МЕТОДЫ РАСПОЗНАВАНИЯ ТЕКСТОВ

***Вовк А.В.**, доцент, кафедра МСТ, ХНУРЕ*
***Сиденко Д.А.**, магистр, кафедра МСТ, ХНУРЕ*

Несмотря на то, что в настоящее время большинство документов составляется на компьютерах, задача создания полностью электронного документооборота ещё далека до полной реализации. Как правило, существующие системы охватывают деятельность отдельных организаций, а обмен данными между организациями осуществляется с помощью традиционных бумажных документов.

Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид.

Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл. Более предпочтительным, по сравнению с графическим, является текстовое представление информации. Этот вариант позволяет существенно сократить затраты на хранение и передачу информации, а также позволяет реализовать все возможные сценарии использования и анализа электронных документов. Поэтому наибольший интерес с практической точки зрения представляет именно перевод бумажных носителей в текстовый электронный документ [1-4].

На вход системы распознавания поступает растровое изображение страницы документа. Для работы алгоритмов распознавания желательно, чтобы поступающее на вход изображение было как можно более высокого качества. Если изображение зашумлено, нерезко, имеет низкую контрастность, то это усложнит задачу алгоритмов распознавания. Поэтому перед обработкой изображения алгоритмами распознавания проводится его предварительная обработка, направленная на улучшение качества изображения [5]. Она включает фильтрацию изображения от шумов, повышение резкости и контрастности изображения, выравнивание и преобразование в используемый системой формат (в нашем случае 8-битное изображение в градациях серого).

Подготовленное изображение попадает на вход модуля сегментации. Задачей этого модуля является выявление структурных единиц текста – строк, слов и символов. Выделение фрагментов высоких уровней, таких как строки и слова, может быть осуществлено на основе анализа промежутков между тёмными областями.

К сожалению, такой подход не может быть применён для выделения отдельных букв, поскольку, в силу особенностей начертания или искажений, изображения соседних букв могут объединяться в одну компоненту связанности или наоборот – изображение одной буквы может распадаться на

отдельные компоненты связанности. Во многих случаях для решения задачи сегментации на уровне букв используются сложные эвристические алгоритмы.

Полагаем, что для принятия окончательного решения о прохождении границы букв на таком раннем этапе обработки, системе распознавания недостаточно информации. Поэтому задачей модуля сегментации на уровне букв в разработанном алгоритме является нахождение возможных границ символов внутри буквы, а окончательное решение о разбиении слова принимается на последнем этапе обработки, с учётом идентификации отдельных фрагментов изображения как букв. Дополнительным преимуществом такого подхода является возможность работы с начертаниями букв, состоящих из нескольких компонент связанности без специальной обработки таких случаев.

Результатом работы модуля сегментации является дерево сегментации – структура данных, организация которой отражает структуру текста на странице. Самому верхнему уровню соответствует объект страница. Он содержит массив объектов, описывающих строки. Каждая строка в свою очередь включает набор объектов слов. Слова являются листьями этого дерева. Информация о возможных местах разделения слова на буквы храниться в слове, однако отдельные объекты для букв не выделяются. В каждом объекте дерева хранится информация об области, занимаемой соответствующим объектом на изображении. Данная структура легко может быть расширена для поддержки других уровней разбиения, например колонок, таблиц.

Выявленные фрагменты изображения подаются на вход классификатора, выходом которого является вектор возможности принадлежности изображения к классу той или иной буквы. В разработанном алгоритме используется классификатор составной архитектуры, организованный в виде дерева, листьями которого являются простые классификаторы, а внутренние узлы соответствуют операциям комбинирования результатов низлежащих уровней.

Результатом работы классификатора является нечёткое множество, полученное в результате комбинирования на самом верхнем уровне. На последнем этапе принимается решение о наиболее правдоподобном варианте прочтения слова. Для этого используются уровни возможности прочтения отдельных букв, межбуквенной сегментации и частоты сочетаний букв в русском языке.

Список литературы

1. Арлазаров В.Л., Куратов П.А., Славин О.А. Распознавание строк печатных текстов // Методы и средства работы с документами: Сб. трудов ИСА РАН. С. 31-51.

2. Квасников В.П., Дзюбаненко А.В. Улучшение визуального качества цифрового изображения путем поэлементного преобразования // Авиационно-космическая техника и технология. 2009. № 8. С. 200-204.

3. Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms / Melin P., Urias J., Solano D., Soto M., Lopez M., Castillo O. // Engineering Letters. 2006. № 13:2.

4. Выбор признаков для распознавания печатных кириллических символов / Багрова И.А., Грицай А.А., Сорокин С.В., Пономарев С.А., Сытник Д.А. // Вестник Тверского Государственного Университета. 2010. № 28. С. 59-73.

5. Orobinskyi P., Deineko Z., Lyashenko V. Comparative Characteristics of Filtration Methods in the Processing of Medical Images // American Journal of Engineering Research. 2020. Vol. 9(4). P. 20-25.