Способы распознавания текстов на разных изображениях

Х. А. Примова, email: xolida_primova@mail.ru
Р. Т. Рахимов, email: rustamjonraximov@gmail.com

Самаркандский филиал Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий

Аннотация. *В последние годы извлечение текста из изображений документов является одной из наиболее широко изучаемых тем в области анализа изображений и оптического распознавания символов. Эти извлечения изображений документов можно использовать для анализа документов, анализа содержимого, поиска документов и многого другого. Для извлечения документа из изображений мы использовали очень простой подход, основанный на алгоритме FAST. Сначала мы разделили изображение на блоки и проверили их плотность в каждом блоке. Более плотные блоки были помечены как текстовые блоки, а менее плотные - область изображения или шум. Затем мы проверяем возможность соединения блоков, чтобы сгруппировать блоки, чтобы текстовую часть можно было изолировать от изображения. Этот метод очень быстрый и универсальный, его можно использовать для обнаружения различных языков, почерка и даже изображений с большим количеством шума и размытия. В заключение, этот метод помогает более точно и менее сложно обнаруживать текст на изображениях документа.*

Ключевые слова: *FAST (Features from accelerated segment test),многоязычные документы, рукописные документы.*

Введение

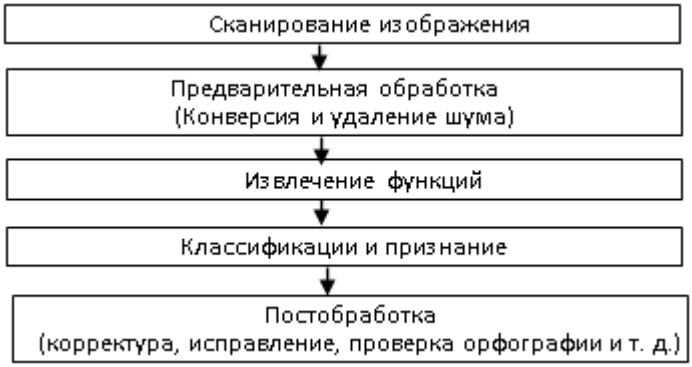
В последние годы появилась тенденция к оцифровке документов. С цифровизацией мира бумажные документы необходимо преобразовать в цифровые, чтобы сделать их более удобными, доступными для поиска и для сохранения документов. Для этого используется оптическое распознавание символов. Оптическое распознавание текста можно описать как механическое или электронное преобразование сканированных изображений, при котором изображения могут быть рукописными, машинописными или печатными [2]. Более полувека ведутся исследования в этой области, и уровень распознавания символов в современном OPC превышает 99% для высококачественных документов и 90% для рукописных документов. Для устаревших

документов и книг эффективность ОРС снижается до 80%. В последнее время для извлечения текста из изображений документов использовалось множество методов. Здесь мы будем использовать очень простой подход, основанный на алгоритме точки FAST. Во-первых, мы разделяем изображение документа на более мелкие неперекрывающиеся блоки фиксированного размера.

Оптическое распознавание символов(ОРС)

Оформление работы должны строго выполняться с использованием только стилей, которые представлены в данном шаблоне. Развитие распознавания символов в последнее десятилетие является значительным, а методы обнаружения символов обширны. Достижения в области распознавания символов очевидны в оптическом распознавании символов (ОРС), классификации документов, компьютерном зрении, интеллектуальном анализе данных, распознавании форм и биометрической аутентификации [2]. Распознавание символов - это процесс классификации входящего символа по предопределенному классу символов [1]. Распознавание символов находит свое применение в идентификации текста на изображениях. Текст может быть сканированным документом или рукописным текстом.

В последние годы появилась тенденция к оцифровке документов. С цифровизацией мира бумажные документы должны быть преобразованы в цифровые для более удобного использования, поиска и сохранения документов. Для этого используется оптическое распознавание символов. Оптическое распознавание текста можно описать как механическое или электронное преобразование сканированных изображений, при котором изображения могут быть рукописными, машинописными или печатными [2]. Более полувека ведутся исследования в этой области, и уровень распознавания символов в современном ОРС превышает 99% для высококачественных документов и 90% для рукописных документов. Для устаревших документов и книг эффективность ОРС снижается до 80%. В последнее время многие организации полагаются на ОРС для повышения производительности и эффективности. ОРС может выполняться офлайн и / или онлайн. Распознавание в режиме онлайн: процессор ОРС распознает символы по мере их ввода. В автономном режиме процессор может распознавать как документ, так и рукописные символы, но распознавание в автономном режиме сильно зависит от качества отсканированных изображений [6].



OPC состоит из многих этапов, таких как сканирование изображения, предварительная обработка, сегментация, извлечение признаков, классификация и распознавание, постобработка. Задача предварительной обработки связана с удалением шума и вариаций на изображении [3]. На этапе сканирования изображение получается. Качество изображения сильно зависит от используемого сканера. В практических приложениях отсканированные изображения не идеальны, из-за некоторых ненужных деталей в изображении может присутствовать некоторый шум, который может вызвать нарушение обнаружения символов на изображении. После завершения процесса OPC необходимо выполнить несколько шагов постобработки в зависимости от приложения, например тегирование документов метаданными (автор, год и т. д.) или проверка документов для исправления ошибок распознавания текста и орфографических ошибок [4].

2. Предлагаемые работы

В предложенном подходе для извлечения документа из изображений мы использовали очень простой алгоритм FAST. Сначала мы разделили изображение на блоки и проверили их плотность в каждом блоке. Более плотные блоки были помечены как текстовые блоки, а менее плотные - область изображения или шум. Затем мы проверяем связность блоков, чтобы сгруппировать блоки, чтобы текстовая часть могла быть изолирована от изображения. Этот метод очень быстрый и универсальный, его можно использовать для обнаружения различных языков, почерка и даже изображений с большим количеством шумов. и размытие. В заключение, этот метод

помогает более точно и менее сложно обнаруживать текст на изображениях документа.

Блок-схема на рисунке 2 показывает этапы предлагаемого подхода. Подробности шагов приведены ниже.

Шаг 1: изображение сканируется и преобразуется в оттенки серого. Изображение в градациях серого может быть преобразовано в двоичное изображение. Этот процесс называется оцифровкой изображения (бинаризация). Шум возникает из-за сканера. В проекте мы использовали фильтр Гаусса. Фильтрация Гаусса используется для размытия изображений, удаления шума и удаления нежелательных деталей на изображении [8] [9].

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^4 + y^4}{2\sigma^2}}$$

Шаг 2: угловые точки определяются алгоритмом FAST [5].

Шаг 3: разделите изображение на неперекрывающиеся блоки и рассчитайте количество угловых точек

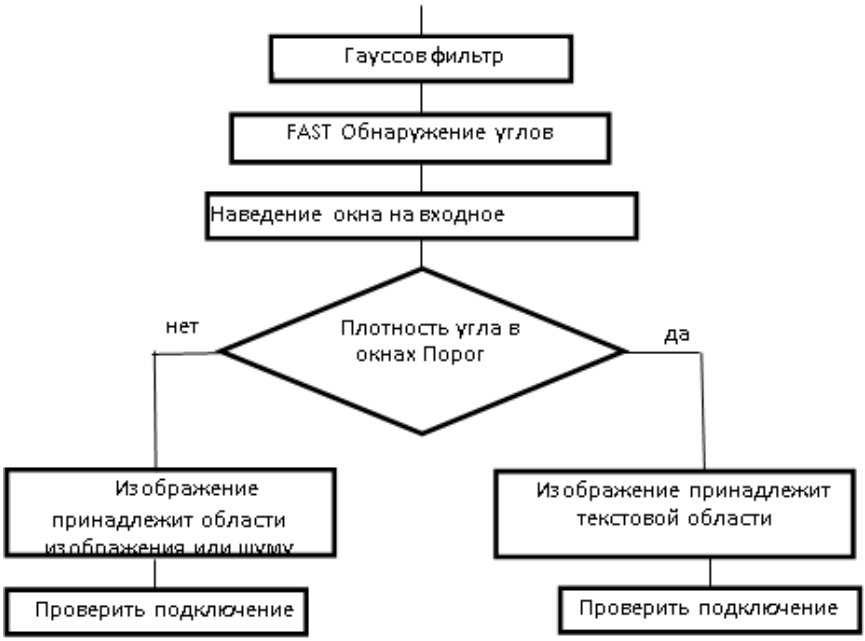


Рис. 2. Блок-схема алгоритма

Шаг 4: из блока найдите блок, который имеет максимальное количество угловых точек ($N\max$), определите порог, используя выбранный блок, порог используется как $T = 0.2 * N\max$. (20% от максимального значения).

Шаг 5: разделите блоки, имеющие большее количество углов, чем пороговое значение, принадлежащие текстовым областям, а блоки, имеющие меньшее пороговое значение, принадлежат области изображения или фона.

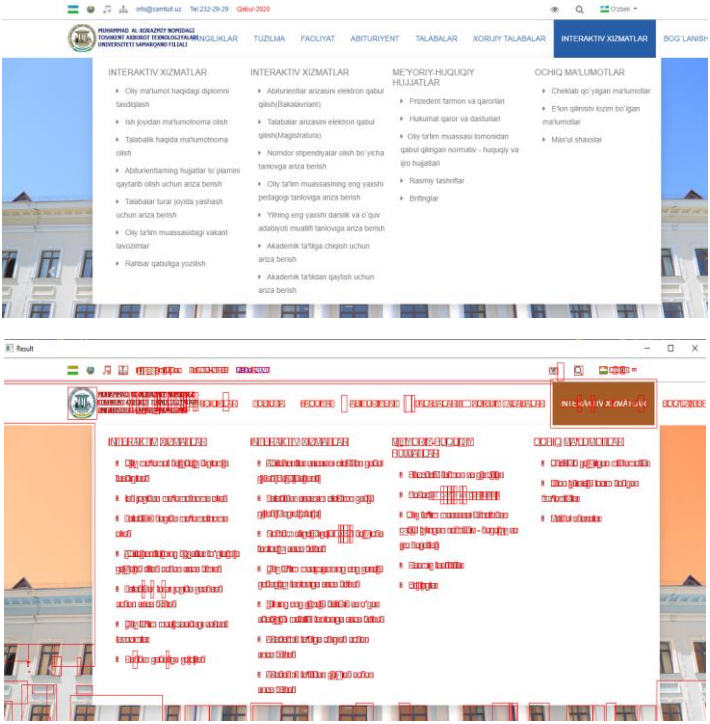


Рис. 3. Изображение узбекского документа (Исходное изображение, Обнаружение текстового изображения)

Шаг 6: после обнаружения текстовых блоков из угловой точки, проверьте возможность соединения этих блоков (8-связанных областей), чтобы восстановить текстовые области [9].

Экспериментальный результат

Формулы это простой метод с точностью и отзывчивостью более 90%, а часто в среднем 95%. Однако этот метод не очень эффективен для шрифтов большого размера, а также для некоторых конкретных изображений, для которых углы слишком сильно реагируют.

Несмотря на эти проблемы, он быстрый (и его можно распараллеливать) и менее сложный по сравнению с другими инструментами OPC, и в будущем его можно будет улучшить.

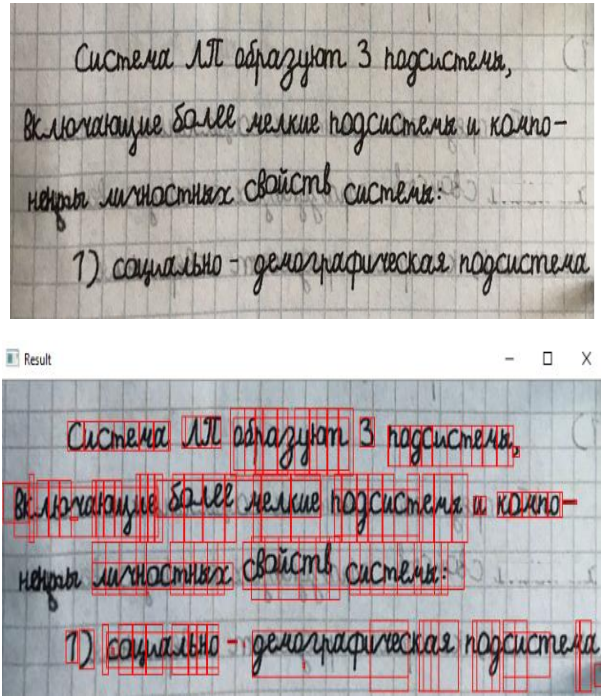


Рис. 4. Русское рукописное изображение (Исходное изображение, Обнаружение текстового изображения)

Заключение

При таком подходе мы увидели, что с помощью угловых точек на изображениях документов любого качества, ориентации или рукописных можно очень просто получить точное извлечение текста с низкими затратами и без особых параметров. Для извлечения текста из изображений мы используем очень простой подход, основанный на алгоритме FAST. Сначала мы разделили изображение на блоки и проверили их плотность в каждом блоке. Более плотные блоки были

помечены как текстовые блоки, а менее плотные - область изображения или шум. Затем мы проверяем возможность соединения блоков, чтобы сгруппировать блоки, чтобы текстовую часть можно было изолировать от изображения.

Список литературы

2. Ghai, D. Text Extraction from Document Images- A Review/ D.Ghai, N.Jain// International Journal of Computer Applications (0975 – 8887) , Volume 84 – No 3, -December, -2013, pp. 40- 48.

3. Automatic license plate recognition using extracted features / Nauman Saleem [et. al] // In 4th International Symposium on Computational and Business Intelligence,September 5-7, 2016. – Olten, Switzerland, 2016. – pp. 221-225.

4. Verma, R. A Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques / R.Verma, J.Ali // International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10. – October, 2013. – pp. 617-622.

5. Pansare, S. A Survey on Optical Character Recognition Techniques / S.Pansare, Dh.Joshi // International Journal of Science and Research (IJSR), Volume 3 Issue 12, December 2014, pp.1247-1249.

6. Viswanathan, Features from Accelerated Segment Test (FAST). / Viswanathan, Deepak Geetha//-2009, pp. 1-5.

7. Wang, Yao. Image Filtering: Noise Removal, Sharpening, Deblurring / Yao Wang,//EE 3414 Multimedia Communication Systems, Polytechnic University, Brooklyn, NY11201,

8. Primova, H. A. Computing fuzzy integral of the basis of fuzzy mesure/ H. A. Primova, D. M. Sotvoldiyev, R. T. Raximov, X. Bobabekova// Conference Series 1441 (1), 2020, doi: http://dx.doi.org/10.1088/1742-6596/1441/1/012161.

9. Shalin A. Optical Character Recognition/ Shalin A.[et, al] // International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 3, Issue 1, January, -2014, pp. 4956-4958.

10. Junga, K. Text information extraction in images and video: a survey / K.Junga, K.I. Kim, A. K. Jain // Pattern Recognition, 37, pp. 977-997, 2004.