Реализация и сравнительный анализ алгоритмов распознавания текста

Талалайкина Е.И.[1], Огнева М.В.[2], Лаптев Ю.В.[3]
[1]talalaykinaei@gmail.com, [2]ognevamv@mail.ru, [3]laptev.iurij2016@yandex.ru
Саратовский государственный университет имени Н.Г. Чернышевского

Данная статья описывает сравнительный анализ трех реализованных алгоритмов распознавания текста, проведенного для выбора алгоритма, который будет использован в мобильном приложении-сканере с функцией распознавания текста на сложном фоне. Такие приложения являются удобной альтернативой настольному сканеру для школьников и студентов, так как упрощают и удешевляют процесс сканирования.

Ключевые слова: оптическое распознавание символов, мобильное сканирование, контурный анализ, алгоритм масштабного пространства кривизны, нейросеть.

В современном мире перевод информации, представленной на бумажных носителях, в электронный вид является задачей, актуальной не только для отдельно взятых пользователей, но и целых ведомств и организаций, в том числе образовательных. Высокая скорость обработки, возможность интеграции с внешними ресурсами, быстрота документооборота между организациями, а также между организациями и пользователями (например, отправка документов в образовательные организации учащимися), возможность создания резервных копий – все это является достоинствами электронного представления документов. Высокая скорость коммуникации между организациями и пользователями.

Переход к цифровому документообороту требует инструмента, который бы минимизировал усилия и временные затраты на перепечатывание, а также помогал исключить ошибки, связанные с необходимостью ручной оцифровки бумажных носителей. Таким инструментом является сканер.

Однако необходимость наличия сканера создает некоторые трудности процессу в виде покупки, размещения и настройки устройства. Кроме того, сканеры привязаны к формату документов – на устройстве, подходящем для сканирования документов формата А4, не получится обработать формат А3.

Альтернативой привычному настольному устройству может стать приложение для сканирования на смартфоне. Такое приложение позволяет сфотографировать документ, после чего автоматически определяет его границы, настраивает резкость и четкость изображения. Пользователь может создавать многостраничные документы в разных форматах и сразу отправлять их по почте или через мессенджеры. Неоспоримыми достоинствами приложения-сканера являются компактность и мобильность технологии – «сканирование» возможно где угодно, необходим только смартфон. Также для приложения не важен формат листа исходного документа.

Такие приложения очень удобны для школьников и студентов особенно в условиях дистанционной учебы, так как работают на любом современном смартфоне, позволяют сделать быстрым и легким процесс получения скан-копий домашних заданий и конспектов для отправки на проверку и исключают необходимость наличия настольного сканера для этого.

Сканирование предполагает не только получение «фотографии» текста, но и его перевод в редактируемый формат с возможностью поиска – иными словами, оптическое распознавание символов (англ. Optical Character Recognition – OCR).

Точное распознавание латинских символов в печатном тексте в настоящее время возможно для изображений в хорошем качестве, например, сканированных печатных документов, набранных стандартным шрифтом. Распознавание текста на изображениях с неоднородным фоном является более сложной задачей, так как фон затрудняет нахождение текстовых областей. Сложное форматирование документа также является причиной ошибок при работе систем распознавания текста.

Нами были рассмотрены 14 представленных на рынке мобильных приложений для сканирования документов, имеющих функцию распознавания текста. Только три приложения из рассмотренных дали приемлемый результат для документа с неоднородным фоном. Но ограниченный функционал и очень короткий пробный период бесплатных версий этих приложений делает их фактически бесполезными с точки зрения постоянного практического использования.

Рис.1 Изображение, на котором проверялся функционал распознавания текста представленных на рынке приложений

Таким образом, распознавание текста в документах со сложным фоном и форматированием на сегодняшний день является нерешенной задачей в области мобильного сканирования.

В целом реализация системы распознавания текста зависит от целей ее применения и устройств, для которых она создается. Но центральным звеном любой такой системы является алгоритм распознавания текста.

Для создания собственного приложения-сканера для смартфона с функцией распознавания текста, которое поддерживало бы распознавание для документов со сложным фоном и/или форматированием, были реализованы и протестированы три алгоритма.

Для реализации первого алгоритма распознавания текста был использован метод сравнения моментов контуров букв в тексте с моментами заранее подготовленных шаблонов букв.

Любой объект на изображении имеет границы, которые человек воспринимает как резкий перепад яркости между двумя областями, – контуры. Контурный анализ – метод описания, хранения, распознавания, сравнения и поиска объектов на изображении.

Этот метод имеет слабую устойчивость к помехам, а любое пересечение объектов приводит либо к невозможности детектирования, либо к неправильным результатам, но простота и быстродействие контурного анализа позволяют успешно применять данный подход (при чётко выраженном объекте на контрастном фоне и отсутствии помех).

Для сравнения двух контуров предварительно рассчитываются их моменты. Момент – это характеристика контура, объединённая (суммированная) со всеми пикселями контура [1].

Как уже отмечалось выше, данный алгоритм чувствителен к ошибкам сегментации: слияние букв или разрывы в контурах приводят к неправильным

результатам распознавания. Также из-за неправильной сегментации ошибочным будет результат распознавания слов с буквами, состоящими из нескольких частей, такими как «ё», «й», «ы». Ошибки сегментации возникают и при распознавании изображений с низкой контрастностью фона и текста, становясь причинами дальнейших ошибок в распознавании.

Второй из реализованных вариантов алгоритма распознавания текста основан на алгоритме масштабного пространства кривизны (culvature scale space, CSS), который разглаживает контур буквы с помощью Гауссовой функции ядра и отслеживает её точки перегиба [2]. Для расчета кривизны кривых используются вейвлеты Гаусса. На вход для обработки подается текстовая область, представляющая собой текст на сложном фоне, а на выходе генерируется распознанный текст [3].

Для решения проблемы ошибочного распознавания составных букв была реализована функция постобработки ответа, но к ошибкам сегментации по-прежнему приводила низкая контрастность фона и текста.

Третий реализованный алгоритм распознавания текста основан на нейросетях типа «долговременная-кратковременная память» (англ. Long short-term memory, LSTM), представленных в библиотеке распознавания текста с открытым исходным кодом Tesseract [4]. Tesseract плохо подходит для случаев, когда изображение сильно зашумлено или имеет множество объектов, а не только чистый предварительно обработанный текст, поэтому для данного алгоритма была реализована система предобработки изображения и обнаружения текста, но она справилась не со всеми случаями низкой контрастности фона и текста.

Для тестирования реализованных алгоритмов был создан набор из 120 изображений. Результаты тестирования представлены в таблице.

Таблица – Результаты тестирования реализованных алгоритмов

	Текст	черный (контрастный)				цветной (неконтрастный)		
	Фон	белый	цветной монотонный	цветной с плавными переходами	цветной с резкими переходами	монотонный (белый и цветной)	цветной с плавными переходами	цветной с резкими переходами
	Общее количество изображений	20	20	20	20	20	20	20
Алг. 1	Количество ошибочно распознанных изображений	6	7	8	6	12	20	20
	Процент ошибок	30%	35%	40%	30%	60%	100%	100%
Алг. 2	Количество ошибочно распознанных изображений	0	0	0	0	7	20	20
	Процент ошибок	0%	0%	0%	0%	35%	100%	100%

Алг. 3	Количество ошибочно распознанных изображений	0	0	0	0	4	15	17
	Процент ошибок	0%	0%	0%	0%	20%	75%	85%

Из полученных результатов тестирования реализованных алгоритмов можно сделать ряд выводов.

Второй и третий реализованные алгоритмы отлично справились с распознаванием черного текста на белом и цветном фоне. Однако стоит отметить, что при использовании алгоритма CSS некоторые буквы в словах при распознавании поменяли регистр. Причиной этому является идентичное CSS-изображение между малым и большим регистром.

Первый реализованный алгоритм такой особенности не имеет, поэтому он более точно находит некоторые буквы, но при этом слова, содержащие буквы «ё», «й» и «ы», он распознает с ошибками.

Алгоритм на основе Tesseract OCR не только дает отличные результаты распознавания в случае черного текста на белом и цветном фоне, в том числе при наличии букв, состоящих из нескольких частей, таких как «ё», «й», «ы», но и показывает более качественные результаты в случае изображений с низкой контрастностью текста и фона.

Из проведенного исследования сделан вывод использовать в дальнейшей работе (разработке приложения-сканера) алгоритм распознавания текста на основе Tesseract OCR, доработав процесс отделения текста от фона, чтобы повысить качество распознавания в случае низкой контрастности текста и фона.

Список литературы

[1] *Flusser, J., Suk, T., Zitová, Б.* Moments and Moment Invariants in Pattern Recognition / J. Flusser, T. Suk, Б. Zitová. – New York: John Wiley & Sons Ltd, 2009. – 321 p. – P. 6.

[2] *Kopf S., Haenselmann T., Effelsberg, W.* Enhancing curvature scale space features for robust shape classifi-cation / S. Kopf, T. Haenselmann, W. Effelsberg // IEEE International Conference, 2005.

[3] *Балахонцева А., Гободоба А., Нгуен Т.* Система распознавания символов на изображениях со сложным фоном / А. Балахонцева, А. Гободоба, Т. Нгуен // The 23rd International Conference on Computer Graphics and Vision. – 2013.

[4] tesseract-ocr GitHub [Электронный ресурс]. URL: https://github.com/tesseract-ocr/ (Дата обращения 15.09.2021).