Министерство образования Республики Беларусь

Учреждение образования

Белорусский государственный университет

информатики и радиоэлектроники

УДК 004.93'12

Калько

Алексей Игоревич

Система идентификации ручного почерка

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники

по специальности 1–40 81 01 – Информатика и технологии разработки программного обеспечения

_____

Научный руководитель

Наранович Оксана Ивановна

доцент, кандидат

физико–математических наук

_____

Минск 2018

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

В настоящее время существует множество текстовых документов в бумажном рукописном виде. Одним из способов организации человеко-машинного взаимодействия является передача компьютерной системе инструкций пользователя в формате рукописного текста.

Это связано с тем, что компьютеры появились сравнительно недавно, а история науки, так или иначе связанная с математикой, существует уже много веков. С развитием информационных технологий рано или поздно требуется перевод документов в электронный формат.

Уже существуют системы, которые с достаточно высокой точностью распознают обычные печатные тексты со сканированных изображений. Системы, распознающие и идентифицирующие рукописный текст, находятся в состоянии доработки готовых систем или разработки новых аналогов.

Для перевода таких текстов в электронный формат требуется использовать рукописный способ ввода текста. Если текст содержит много символов, то эта работа является достаточно трудоемкой. Учитывая то, что большинству пользователей с трудом удается вводить большие объемы текста с клавиатуры, то для этой работы нужны специально подготовленные люди.

С постепенной автоматизацией многих процессов в различных сферах, более приоритетным является обучение компьютера идентификации рукописного текста, а не подготовка специально обученных людей. Ввод рукописного текста в компьютер является достаточно сложной задачей, поэтому в качестве основной цели рассматривается возможность автоматического распознавания сканированных рукописных текстов.

Таким образом, задача распознавания и идентификации рукописных текстов при помощи нейронных сетей и математического аппарата в настоящее время является актуальной.

Основная сложность автоматической идентификации рукописного текста состоит в сложности их иерархической структуры и их большом разнообразии. Для анализа сложной структуры предлагается использование процедур, на основе которых будет построена модель изображения текста. Так как все люди имеют уникальный почерк, то каждый символ имеет множество вариантов написания. Кроме того, распечатанные на бумаге тексты со временем могут потерять свое качество из-за влияния внешних факторов. Следовательно, сканированное изображение будет иметь некоторые дефекты, связанные с появлением шумов или потерей фрагментов

текста. В связи с этим при распознавании отдельных символов могут быть получены некоторые неопределенности, т.е. такие ситуации, в которых однозначно нельзя определить класс символа, к которому относится та или иная компонента изображения.

Для выбора наиболее правильного варианта в данном случае предлагается использовать специальные текстовые шаблоны, называемые регулярными выражениями.

Степень разработанности объекта.

Исследования в области анализа и распознавания текстов рассматривали в своих работах H.S. Baird, H. Saiga, N. Matsakis, K.-F. Chan, D.-Y. Yeung, M. Suzuki, Садыков С.С., Самандров И.Р., Салюм Саид Салех, A. Belaid, J.P. Haton, R.H. Anderson, Фу К.С., Исупов Н.С., Кучуганов А.В., Костюк Ю.Л. и др.

В настоящее время существуют следующие приложения, идентифицирующие рукописный текст и рукописные формулы: Math input panel, InftyReader, GOCR. Math input panel является системой динамического распознавания рукописных формул, поэтому не может быть применена к обработке сканированных текстов. InftyReader и GOCR предназначены для распознавания печатных текстов с растровых изображений. Требуемое качество не достигается при сканировании изображений, поэтому точность распознавания не высокая. Ни одна из существующих систем не идентифицирует сканированные рукописные тексты.

Объектом исследования являются растровые изображения текстов, с содержащимся в них рукописным текстом.

Предметом исследования являются методы анализа изображений и построения моделей изображений с рукописным текстом.

Цель работы — разработка и исследование средств анализа изображений со сложно структурированными объектами и автоматической идентификации рукописных текстов со сканированных изображений, в которых одновременно могут находиться и обычный печатный текст.

Для достижения этой цели требуется решить следующие основные задачи:

1. Выполнить подробный анализ проблем, возникающих при обработке изображений со сложно структурированными объектами, с целью выявления возможных вариантов их решения. Исследовать существующие системы, решающие задачи, связанные с распознаванием текстов.

2. Предложить способ анализа графических изображений со сложной иерархической структурой на основе метода секущих плоскостей. Построить

модель изображения символов. Разработать алгоритм идентификации иерархической структуры изображения рукописного символа.

3. Предложить способ использования регулярных выражений для уточнения результатов обработки текстов при наличии в них неопределенностей на основе использования нейронных сетей. Разработать алгоритм нахождения пересечения множеств, заданных регулярными выражениями. Рассмотреть варианты адаптации алгоритма к различным условиям с целью повышения точности распознавания.

4. Произвести экспериментальные исследования предложенных алгоритмов.

Методы исследования. Теоретические исследования выполнены с использованием методов теории множеств (множество пикселей в изображении, множество блоков в сегментации), теории алгоритмов (алгоритмы бинаризации, сегментации, утонышения контуров текста) и теории искусственного интеллекта. При решении практической части использовалась технология объектно-ориентированного программирования.

Достоверность и обоснованность полученных в работе результатов подтверждается корректностью построенных моделей с использованием регулярных выражений, положительными результатами при выполнении экспериментальных исследований.

Научная новизна.

1. Предложен способ описания изображений сложно структурированных текстов. Для этого используется понятие двумерно ориентированного графа, в котором ребра могут задавать не только общую информацию о наличии связи между вершинами, но и характер этой связи с точки зрения направления на плоскости. Кроме того, направление может задаваться сразу целым набором двумерных векторов, что позволяет более детально описывать траекторию движения от одной вершины к другой.

2. На основе метода Оцу модифицирован граф алгоритма бинаризации изображения, содержащего рукописный текст.

3. Предложен способ задания двух вариантов распознавания текста, который позволяет учитывать вероятности исходов идентификации как всего текста, так и отдельных фрагментов и символов. Для этого вводится понятие взвешенных регулярных выражений, порождающих взвешенные регулярные множества, отличающиеся от обычных регулярных множеств тем, что каждому элементу задается некоторый числовой вес, отражающий превосходство одних элементов над другими. Результат распознавания задается взвешенным регулярным выражением.

4. Разработан алгоритм на основе использования нейронных сетей для нахождения пересечения множеств, заданных регулярными выражениями, который применяется для проверки соответствия результата распознавания используемому образу. Предложенный алгоритм позволяет использовать словарь с бесконечным числом элементов, который задается в виде регулярного выражения.

Результаты исследования опубликованы в 8 научных работах и докладывались на конференциях:

− международная научно-практическая конференция «Экономика, технологии и право в современном мире», г. Барановичи;

− международная научно-практическая конференция «Техника и технологии: инновации и качество», г. Барановичи.

Результаты апробированы и внедрены в образовательный процесс кафедры государственного управления и уголовно-правовых дисциплин Барановичского государственного университета с ноября 2016 г.