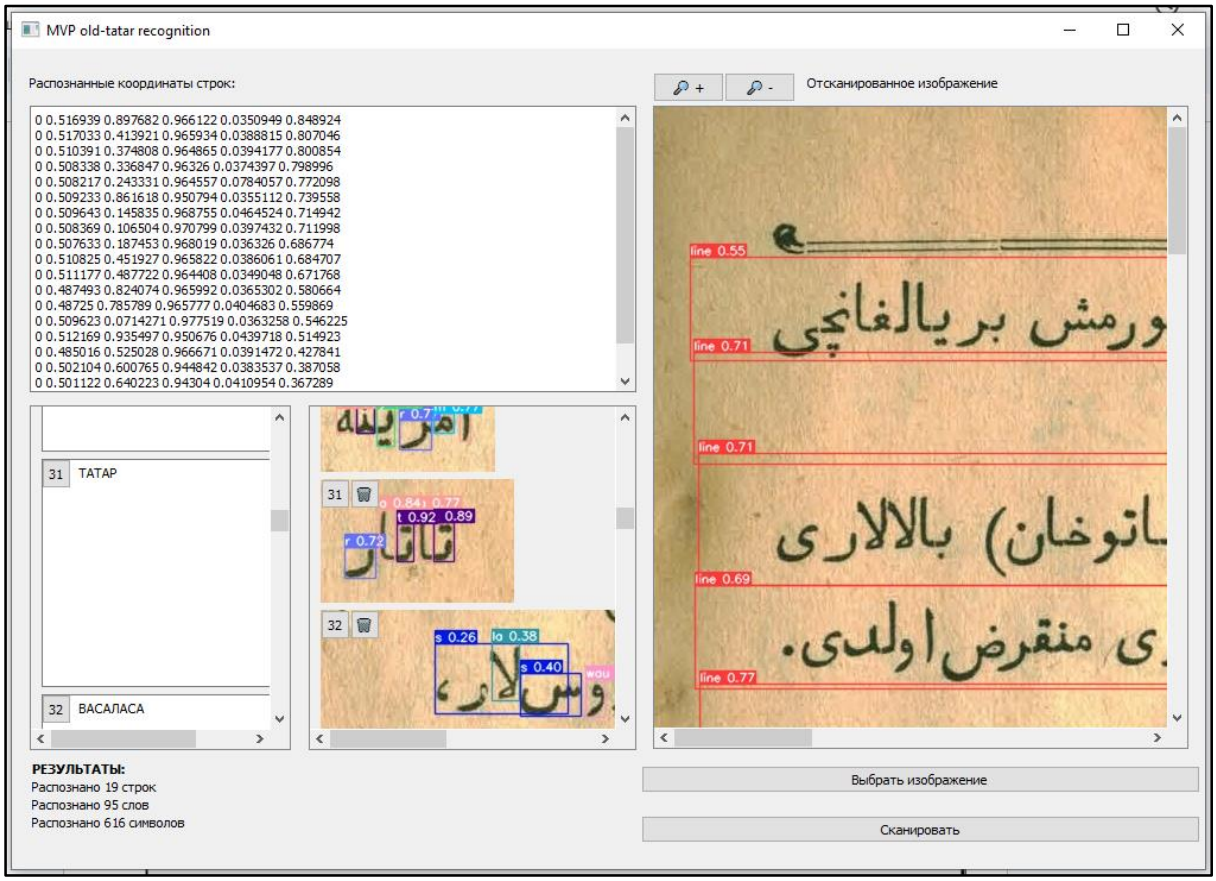

Рис. 27. Результаты работы программы

Разработанный графический интерфейс позволяет пользователям с удобством работать с процессом распознавания старотатарского текста. Созданные окна с транслитерацией старотатарских слов позволяют пользователям оперировать с результатами работы программы. Окна с распознанными строками и словами позволяют оперировать с текстом и находить необходимые слова.

ЗАКЛЮЧЕНИЕ

Таким образом, все поставленные задачи решены и достигнута цель исследования – разработка модели, способной с помощью методов компьютерного зрения распознать старотатарский текст.

Результатом работы стали:

- Создание набора данных для распознавания строк на старотатарском языке;
- Создание набора данных для распознавания слов на старотатарском языке;

- Создание набора данных для распознавания символов на старотатарском языке;
- Разработанная модель распознавания строк на старотатарском языке;
- Разработанная модель распознавания слов на старотатарском языке;
- Разработанная модель распознавания символов на старотатарском языке;
- Разработанный графический интерфейс модели.

Представленные разработанные методы и подходы в области распознавания старотатарского текста позволяют успешно решать также и задачи распознавания арабского текста. Модель протестирована и может быть полезна для исследователей древних рукописей, в которых использован старотатарский текст. Созданные датасеты могут быть востребованы в научной среде в задачах распознавания арабоязычных и иных текстов.

Исходные коды и файлы программы, которые были использованы в исследовательской работе, доступны по ссылке
URL: https://github.com/iskander1998/old-tatar

СПИСОК ЛИТЕРАТУРЫ

1. *Старовойтов В.В.* О цифровой реставрации исторических текстовых документов // Системный анализ и прикладная информатика. 2015. №1.
URL: https://cyberleninka.ru/article/n/o-tsifrovoy-restavratsii-istoricheskih-tekstovyh-dokumentov (дата обращения: 24.04.2024).

2. Announcing Tesseract OCR – The official Google Code blog. URL: https://googlecode.blogspot.com/2006/08/announcing-tesseract-ocr.html (дата обращения: 24.04.2024).

3. *Rice S., Jenkins F., Nartker T.* The Fourth Annual Test of OCR Accuracy. 2012.
URL: https://www.researchgate.net/publication/247886491_The_Fourth_Annual_Test_of_OCR_Accuracy (дата обращения: 24.04.2024).

4. *Андрианов А.И.* Сравнение OCR-систем на основе точности анализа изображения // Бизнес-информатика. 2009. №4.
URL: https://cyberleninka.ru/article/n/sravnenie-ocr-sistem-na-osnove-tochnosti-analiza-izobrazheniya (дата обращения: 01.05.2024)

5. *Нестеров А.С.* Анализ рынка современных информационных систем оптического распознавания символов (OCR) // Вопросы науки и образования. 2020. №23 (107). URL: https://cyberleninka.ru/article/n/analiz-rynka-sovremennyh-informatsionnyh-sistem-opticheskogo-raspoznavaniya-simvolov-ocr (дата обращения: 01.05.2024).

6. *Pechwitz M., Maddouri S.S., Märgner V., Ellouze N., Amiri H.* IFN/ENIT-database of handwritten Arabic words // Proc. of CIFED, Citeseer, 2002. P. 127–136.

7. *Lawgali A., Angelova M., Bouridane A.* HACDB: Handwritten Arabic characters da-tabase for automatic character recognition // European Workshop on Visual Information Processing (EUVIP), 2013. P. 255–259.

8. *Altwaijry N., Al-Turaiki I.* Arabic handwriting recognition system using convolutional neural network // Neural Comput. Appl. 2021. Vol. 33, No. 7. P. 2249–2261.

9. *Balaha H.M., Ali H.A., Saraya M., Badawy M.* A new Arabic handwritten character recognition deep learning system (AHCR-DLS) // Neural Comput. Appl. 2021. Vol. 33, no. 11. P. 6325–6367.

10. *Nayef B.H., Abdullah S.N.H.S., Sulaiman R., Alyasseri Z.A.A.* Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks// Multimed. Tools Appl. 2022. Vol. 81, No. 2. P. 2065–2094.

11. *Elkhayati M., Elkettani Y.* UnCNN: A New Directed CNN Model for Isolated Arabic Handwritten Characters Recognition // Arab J. Sci. Eng. 2022. Vol. 47, No. 8. P. 10667–10688.

12. *Balaha H.M.* HMBD: Arabic Handwritten Characters Dataset.
URL: https://github.com/HossamBalaha/HMBD-v1 (дата обращения: 09.05.2024).

13. *Balaha H.M., Ali H.A., Saraya M.* A new Arabic handwritten character recognition deep learning system (AHCR-DLS) // Neural Comput. Appl. 2021. Vol. 33. P. 6325–6367.

14. *Zou Z., Chen K., Shi Z., Guo Y., Ye J.* Object Detection in 20 Years: A Survey // Proceedings of the IEEE. 2023. Vol. 111, No. 3. P. 257–276.

15. *Закирьянов И.И., Хаялеева И.З., Валишин И.А., Курито Е.Д., Фасхутдинов А.Н.* Инструмент для распознавания языка жестов из видеопотока в режиме реального времени // Электронные библиотеки. 2023. Т. 26, № 6.
URL: https://rdl-journal.ru/article/view/804/876 (дата обращения: 01.05.2024).

16. *Mulyana D., Rowis M.* Optimization of Text Mining Detection of Tajweed Reading Laws Using the Yolov8 Method on the Qur'an // QALAMUNA: Jurnal Pendidikan, Sosial, Dan Agama. 2022. Vol. 14, No. 2. P. 1089–1110.

17. *Badr Al-Badr., Sabri A.M.* Survey and bibliography of Arabic optical text recognition // Signal Processing. 1995. Vol. 41, Issue 1. P. 49–77.

18. *Turki H., Elleuch M., Kherallah M., Damak A.* Arabic-Latin Scene Text Detection based on YOLO Models // International Conference on Innovations in Intelligent Systems and Applications (INISTA), Hammamet, Tunisia, 2023. P. 1–6.

19. *Rahal N., Tounsi M., Hussain A., Alimi A.M.* Deep Sparse Auto-Encoder Features Learning for Arabic Text Recognition // IEEE Access. 2021. Vol. 9. P. 18569–18584.

20. *Султанов М.* Татар ханлары (تاتار خانلاري) // Типография братьев Каримвых. 1911.
URL: https://darulkutub.com/uploads/books/820d9f6dcf1e868ee899d47e487b06189c2b816a.pdf (дата обращения: 01.05.2024).

21. *Камал Ш.* Акчарлаклар (آقچارلاقلر) // "Вакыт" матбагасы. 1915.
URL: https://miras.info/projects/mirasxane/books/425-akcharlaklar-.html (дата обращения: 01.05.2024).

22. Аек бул (آيق بول) // Типо-литография Императорского университета. 1907.
URL: https://darul-kutub.com/uploads/books/c3032b3c9136803dc0e38db69cd15541fb24064b.pdf (дата обращения: 01.05.2024).

APPLICATION OF COMPUTER VISION METHODS TO OLD TATAR TEXT RECOGNITION

I. A. Valishin[0009-0006-6891-031X]

Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, st. Kremlevskaya, 35, Kazan, 420008

iskander1998@list.ru

Abstract

A developed tool that recognizes strings, words and Arabic characters from scanned images. The possibilities and prospects for using the tool in research activities

are considered. The results of experiments on the operational performance of the instrument are presented using the example of Old Tatar digitized images.

Keywords: YOLO, Arabic character recognition, neural networks, computer vision.

REFERENCES

1. *Starovoitov V.V.* O cifrovoj restavracii istoricheskih tekstovyh dokumentov // Sistemnyj analiz i prikladnaya informatika. 2015. № 1.
URL: https://cyberleninka.ru/article/n/o-tsifrovoy-restavratsii-istoricheskih-tekstovyh-dokumentov

2. Announcing Tesseract OCR – The official Google Code blog.
URL: https://googlecode.blogspot.com/2006/08/announcing-tesseract-ocr.html

3. *Rice S., Jenkins F., Nartker T.* The Fourth Annual Test of OCR Accuracy. 2012. URL: https://www.researchgate.net/publication/247886491_The_Fourth_Annual_Test_of_OCR_Accuracy

4. *Andrianov A.I.* Sravnenie OCR-sistem na osnove tochnosti analiza izobrazheniya // Biznes-informatika. 2009. № 4.
URL: https://cyberleninka.ru/article/n/sravnenie-ocr-sistem-na-osnove-tochnosti-analiza-izobrazheniya

5. *Nesterov A.S.* Analiz rynka sovremennyh informacionnyh sistem optich-eskogo raspoznava-niya simvolov (OCR) // Voprosy nauki i obrazovaniya. 2020. № 23 (107).
URL: https://cyberleninka.ru/article/n/analiz-rynka-sovremennyh-informatsionnyh-sistem-opticheskogo-raspoznavaniya-simvolov-ocr

6. *Pechwitz M., Maddouri S.S., Märgner V., Ellouze N., Amiri H.* IFN/ENIT-database of handwritten Arabic words // Proc. of CIFED, Citeseer, 2002. P. 127–136.

7. *Lawgali A., Angelova M., Bouridane A.* HACDB: Handwritten Arabic characters da-tabase for automatic character recognition // European Workshop on Visual Information Processing (EUVIP). 2013. P. 255–259.

8. *Altwaijry N., Al-Turaiki I.* Arabic handwriting recognition system using convolutional neural network // Neural Comput. Appl. 2021. Vol. 33, No. 7. P. 2249–2261.

9. *Balaha H.M., Ali H.A., Saraya M., Badawy M.* A new Arabic handwritten character recognition deep learning system (AHCR-DLS) // Neural Comput. Appl. 2021. Vol. 33, no. 11. P. 6325–6367.

10. *Nayef B.H., Abdullah S.N.H.S., Sulaiman R., Alyasseri Z.A.A.* Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks// Multimed. Tools Appl. 2022. Vol. 81, No. 2. P. 2065–2094.

11. *Elkhayati M., Elkettani Y.* UnCNN: A New Directed CNN Model for Isolated Arabic Handwritten Characters Recognition // Arab J. Sci. Eng. 2022. Vol. 47, No. 8. P. 10667–10688.

12. *Balaha H.M.* HMBD: Arabic Handwritten Characters Dataset.
URL: https://github.com/HossamBalaha/HMBD-v1 (дата обращения: 09.05.2024).

13. *Balaha H.M., Ali H.A., Saraya M.* A new Arabic handwritten character recognition deep learning system (AHCR-DLS) // Neural Comput. Appl. 2021. Vol. 33. P. 6325–6367.

14. *Zou Z., Chen K., Shi Z., Guo Y., Ye J.* Object Detection in 20 Years: A Survey // Proceedings of the IEEE. 2023. Vol. 111, No. 3. P. 257–276.

15. *Zakiryanov I.I., Khayaleeva I.Z., Valishin I.A., Kurito E.D., Faskhutdinov A.N.* Instrument dlya raspoznavaniya yazyka zhestov iz videopotoka v rezhime real-nogo vremeni // Elektronnye biblioteki. 2023. T. 26, № 6.

16. *Mulyana D., Rowis M.* Optimization of Text Mining Detection of Tajweed Reading Laws Using the Yolov8 Method on the Qur'an // QALAMUNA: Jurnal Pendidikan, Sosial, Dan Agama. 2022. Vol. 14, No. 2. P. 1089–1110.

17. *Badr Al-Badr., Sabri A.M.* Survey and bibliography of Arabic optical text recognition // Signal Processing. 1995. Vol. 41, Issue 1. P. 49–77.

18. *Turki H., Elleuch M., Kherallah M., Damak A.* Arabic-Latin Scene Text Detection based on YOLO Models // International Conference on Innovations in Intelligent Systems and Applications (INISTA), Hammamet, Tunisia, 2023. P. 1–6.

19. *Rahal N., Tounsi M., Hussain A., Alimi A.M.* Deep Sparse Auto-Encoder Features Learning for Arabic Text Recognition // IEEE Access. 2021. Vol. 9. P. 18569–18584.

20. *Sultanov M.* Tatar hanlary (يارالنخ راتات) // Tipografiya bratev Karimvyh. 1911.
URL: https://darulkutub.com/uploads/books/820d9f6dcf1e868ee899d47e487b06189c2b816a.pdf

21.*Kamal Sh.* Akcharlaklar (آقچارالقلر) // "Vakyt" matbagasy. 1915. URL: https://miras.info/projects/mirasxane/books/425-akcharlaklar-.html

22.Aek bul (آيق بولا) // Tipo-litografiya Imperatorskogo universiteta. 1907. URL: https://darul-kutub.com/uploads/books/c3032b3c9136803dc0e38db69cd15541fb24064b.pdf

СВЕДЕНИЯ ОБ АВТОРЕ

ВАЛИШИН Искандер Айратович – выпускник магистратур Института информационных технологий и интеллектуальных систе Казанского (Приволжского) федерального университета.

Iskander Airatovich VALISHIN – Master's at the Institute of Info mation Technology and Intelligent Systems, Kazan Federal University.

email: iskander1998@list.ru

ORCID: 0009-0006-6891-031X

Материал поступил в редакцию 24 июня 2024 года