СОВРЕМЕННЫЕ
ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ.
Том 4 (2021). С. 26–33

УДК 004.432.2

ОБЗОР СРЕДСТВ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕКСТА НА ПРИМЕРЕ ЗАДАЧИ ОЦИФРОВКИ РУССКОЯЗЫЧНЫХ АРХИВНЫХ ЖУРНАЛОВ

© 2021 г. **А. А. КАПРАНЧИКОВА, А. С. КЛЕВЦОВА**

Аннотация. Особенности задачи оптического распознавания русскоязычного текста со страниц архивных журналов и обзор существующих средств оптического распознавания текста.

Ключевые слова: OCR, оптическое распознавание текста, русскоязычные журналы, библиотека, облачный сервис, .NET.

Введение

Оптическое распознавание символов (англ. optical character recognition, OCR) – процесс классификации оптических шаблонов, содержащихся в цифровом изображении. Оно используется для перевода изображений рукописного, машинописного или печатного текста в текстовые данные [2]. Распознавание символов имеет широкую сферу применения - от создания электронных версий книг и документов до решения задач автоматизации систем учета данных.

Задаче оптического распознавания символов было уделено много внимания в последние несколько десятилетий, в результате чего появилось большое количество библиотек, сервисов и приложений, решающих данную задачу с разной степенью успешности. В данной статье рассмотрены некоторые из них на примере задачи оцифровки русскоязычных архивных газет и журналов.

Несмотря на то, что задачи распознавания текста широко распространены, данная задача имеет свои особенности, обусловленные спецификой источника информации. Типографика газет и журналов обладает набором особенностей и правил, отличающих ее от типографики книг. Эти особенности обуславливают ряд сложностей, с которыми могут столкнуться существующие сервисы и библиотеки для распознавания текста.

Данная задача актуальна для учреждений, имеющих в своем распоряжении архивы газет и журналов и желающих предоставить полноценный доступ к их цифровым версиям. Примером такого учреждения является «Воронежская областная детская библиотека».

1. Специфика типографики газет и журналов

Типографика – искусство расположения композиции из наборного материала на плоскости листа [1]. Именно оно обуславливает особенности расположения текста на страницах газет и журналов, выбор шрифтов и множество других нюансов дизайна печатных изданий. При этом, «принятые в типографике правила – обобщение опыта, накопленного за века существования печатного дела» [1]. Типографика не является строгой наукой, поэтому любое печатное издание имеет уникальный дизайн. Однако можно проследить ряд

Рис. 1. Пример страниц журналов с разным количеством колонок

особенностей, которыми обладает типографика разных газет и журналов, и которые могут оказать существенное влияние на процесс оптического распознавания текста.

Расположение текста в газетах и журналах значительно отличается от расположения текста в книгах. Такой текст часто располагается в колонках, на которые разделена страница, при этом их количество обычно одинаково для каждой страницы конкретного номера газеты или журнала, но может отличаться для разных изданий. Наиболее часто используются 2 – 4 колонки. Возможны и другие значения количества колонок, так как они определяются дизайном конкретного издания. На рис. 1 представлены примеры страниц журналов с разным количеством колонок.

Ширина колонки зависит прежде всего от характера текста [1], и даже в рамках одного номера журнала эта ширина может изменяться. Вертикальный пробел между двумя колонками текста – межколонник – тоже может быть разной ширины [1]. При этом, строки текста в соседних колонках принято располагать на одном уровне. Эти особенности также можно увидеть на рис. 1.

В процессе распознавания текста перечисленные выше особенности могут стать проблемой: система распознавания текста может некорректно разметить блоки текста, например, восприняв строку второй колонки продолжением строки первой колонки. В таком случае распознанные текстовые данные потеряют правильную последовательность, и появится задача ее восстановления. Подобную задачу нельзя назвать простой, так как нельзя утверждать, что проблема будет возникать всегда или определенным образом. Некоторые системы распознавания текста предоставляют возможность пользователю разметить блоки текста самостоятельно, но автоматизировать данный процесс – тоже достаточно сложная задача с учетом того, что разные издания могут обладать уникальными особенностями расположения текста. Ручная же разметка подобных блоков является достаточно трудоемким процессом.

Можно выделить и другие особенности типографики газет и журналов. Например, часто используется буквица – первая прописная буква текста, главы и т.п. (обычно увеличенного размера), служащая элементом художественного оформления издания [3]. В газетах и журналах она появляется чаще, чем в книгах, порой несколько раз на одной странице. Ее

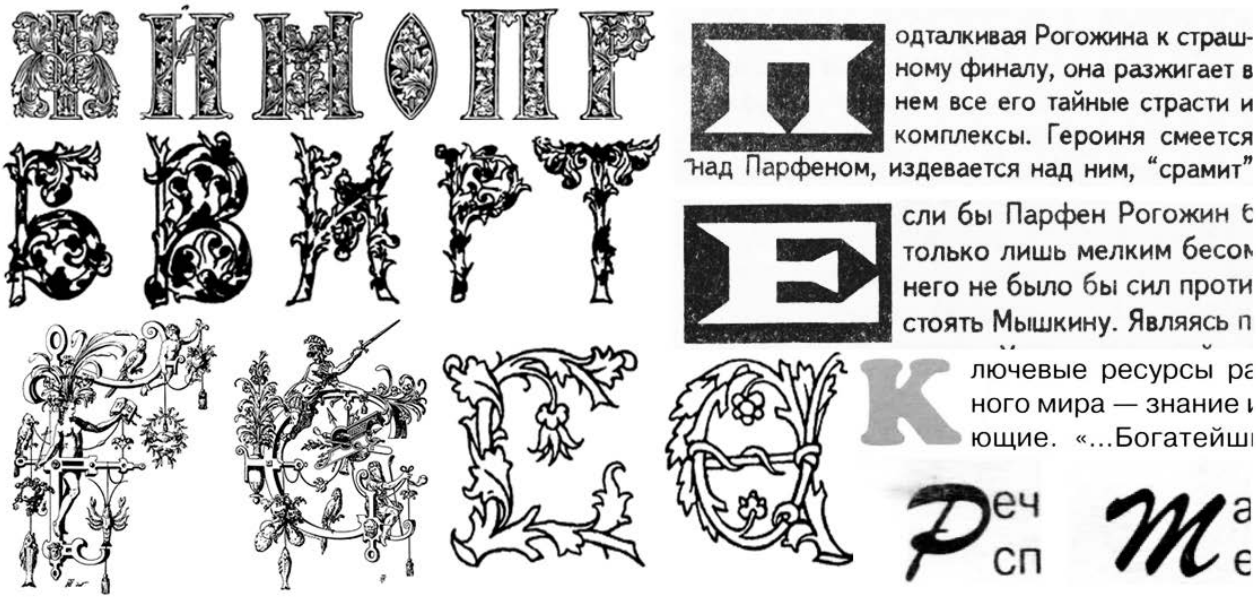

Рис. 2. Пример буквиц в печатных изданиях

выделяющееся художественное оформление может привести к некорректному распознаванию символа. На рис. 2 представлены примеры буквиц.

Кроме того, частота появления изображений на страницах газет и журналов гораздо выше, чем на страницах книг. Это тоже может осложнить процесс распознавания текста.

2. Обзор существующих решений

На рис. 3 представлен пример страницы журнала, который был использован для тестирования всех решений, представленных ниже.

2.1. Tesseract. Tesseract – средство для оптического распознавания символов с открытым исходным кодом, которое было разработано между 1984 и 1994 годами. С 2006 года Tesseract принадлежит компании Google, которая купила его и открыла исходный код под лицензией Apache 2.0 [4].

На данный момент, наиболее актуальной версией является версия 4.1.1. Tesseract можно использовать как из командной строки, так и через API, чтобы распознавать текст с цифровых изображений [8]. Он предоставляет возможность использовать в качестве основы для распознавания один из трех наборов тренировочных данных, предоставленных компанией Google, а также обучить модель на основе собственного набора данных чтобы добавить поддержку новых языков или повысить качество распознавания уже представленных. При использовании Tesseract для распознавания на русском языке достаточно использовать базовые наборы данных.

Ниже представлен пример использования Tesseract для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
using (var engine = new TesseractEngine(@"tessdata-master", "rus",
EngineMode.Default))
{
    using (Pix img = Pix.LoadFromFile(imageFilename))
    {
        using (Page recognizedPage = engine.Process(img))
```

мя демонстрировал свое традиционное предназначение и одновременно новый «вариант использования». Издания на перевернутом стеллаже (вспомним игру в перевертыши), стояли не чинно в ряд, как на обычной полке, но в виде стопок книг. Одна из них напоминала по форме винтовую лестницу, другая — построенный из книг домик. Журналы — с закладками «по месту прописки» интересных литературных произведений, а также статей и вовсе были беспорядочно разбросаны по подиуму.

Визуальным центром выставки стал портрет «грустного клоуна» с серьезным усталым лицом — Юрия Никулина. Рядом стояли аксессуары из фильмов с его участием: разноцветный женский зонтик, шарфик, пустая бутылка. Библиотекари — творцы выставки — рассматривали их как «приманку», направленную на привлечение непроизвольного внимания.

Несомненно, беспорядок, как и вроде бы случайный отбор книг по одинаковому формату, были кажущимися. Поскольку грустное и веселое всегда соседствуют в жизни, в экспозиции «участвовали» книги разных отраслей и тем, объединенные не столько жанром — юмористической литературой, книгами об искусстве комедии и пр., сколько другим признаком: их незаслуженно редко спрашивали читатели. Описанная выставка представляла собой книжно-журнальный развал, который, по данным исследований еще 1970—1980-х гг., особенно любят молодые читатели. Поскольку развалы по-своему противостоят привычной и нередко наводящей скуку «упорядоченности» традиционных форм библиотечной рекомендации.

Неординарный подход к оформлению развала в день 1 апреля, самоиронию библиотекарей, которые «покусились на святое», не побоялись нарушить устоявшийся порядок и даже перевернуть стеллаж и книги вверх тормашками, подростки оценили адекватно. Развал пользовался громадным успехом, и возведенные библиотекарем книжные пирамиды рушились и исчезали на глазах создателей. Сотрудники не уставали подкладывать заранее подготовленные «резервные» издания.

Данный пример хорошо иллюстрирует довод о самодостаточной роли оформления. В авторских выставках оно является частью их содержательно-смысловой основы. Детей, растущих у экранов (телевизоров, видеоплееров, компьютеров), нелегко убедить в необходимости взять ту или иную книгу, если она напрямую не связана с процессом обучения. Их надо заинтересовать, более того — УДИВИТЬ. Не случайно ивановские библиотекари причислили свою первоапрельскую шутку к разряду *выставок-сюрпризов*.

«Замешанные» на веселой фантазии, эти выставки, пользуясь фразеологией их создателей, можно причислить и к так называемым *«фигурным»*, в основе которых находится игра объемами. Последние ориентированы на необычную форму представления. Вот, например, как можно из обычной выставки новых поступлений для дошкольников создать «фигурную». Выставку *«К нам новая книга пришла»* размещают на любимом детьми подиуме «лесенкой»: от нижних ступенек — сказок и рассказов для самых маленьких — до верхней, где представлены энциклопедии для дошкольников. По замыслу авторов, такое расположение книг по-своему отражает ступени читательского роста.

Игра объемами в детской библиотеке нередко сочетается с игрой цветом. Например, необычно выглядит выставка *«Зеркало женской души»*. Разработанная талантливыми библиотекарями из Кировской областной детской библиотеки имени А. С. Грина, она была адресована девочкам-подросткам*. Читательницы могли увидеть на столе задрапированные белыми и черными шелковыми шторами коробки разной величины с книгами русских поэтесс разных эпох.

Попеременно чередовавшийся белый и черный цвета позволяли визуально подчеркнуть различия поколений поэтесс и одновременно преемственность их творчества. Данный прием также подчеркивал неоднозначный, часто трагический жизненный путь женщин-поэтов. Сделанные тушью на белом ватмане цитаты, рисунки, виньетки хорошо смотрелись на фоне «черного ящика». На белой драпировке соседнего раздела аналогичный изобразительный ряд был представлен белыми буквами на черном листе.

Данная выставка отражала и игру линий, которые выступали своеобразным знаком эпохи. Например, портретам поэтесс XIX в. соответствовали виньетки. Серебряный век был представлен черными женскими головками-профилями на белом листе. Наши современницы — черно-белыми фотографиями. Книжный ряд был также проиллюстрирован с помощью игры знаками. В качестве последних выступали «старинный» подсвечник, цветок в бокале, перчатки, шкатулка и другие аксессуары, создающие романтичный женский образ.

К сожалению, авторы выставки обошли вниманием выигрышное изобразительное решение, которое позволяло уточнить смыслы ее заглавия. Мы имеем в виду зеркало: при оформлении такого рода экспозиций им умело «играют» профессиональные дизайнеры. Семиотика свидетельствует, что зеркало «отсылает вперед» и одновременно замыкает пространство, воплощая его интимный характер*.

Желание отступить от стандарта, помочь читателям развить воображение, творческую фантазию приводит к появлению таких игровых по своей природе книжно-иллюстративных экспозиций, как *«выставка-ситуация»*. Ее методику также описывает частинский библиотекарь. Организуемая для подростков, она моделирует необычную, и иногда даже фантастическую ситуацию, о которой мечтают тинэйджеры:

* Информационная среда детской библиотеки Вып. 2: Книжная выставка: традиции и инновации. Из опыта работы ОДБ имени А. С. Грина. Киров, 2003. 39 с.

* Бодрийяр Ж. Система вещей. М.: Рудомино. 1995. С. 18.

Рис. 3. Страница журнала, использованная для тестирования решений

```
        {
            string recognizedText = recognizedPage.GetText();
        }
    }
}
```

Даже в конфигурации по умолчанию Tesseract достаточно хорошо справляется с распознаванием текста с подобной страницы. Он обрабатывает блоки текста в правильной последовательности и корректно распознает большую часть символов. На рис. 4 представлены распознанные текстовые данные.

Можно заметить, что Tesseract не удалось корректно распознать буквицы, что обусловлено их внешним отличием от остального текста. Также стоит отметить, что итоговые текстовые данные нуждаются в обработке перед дальнейшим использованием: удалении лишних символов переноса слов и переноса строки. Только после подобной обработки текстовые данные можно будет использовать, например, для полнотекстового поиска.

2.2. Iron OCR. IronOCR – библиотека на языке C# для оптического распознавания символов с цифровых изображений и PDF документов. Она использует последнюю версию Tesseract, а также предоставляет возможность использовать предыдущие версии [6]. Для данной библиотеки существует пакет для распознавания текста на русском языке.

Помимо функциональности распознавания текста, данная библиотека предоставляет доступ к ряду методов предварительной обработки изображения, которые могут улучшить

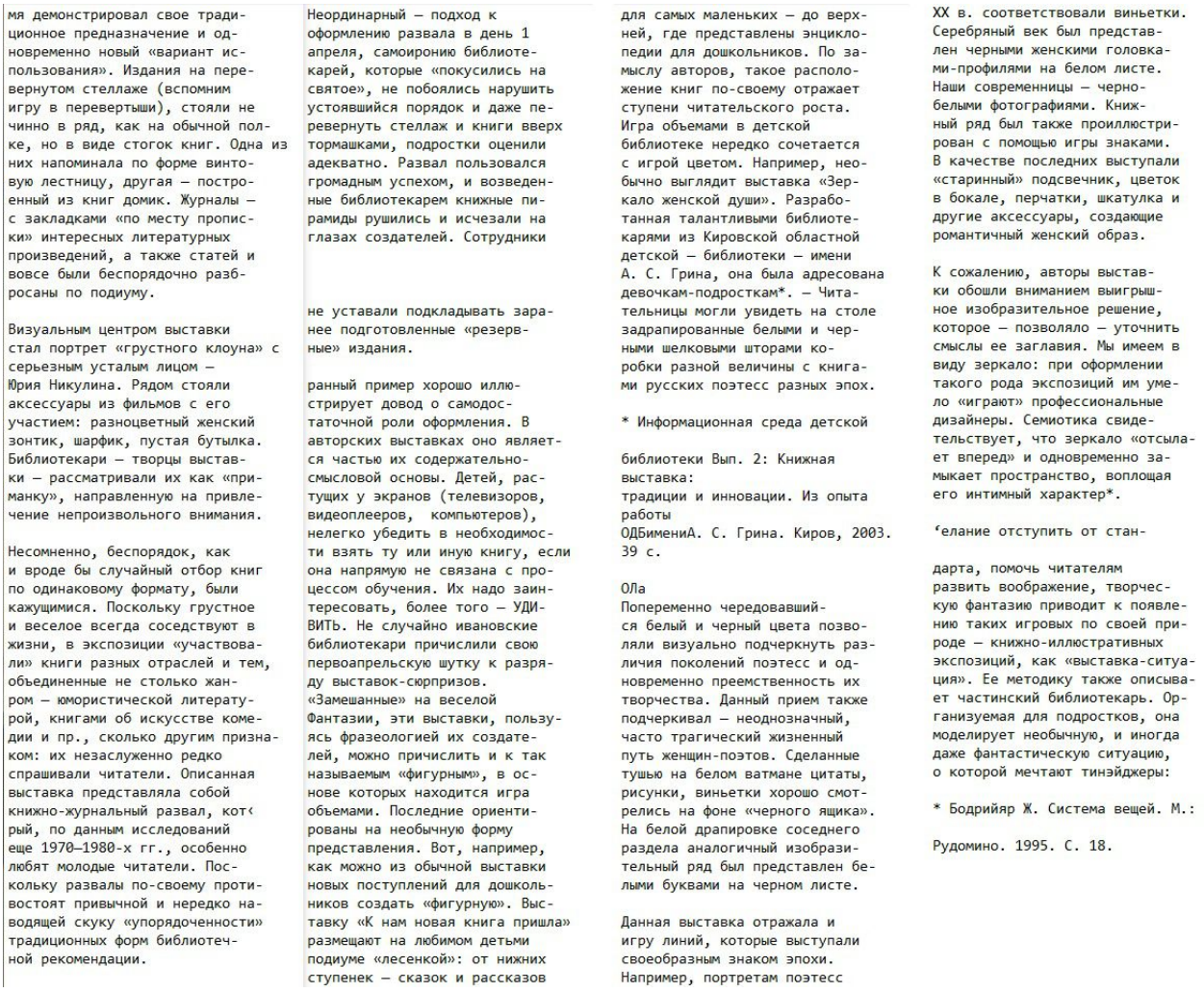

Рис. 4. Результат распознавания с помощью Tesseract

результаты распознавания текста, а также к методам конвертации файлов разных форматов.

Данная библиотека бесплатна для разработки и тестирования, но для коммерческого использования необходимо приобрести лицензию.

Ниже представлен пример использования IronOCR для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
var ocr = new IronTesseract();
ocr.Language = OcrLanguage.Russian;
string Text = ocr.Read(imageFilename).Text;
```

Можно заметить, что результаты распознавания данной библиотеки близки к результатам распознавания Tesseract. Блоки текста также обработаны в правильном порядке, а большинство символов распознано корректно. У данной библиотеки тоже возникли трудности с распознаванием буквиц, и можно заметить, что результаты их распознавания отличаются от результатов распознавания Tesseract. На рис. 5 представлены распознанные текстовые данные.

Для использования итоговых текстовых данных, например, для полнотекстового поиска, тоже потребуется их обработка – удаление лишних символов переноса строк и слов, а также некоторых лишних символов, появившихся в итоговых данных.

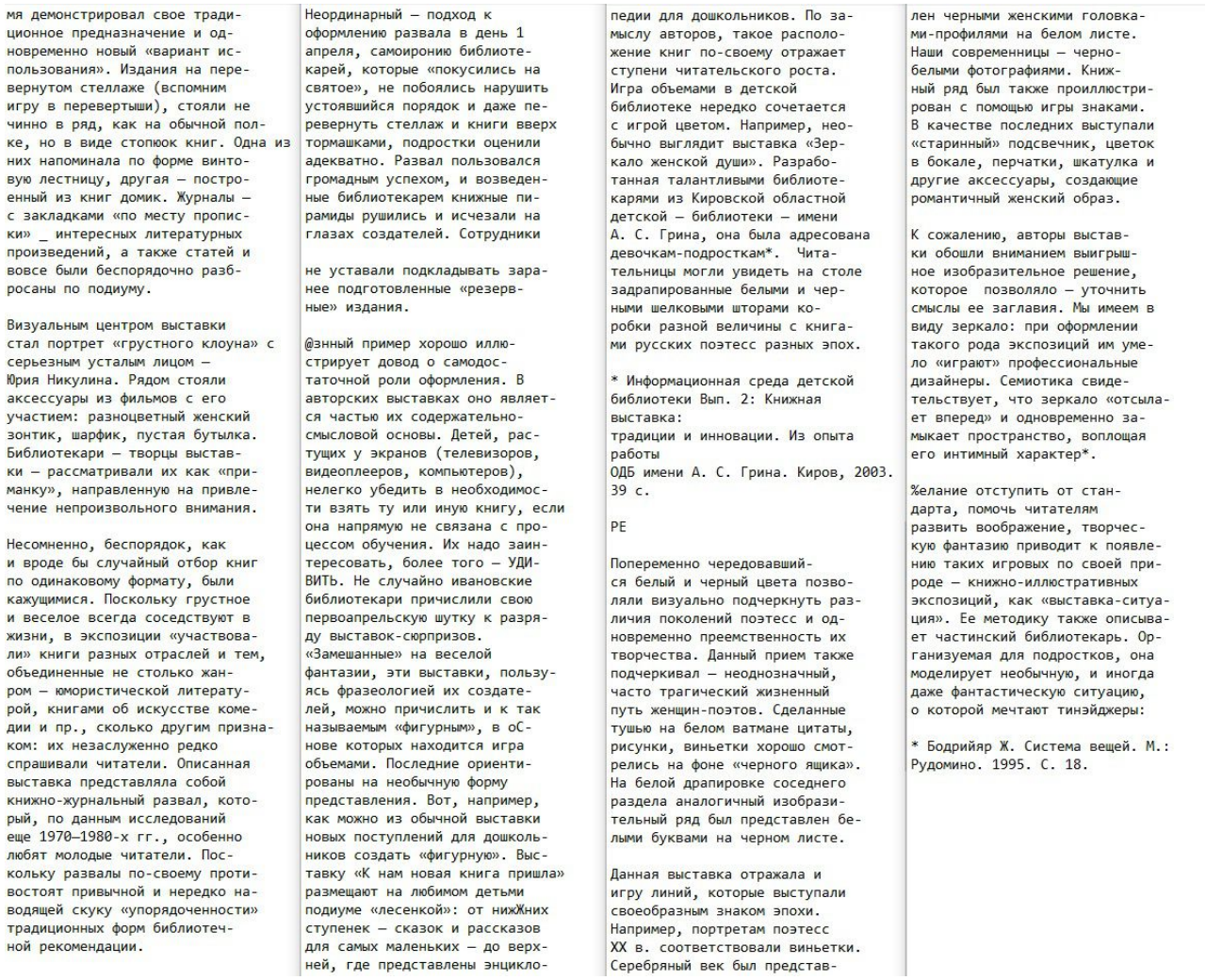

Рис. 5. Результат распознавания с помощью IronOCR

2.3. Google Cloud Vision API. Google Cloud Vision API – набор сервисов в сфере компьютерного зрения от компании Google, доступ к которым предоставляется по подписке. Среди них есть OCR сервис – сервис для оптического распознавания текста. Доступ к ним осуществляется через REST API.

Для использования данного сервиса необходима регистрация аккаунта в Google Cloud Platform. OCR сервис можно использовать бесплатно при ограничении в 1000 изображений в месяц [5].

Данный сервис автоматически распознает язык текста и может работать с русским языком.

Ниже представлен пример использования Google Cloud Vision API для распознавания текста с отсканированной страницы журнала в .NET Core приложении.

```
Image image = Image.FromFile(imageFilename);
var client = ImageAnnotatorClient.Create();
TextAnnotation text = client.DetectDocumentText(image);
string detectedText = text.Text;
```

Результаты распознавания с помощью Google Cloud Vision API представлены на рис. 6.

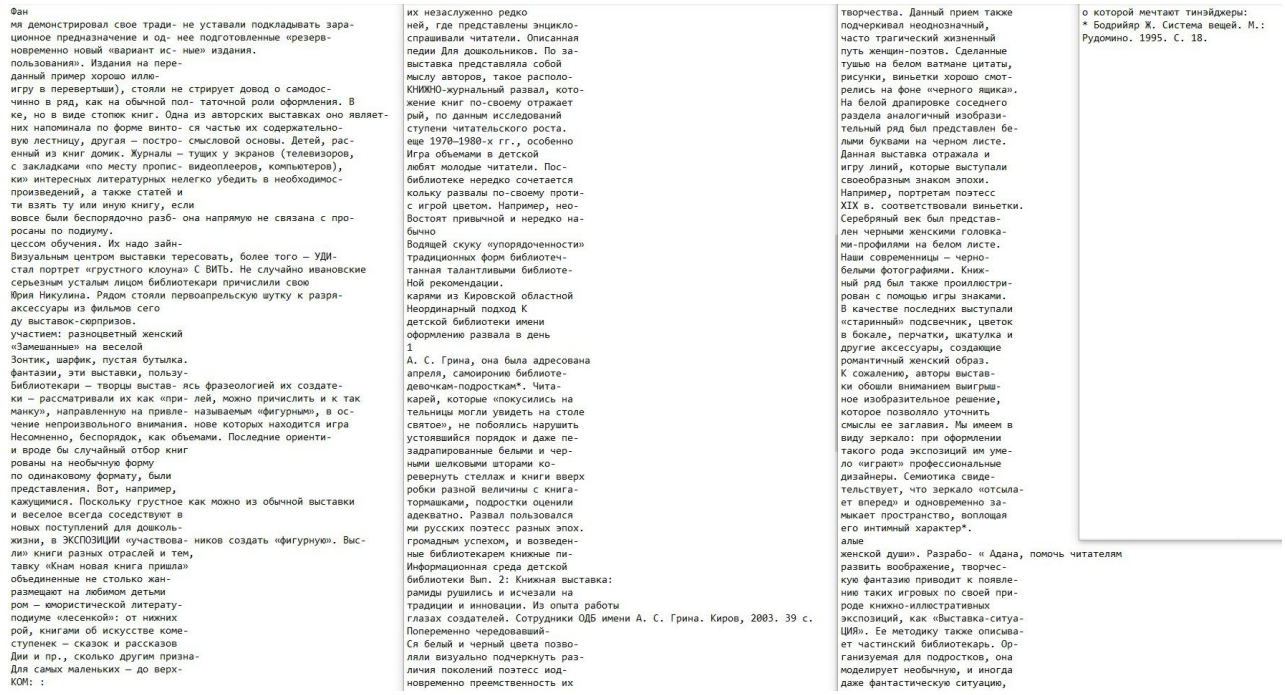

Рис. 6. Результат распознавания с помощью Google Cloud Vision API

Можно заметить, что Google Cloud Vision API, в отличие от ранее рассмотренных библиотек, не удалось сохранить корректный порядок текстовых блоков. Например, в части случаев он распознал соответствующую строку второй колонки текста как продолжение строки первой колонки. Приведение подобных текстовых данных в естественный вид станет достаточно сложной задачей.

Стоит также заметить, что в отличие от ранее рассмотренных библиотек, этот сервис корректно распознал присутствующую в тексте буквицу.

2.4. Microsoft Azure Computer Vision API. Microsoft Azure Computer Vision API – набор облачных решений от компании Microsoft в сфере компьютерного зрения. Они входят в облачную платформу Microsoft Azure. Среди API данного набора присутствует сервисы оптического распознавания символов.

Read API – наиболее современный и актуальный сервис в данном облаке, способный распознавать печатные и рукописные символы с цифровых изображений и PDF-документов [7]. К сожалению, данный сервис не поддерживает распознавание на русском языке, что делает невозможным его использование для решения данной задачи.

Предыдущая версия данного сервиса, OCR API, использующая более старую модель распознавания, к сожалению, является устаревшей и не рекомендованной к использованию, так как скоро прекратится ее поддержка [7]. Но она может работать с русским языком, что позволяет надеяться, что в скором времени и Read API будет поддерживать распознавание на русском языке.

Заключение

Среди рассмотренных в данной статье средств оптического распознавания текста наиболее подходящими для решения задачи оцифровки русскоязычных архивных газет и журналов представляются решения на основе Tesseract. Распознанные ими текстовые данные

отличаются корректным порядком слов и большим процентов правильно распознанных символов, что значительно упрощает необходимую обработку результатов.

Облачное решение Google Cloud Vision API в ряде случаев испытывает сложности с сохранением корректного порядка слов, что значительно усложнит необходимую обработку текстовых данных, полученных в результате распознавания, по сравнению с другими решениями. А облачное решение от Microsoft на данный момент не предоставляет возможность работы с текстом на русском языке в актуальной версии сервиса.

СПИСОК ЛИТЕРАТУРЫ

1. Королькова А. Живая типографика / А. Королькова. — Москва : IndexMarket, 2012. — 224 с.
2. Chaudhuri A. Optical Character Recognition Systems / A. Chaudhuri, K. Mandaviya, P. Badelia, S. K. Ghosh. — Cham : Springer International Publishing AG, 2017. — 248 p.
3. Словари и энциклопедии на Академике [Электронный ресурс] / Академик — Режим доступа: https://dic.academic.ru/dic.nsf/es/97444/
4. An Overview of the Tesseract OCR Engine [Электронный ресурс] / Ray Smith, Google Inc. — Режим доступа: https://tesseract-ocr.github.io/docs/tesseracticdar2007.pdf
5. Cloud Vision documentation [Электронный ресурс] / Google Inc. — Режим доступа: https://cloud.google.com/vision/docs
6. IronOCR for .NET [Электронный ресурс] / Iron Software LLC — Режим доступа: https://ironsoftware.com/csharp/ocr/docs/
7. Microsoft Computer Vision Documentation [Электронный ресурс] / Microsoft — Режим доступа: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr
8. Tesseract User Manual [Электронный ресурс] / Tesseract OCR — Режим доступа: https://tesseract-ocr.github.io/tessdoc/

Капранчикова Алисия Александровна

Воронежский государственный университет, факультет прикладной математики и механики, кафедра программного обеспечения и администрирования информационных систем, магистрант (ВГУ, ПММ, ПОиАИС)

E-mail: kapranchikovaal@gmail.com

Клевцова Александра Сергеевна

Воронежский государственный университет, факультет прикладной математики и механики, кафедра программного обеспечения и администрирования информационных систем, магистрант (ВГУ, ПММ, ПОиАИС)

E-mail: alexandra.klevtsova59@gmail.com