УДК 004.91

Быстрый метод распознавания текста официальных документов

Мазайшвили Е.К. (ИТМО)
Научный руководитель – кандидат педагогических наук, доцент Авксентьева Е.Ю. (ИТМО)

Введение. В юридических документах РФ, как правило, используются один из двух шрифтов: Arial и Times New Roman, а также междустрочный интервал и отступы, определяемые ГОСТ Р 7.0.97-2016 [1]. Для преобразования печатного документа в цифровой формат используются технологии распознавания текста — Optical Character Recognition (далее — OCR). Однако, скорость распознавания у большинства систем может быть не такой высокой, за счет того, что они пытаются распознать обширное множество шрифтов и сканируют весь лист [2]. В статье предложен метод быстрого оптического распознавания текста таких документов с использованием набора ключевых областей символов известных шрифтов. Также сканируется не весь лист, а лишь области, в которых возможно появление текста при использовании этих отступов и шрифтов.

Основная часть. Предлагаемый метод опирается на факт, что документ будет оформлен по ГОСТ Р 7.0.97-2016, а, значит, будет использовать шрифт Arial 12 или Times New Roman 13, абзацный отступ 1,25 см, левое поле — 20 мм, правое — 10 мм, верхнее — 20 мм, нижнее — 20 мм
Исходя из этих знаний, можно использовать алгоритм на основе сравнения с шаблоном [3], перемещая скользящее окно точно по местам возможного нахождения текста по ГОСТу, а не по всему листу. Также для ускорения работы предлагается для распознавания использовать модификацию метода сравнения с шаблоном.

Предлагаемый метод принимает на вход изображение сканированного документа и состоит из трех шагов.

На первом шаге происходит определение границ страницы и коррекция ее ориентации, если лист был сканирован с наклоном. Это реализуется с помощью нахождения краев белого листа на скане или с помощью нанесения специальных меток на страницу.

На втором шаге по всем возможным местам нахождения текста на изображении двигается скользящее окно, внутри которого перебираются шаблоны букв. Шаблоны состоят из набора ключевых областей внутри рамки, которые должны быть белые или черные.

Все ключевые области просчитаны заранее и хранятся в памяти программы. В этой статье не приводятся изображения всех областей для всех символов, но области могут быть определены алгоритмически. Каждый шрифт имеет свои, незначительно отличающиеся ключевые области. Определение используемого шрифта выходит за рамки этой статьи.

Алгоритм создания ключевых областей:

1. Создать bounding box (ограничивающую прямоугольную рамку), по размеру самой крупной заглавной буквы, расширенная снизу до вертикального размера самого длинного нижнего выносного элемента — нижней части прописной буквы «у».

2. Найти все белые области и все черные области. Например, с помощью алгоритма заливки.

3. Каждую черную область разделить на части любым алгоритмом тесселяции. Цель — сделать каждую часть близкой к 1/4 буквы. Некоторые буквы, например, знаки препинания, можно разбить только на 1 часть.

4. Центр тяжести каждой черной части, плюс 3-5 пикселей (зависит от размера части) в каждую сторону образуют области, которые должны быть черными на сканируемом изображении этого символа.

5. Центр тяжести каждой белой части, плюс 3-5 пикселей в каждую сторону образуют области, которые должны быть белыми на сканируемом изображении этого символа.

6. Области каждого символа сравниваются с областями всех остальных символов следующим алгоритмом несколько раз, пока результат не перестанет меняться:

    a. Если белая область какого-то символа, наложенная на все остальные символы дает белый цвет, эта область удаляется из всех символов.

    b. Если черная область какого-то символа совпадает с белой областью другого символа, в другом символе появляется дополнительная белая область в этом месте.

На третьем шаге распознанные символы объединяются в текст, который и является результатом работы программы.

Выводы. Предлагаемый метод обладает преимуществом в скорости распознавания текста, а также в устойчивости к плохо пропечатанным краям текста, что часто бывает у лазерных и струйных принтеров.

Список использованных источников:

1. ГОСТ Р 7.0.97-2016. Система стандартов по информации, библиотечному и издательскому делу. Организационно-распорядительная документация. Требования к оформлению документов.

2. Койнова, Т. А. Алгоритмы распознавания символов / Т. А. Койнова. — Текст : непосредственный // Молодой ученый. — 2022. — № 18 (413). — С. 73-76. — URL: https://moluch.ru/archive/413/91060/ (дата обращения: 16.02.2025)

3. Бербасов В. Д. ПРЕОБРАЗОВАНИЕ ИНФОГРАФИКИ В ДАННЫЕ // Экономика и социум. 2023. №8 (111). URL: https://cyberleninka.ru/article/n/preobrazovanie-infografiki-v-dannye (дата обращения: 18.02.2025).