АННОТАЦИЯ

к диссертации Тойганбаевой Назгуль Абеновны на соискание стпени доктора философии (PhD) по образовательной программе 6D060200 Информатика на тему «Разработка средств и создание базы данных для оптического распознавания рукописного текста казахско-русского языка»

Актуальность темы исследования. Несмотря на то, что различные сферы перешли на цифровой обмен данными, многие документы, такие как анкеты, памятки, ответы на экзаменационные вопросы и архивные данные, по-прежнему требуют ручного ввода. В связи с этим необходимо реализовать распознавание рукописного текста – автоматический метод транскрипции рукописей с помощью компьютера. Распознавание рукописного текста (англ. Handwriting recognition, HWR или Handwritten Text Recognition, HTR) – это способность получать и распознавать понятную рукописную информацию из бумажных документов, изображений, сенсорных экранов и других источников.

Офлайн-распознавание рукописного текста – это метод преобразования букв или слов на изображении в цифровой текст, при котором на вход подается двумерное изображение, а на выходе получается последовательность символов. В настоящее время доступные программные решения для распознавания рукописного текста можно разделить на два основных типа: Онлайн-распознавание – включает распознавание текста, написанного на экране различных устройств с помощью специальных инструментов (стилусов). Офлайн-распознавание – включает автоматическое преобразование текста на изображении в буквенные коды, которые можно использовать в компьютерных и системах обработки текстовой информации. При этом задача офлайн-распознавания рукописного текста до конца ещё не решена. За последние 10 лет наблюдается прогресс в области распознавания печатного и рукописного текста. Несмотря на вычислительную сложность, ряд весьма сложных задач можно решить с помощью структур, разработанных для машинного обучения и глубоких нейронных сетей. Проблема распознавания кириллических рукописей на казахском и русском языках до сих пор изучена не в полной мере. Поэтому разработка алгоритмов офлайн-распознавания рукописного текста на казахском и русском языках является крайне важной, и в этой области необходимо вести исследования. Особенно это актуально, учитывая отсутствие наборов данных для рукописного текста на казахском языке. Таким образом, тема является современной и имеет большое практическое значение.

Распознавание рукописного текста – это сложный структурированный процесс, который включает сканирование бумаги с написанным текстом, разделение полученной информации на тексты и изображения, выбор методов для интеллектуального распознавания рукописей и обработку результатов. Для решения задач распознавания рукописного текста применяются нейронные сети. Основанные на биологической структуре человеческого

мозга, нейронные сети благодаря своей высокой вычислительной мощности считаются в несколько раз более эффективными, чем другие алгоритмы обучения.

В последние годы достигнут значительный прогресс в совершенствовании методов и систем распознавания, однако автоматическое распознавание рукописного текста на изображениях по-прежнему остаётся сложной задачей. Всесторонние и неограниченные наборы данных с образцами рукописей становятся как никогда важными. Рукописный текст встречается в памятках, медицинских записях, исторических документах, текстах, введённых с помощью стилуса, экзаменационных ответах и других документах. Поэтому необходимо обеспечить полное решение задачи распознавания рукописного текста на изображениях. Это подчёркивает необходимость проведения исследований в области разработки масштабных систем распознавания рукописей для нескольких языков. Таким образом, создание базы данных и разработка эффективных средств OCR для рукописного текста на казахском и русском языках является актуальной задачей, имеющей как научную, так и высокую практическую значимость.

Цель диссертационной работы – разработка инструментов для оптического распознавания рукописных текстов на казахском и русском языках с применением моделей глубокого обучения, а также создание соответствующей базы данных.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести обзор и анализ решений, использующих методы глубокого обучения для задач оптического распознавания текста.
2. Сформировать корпус данных рукописных текстов на казахском и русском языках.
3. Применить генетический алгоритм для реализации процесса сегментации, разделяющего текст на сканированных страницах на строки, а строки – на слова.
4. Провести эксперименты с использованием различных методов глубокого обучения на основе корпуса данных рукописных текстов на казахском и русском языках.

Объект исследования – процесс распознавания рукописных текстов на казахском и русском языках.

Предмет исследования – создание корпуса данных рукописных текстов на казахском и русском языках и применение моделей глубокого обучения для их распознавания.

Методы исследования – в качестве метода сегментации рукописных текстов используется генетический алгоритм; для решения задачи распознавания рукописных текстов применяются рекуррентные и сверточные нейронные сети.

Научная новизна полученных результатов:

- Разработана автоматизированная технологическая система для сбора корпуса рукописных текстов методом краудсорсинга с целью обучения нейронных моделей.
- Создан генетический алгоритм, адаптированный для точной и эффективной сегментации слов из рукописных текстов на казахском языке.
- Впервые сформирован структурированный корпус офлайн рукописных текстов на казахском языке на уровне слов – Kazakh Offline Handwritten Text Dataset (KOHTD).
- Нейронные архитектуры Flor, Puigcerver, Bluche и Scheidl, предназначенные для оптического распознавания рукописных текстов на казахском и русском языках, были адаптированы под казахский язык и обучены методом тонкой настройки (fine-tuning).

Теоретическая и практическая значимость работы. Цифровизация рукописных записей позволяет автоматизировать бизнес-процессы многих компаний и облегчить труд человека. Кроме того, распознавание рукописных текстов на казахском и русском языках способствует автоматизации работы почтовых центров. Созданный впервые на казахском языке корпус рукописных текстов KOHTD дает возможность оценивать и сравнивать производительность алгоритмов.

Основные положения, выносимые на защиту:

- Разработана и внедрена в экспериментальном порядке полностью автоматизированная система, выполняющая процессы сбора данных рукописных текстов, их предварительной обработки, аннотирования и включения в корпус.
- Адаптирован генетический алгоритм для точной сегментации слов в рукописных текстах на казахском языке, что позволило повысить точность определения межбуквенных границ за счет его эвристических возможностей.
- Создан открытый датасет Kazakh Offline Handwritten Text Dataset (KOHTD), содержащий офлайн рукописные тексты на казахском языке.
- Нейронные модели Flor, Puigcerver, Bluche и Scheidl, предназначенные для оптического распознавания рукописных текстов на казахском и русском языках, были адаптированы под особенности графемного состава казахского языка и обучены соответствующим образом.

Степень достоверности и результаты апробации. Результаты исследования по теме диссертации опубликованы в 5 статьях, а также получено свидетельство о внесении сведений в государственный реестр объектов авторского права. Результаты исследования были представлены на научных семинарах кафедры компьютерных наук Казахского национального университета имени аль-Фараби, а также на международных конференциях.

Публикация в журнале, входящем в базу Scopus:

1) Toiganbayeva N., Kasem M, Abdimanap G., Bostanbekov K., Abdallah A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset. Signal Processing: Image Communication. Elsevier, 2021, vol.108, pp. 116827. doi

https://doi.org/10.1016/j.image.2022.116827 (Scopus: Q2; CiteScore-2.5; Percentile- 73%))

Статьи в журналах, входящих в перечень научных изданий, рекомендуемых Комитетом по обеспечению качества в сфере науки и высшего образования Министерства науки и высшего образования Республики Казахстан, для публикации основных результатов научной работы:

2) Тойганбаева Н.А., Әбдіманап Ғ.С., Алимова А.Н., Жунусова Ж.Х. Қазақ тіліндегі қолжазба мәтіндерінің мәліметтер қорын құру мен тану. «Алматы энергетика және байланыс университетінің Хабаршысы», 3(62), 2023, 105-116б.

3) Тойганбаева Н.А., Алимова А.Н., Әбдіманап Ғ.С., СакыпбековаМ.Ж., Гусманова Ф.Р. Нейронды желілер негізінде қазақ-орыс тілдеріндегі қолжазба мәтіндерді тану. Абай атындағы ҚазҰПУ. Хабаршы. Физика-математика ғылымдары сериясы. Том 84 № 4 (2023). 183-191б.

4) Тойганбаева Н.А. Терең оқыту модельдері негізінде қазақ-орыс тіліндегі оффлайн қолжазба мәтіндерін тану. Қазақстан Республикасы Ұлттық инженерлік академиясының хабаршысы. № 4 (90) – 2023.

Международная конференция:

5) Тойганбаева Н.А., Тойганбаев Б.А. Нейронды желілерді қолжазбаларды интеллектуалды тану есептерінде қолдану мәселелеріне шолу // САТПАЕВ ОҚУЛАРЫ – 2021, Том II, 86-89б.

Свидетельство о внесении сведений в Государственный реестр прав на объекты, охраняемые авторским правом:

6) Нурсеитов Д.Б, Бостанбеков К.А. Әбдіманап Ғ.С., Тойганбаева Н.А., Алимова А.Н. «Kazakh Offline Handwritten Text Dataset (KOHTD)» Свидетельство о внесении сведений в Государственный реестр прав на объект – базу данных рукописных текстов на казахском языке, охраняемую авторским правом. № 34615, 2023 г.

Личный вклад исследователя. В соответствии с поставленными в диссертационной работе целями были решены соответствующие задачи. Выполнены работы по предварительной обработке данных для распознавания рукописных текстов и сегментации строк текста с использованием генетического алгоритма, осуществлено создание и наполнение базы данных KOHTD. Проведены эксперименты с моделями Flor, Puigcerver, Bluche и Scheidl, выполнена их оценка по стандартным показателям эффективности CER, WER и SER. Все модели были реализованы с использованием библиотек Python, TensorFlow, Keras, OpenMP и Matplotlib.

Объем и структура работы. Диссертация выполнена на казахском языке и состоит из введения, четырёх глав и заключения. Общий объём работы составляет 97 страниц, включает 42 рисунок, 10 таблиц и 95 наименования использованной литературы.

Во введении определена актуальность работы и обоснована необходимость проведения исследований в области разработки крупномасштабных систем распознавания рукописных текстов. Также

сформулированы цель, задачи и методы диссертационной работы, изложена научная новизна полученных результатов, теоретическая и практическая значимость работы, а также представлен перечень публикаций.

В первой главе выполнен обзор решений задач оптического распознавания рукописных текстов с использованием возможностей глубокого обучения. Проведен количественный анализ корпусов данных, применяемых в области распознавания рукописных текстов. Даны объемные, структурные и языковые характеристики наиболее распространённых корпусов данных, используемых в данной области. Проведённый анализ может служить основой для оценки эффективности и проведения сравнительных исследований систем HTR (Handwritten Text Recognition).

Во второй главе рассмотрен процесс создания первого корпуса офлайн рукописных текстов на казахском языке – Kazakh Offline Handwritten Text Dataset (KOHTD*)*, включая этапы сегментации текста и аннотирования. В данном разделе также приведён статистический анализ корпуса KOHTD.

В третьей главе представлены основные типы нейронных сетей, широко применяемых для распознавания рукописных текстов, и описаны принципы их работы. Рассмотрены архитектуры и возможности искусственных нейронных сетей, рекуррентных (RNN), сетей с долговременной и кратковременной памятью (LSTM) и сверточных нейронных сетей (CNN). Также приведены методы обучения нейронных сетей и объяснена роль алгоритма Connectionist Temporal Classification (CTC) в распознавании рукописных текстов.

В четвёртой главе рассмотрены метод Gated-CRNN, предложенный Bluche, метод 2-LSTM, предложенный Puigcerver, метод Gated-CNN, предложенный Flor, и метод Attention-Gated-CNN-BGRU, предложенный Abdallah. Кроме того, модели были обучены на корпусе KOHTD и оценены по стандартным показателям эффективности CER, WER и SER.

В заключении приведены основные результаты и выводы диссертационной работы.