

ОСОБЕННОСТИ РАЗРАБОТКИ ФУНКЦИОНАЛЬНОГО КОМПОНЕНТА ИНФОРМАЦИОННОЙ СИСТЕМЫ, РЕАЛИЗУЮЩЕГО РАСПОЗНАВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

FEATURES OF THE DEVELOPMENT OF A FUNCTIONAL COMPONENT OF AN INFORMATION SYSTEM THAT IMPLEMENTS THE RECOGNITION OF TEXTUAL INFORMATION

*Yu. Maximov
A. Kasymov*

Summary. This article discusses the process of developing a functional component of an information system for recognizing text and graphic information. The task of recognizing graphic information is currently one of the most urgent. The most frequently used methods and means of recognition are considered. The development of software for the recognition and processing of graphic and textual information is an important activity of many companies and organizations. To date, there are a large number of text recognition algorithms, including those using neural networks. However, for practical application, an algorithm is needed that can be used to recognize text written in different languages. The issue of creating an application software for recognizing text information, with the aim of further use as an element of an information system, is being considered. To create the program, an algorithm was chosen that allows you to produce a distribution.

Keywords: automation, algorithmization, software, management, information processing, recognition.

Максимов Юрий Максимович

Аспирант, Воронежский Государственный
Технический Университет
yuramaximo@mail.ru

Касымов Алексей Алексеевич

Аспирант, Воронежский Государственный
Технический Университет
kasimlele@live.ru

Аннотация. В статье рассматривается процесс разработки функционального компонента информационной системы для распознавания текстовой и графической информации. Задача распознавания графической информации в настоящее время является одной из наиболее актуальных. Рассматриваются наиболее используемые методы и средства распознавания. Разработка программного обеспечения распознавания и обработки графической и текстовой информации является важным направлением деятельности многих компаний и организаций. На сегодняшний день существует большое количество алгоритмов распознавания текста, в том числе и с применением нейронных сетей. Однако, для практического применения необходим алгоритм, который может быть использован для распознавания текстов, написанных на разных языках. Вопрос создания прикладного программного обеспечения распознавания текстовой информации, с целью дальнейшего использования как элемента информационной системы. Для создания программы был выбран алгоритм, который позволяет производить распознавание, разработать программную реализацию, протестировать и отладить программу, а также разработать интерфейс пользователя. Широкое распространение графических форматов и необходимость конвертирования данных, полученных в ходе выполнения различных процессов, требуют наличия специальных средств для распознавания графических данных. В конце статьи приводится краткое описание интерфейса пользователя для работы с программой распознавания.

Ключевые слова: автоматизация, алгоритмизация, программное обеспечение, управление, обработка информации, распознавание.

Введение

За последнее десятилетие стремительное развитие получили технологии цифровой обработки изображений, а также сопутствующие им технологии распознавания графических объектов.

Трудно переоценить прикладное значение данных технологий, которые нашли свое применение в таких

сферах как оцифровка архивов и библиотечных фондов, также технологии распознавания позволяют задействовать автоматизированные средства обработки для проведения целого комплекса мероприятий, связанных с анализом распознанных графических объектов.

По мнению авторов, программные решения, использующие в своей основе технологии цифровой об-

BRITISH AIRWAYS CHECK-IN INFORMATION FOR PASSENGERS

For all domestic and European flights, check-In is one and a half hours before departure. For international departures, check-in is two hours before departure. You must check in within the recommended time before your departure.

- 1 Your baggage must be securely closed and locked.
- 2 Every piece of your checked baggage must be labelled with your name and destination address. We suggest a label should always be put inside each piece of baggage.
- 3 You must remove all old destination labels from previous trips. Old labels may delay your baggage.
- 4 Do not carry unidentified objects for other people.
- 5 You may carry personal items as cabin baggage. British Airways recommend passengers include important documents, cash, valuable items, medication and laptops (portable computers) in their cabin baggage. Do not pack these in your checked baggage,



Рис. 1. Сегментация текста на блоки, строки, элементы и символы

работки и распознавания графических объектов могут быть крайне полезными в правоохранительной сфере, а также при решении ряда задач, стоящих перед МЧС России, например:

Выезд сотрудников правоохранительных органов на место совершения преступления, сопряжен с необходимостью сбора доказательной базы и улик, которыми зачастую являются документы. В настоящее время зачастую сотрудники используют только лишь копируемую технику, а анализ документов происходит без использования средств автоматизации;

Работа сотрудников МЧС России по ликвидации последствий, наступивших в следствии возникновения чрезвычайных ситуаций сопряжена с обработкой большого массива задокументированных данных, которые были сформировано в процесс функционирования различного рода предприятий. Особую озабоченность вызывают экономические объекты, на которых используются зарубежные технологические решения, зарубежная техника и оборудование, оригинальные спецификации, техническая документация которых написаны на иностранных языках, при этом в рамках одной организации могут использоваться такого рода оборудование, произведенное в различных странах.

Особую тревогу и озабоченность, вызывает тот факт, что на сегодняшний день ведущие мировые компании, занимающиеся технологиями распознавания изобра-

жений (ABBV, Kofax и др.) в настоящее время перестали официально сопровождать свое программное обеспечение на территории Российской Федерации.

Актуальность рассматриваемой темы, обусловлена так же тем, что на сегодняшний день не получили широкого распространения программные решения распознавания текста, реализованные в качестве полнофункциональных подсистем, существующих ведомственных информационных систем.

Особое прикладное значение имеют мобильные программные продукты, реализующие распознавание графические изображений, в том числе текстовой информации. Для решения подобного рода задач используется метод оптического распознавания символов (англ. optical character recognition, OCR) — реализующий механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные, использующиеся для представления символов в компьютере. [1,2]

Аналитическая часть

Одна из вариаций алгоритма оптического распознавание символов подразумевает реализацию следующих этапов:

1. Получение не редактируемого текстового содержимого из отсканированных документов всех типов, от планшетных сканирований корпора-

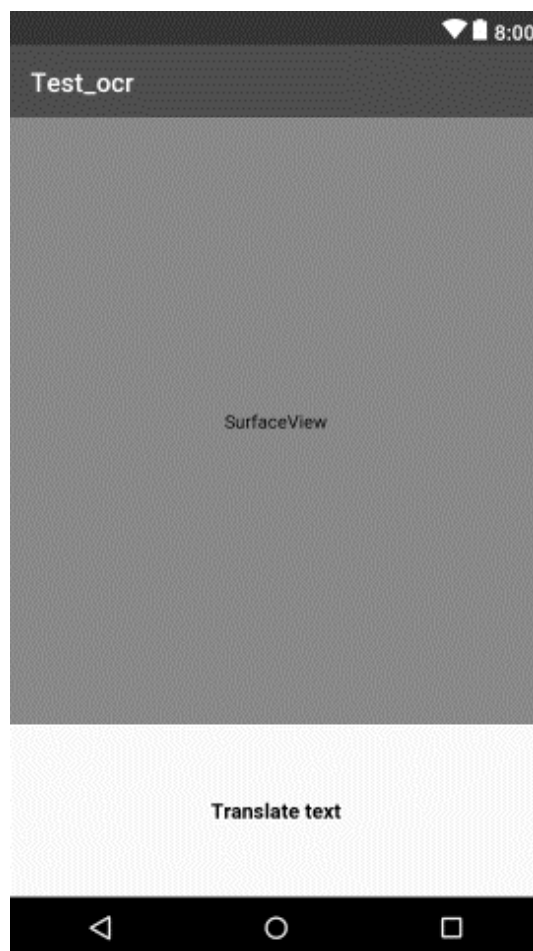


Рис. 2. Интерфейс приложения

тивных архивных материалов до видеозаписей наблюдения в реальном времени и данных мобильной визуализации.

2. Предварительная обработка — это очистка исходных изображений на агрегированном уровне, чтобы текст было легче различить, а шум уменьшался или устранялся.
3. Сегментация и извлечение признаков, подразумевает сканирование содержимого изображения на наличие групп пикселей, которые, вероятно, представляют собой отдельные символы, и отнесение каждого из них к своему собственному классу. Сегментация текста на блоки, строки, элементы и символы происходит поэтапно для извлечения признаков из него (Рис. 1), где:
 - ◆ Блок — это непрерывный набор текстовых строк,
 - ◆ Строка — это непрерывный набор слов на одной оси,
 - ◆ Элемент представляет собой непрерывный набор буквенно-цифровых символов («слово») на одной оси,

◆ Символ — это одиночный буквенно-цифровой символ на оси.

4. Обучение осуществляется после того, как все функции определены, данные обрабатываются в сеансе обучения нейронной сети, где модель попытается разработать обобщенное отображение объекта.
5. Этап проверки и переподготовки может быть реализован с непосредственным участием человека, при этом реализация исправлений подразумевает их возвращение в последующие учебные сессии. На этом этапе, возможно, потребуется пересмотреть качество данных [4,5].

Практическая реализация алгоритма

В качестве языка программирования для реализации алгоритма, был использован язык Java, который позволяет использовать достаточно сложную систему классов, описывающих все объекты, участвующие для решения поставленной задачи, а также реализует под-

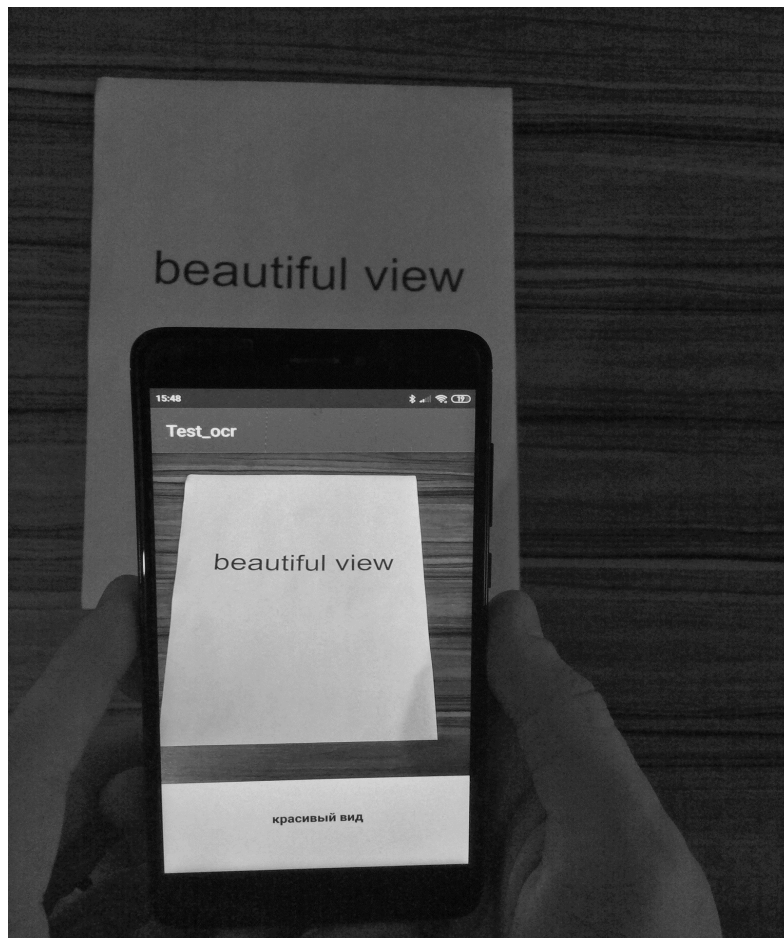


Рис. 3. Результат работы программного обеспечения по обработке текстовой информации

держку библиотек компьютерного зрения OpenCV [6], [7], высокоточную модель распознавания текста на основе глубокого изучения Tesseract [8], библиотеки компьютерного зрения Google ML Kit [9].

Использование среды разработки IntelliJ IDEA при создании проекта позволяет автоматически генерировать все необходимые файлы при этом основные файлы для работы с кодом хранятся в папке «src», файлы, отвечающие за графическое отображение в папке «layout». Для разработки мобильного интерфейса в Android используется расширяемый язык разметки (Extensible Markup Language), сокращенно XML.

Например, использование Android studio позволяет реализовать два режима работы с xml файлами: графический и просмотр кода. На рисунке 2 представлен внешний вид графического режима на примере стартовой страницы создаваемого приложения.

В ходе разработки программного решения были использован вид разметки LinearLayout, который по-

зволяет отражать View-элементы в виде одной строки (если установлен Horizontal) или одного столбца (если установлен Vertical).

Реализация интерфейса подразумевает использование двух основных программных компонента SurfaceView и TextView.

SurfaceView — объект Surface, который поддерживает рисование в фоновом потоке и позволяет использовать OpenGL для изображения трёхмерной графики.

TextView — текстовое поле для отображения результата работы по распознаванию и переводу.

В функциональном модуле было реализовано два класса, а именно MainActivity и NetworkUtils. Каждый класс прописывает название пакета и необходимые модули импорта. Для каждого экрана соответствует отдельный Java класс, для главного экрана изображения соответствует класс «MainActivity», основные части класса:

onCreate() — Метод, который задаёт начальную установку параметров при инициализации активности, здесь же вызывается метод startCameraSource().

startCameraSource() — Основной метод приложения в котором происходит основная работа приложения.

В программном коде был использован класс TranslateTask, который отвечает за направление запросов для перевода текста в фоновом потоке, методы класса TranslateTask:

protected String doInBackground() — основной метод, который выполняется в новом потоке и не имеет доступа к пользовательскому интерфейсу. В этом методе отправляется URL запрос к Yandex Translate API для перевода распознанного текста.

protected void onPostExecute() — выполняется после doInBackground, это метод имеет доступ к пользовательскому интерфейсу и он обновляет текстовое поле, в которое заносит переведенный текст.

Используемый класс NetworkUtil описывает работу с URL запросами, формирование URL из строки и получение результата запросов, основные используемые поля и свойства класса NetworkUtil: YANDEX_API_BASE_URL; YANDEX_JSON_TRANSLATE; PARAM_KEY, PARAM_TEXT; PARAM_LANG, YANDEX_API_KEY.

На рисунке 3 продемонстрирован успешный результат работы приложения по определению текст и его правильного перевода.

Заключение

Рассмотренное в настоящей статье программное решение отличается простотой, может быть реализовано программистами начального уровня, и как следствие финансовые расходы на разработку могут быть ограничены зарплатным фондом штатных специалистов учреждения или организации, что в условиях повсеместной экономии финансовых средств так же является крайне актуальным.

Применение рассмотренного программного решения может осуществляться в рамках реализации служебных обязанностей широкого круга исполнителей, так же оно может быть использовано в качестве функционального элемента государственных или коммерческих информационных систем.

На сегодняшний день особую озабоченность вызывают существующие подходы обработки больших объемов данных, требующие реализацию более эффективных и устойчивых методов распознавания символов.

ЛИТЕРАТУРА

1. Burie JC, J. ICDAR2015 competition on smartphone document capture and OCR (SmartDoc) / Burie JC, J. Chazalon, M. Coustaty et al. // 3th International Conference on Document Analysis and Recognition (ICDAR). — Tunis: IEEE, 2015;(3): 1161–1165.
2. Herbert, F.S. History of OCR, Optical Character Recognition / F.S. Herbert. — 4th ed. — Manchester Center: Recognition Technologies Users Association, 1982.
3. Khaoula Text Recognition in Multimedia Documents: A Study of two Neural-based OCRs Using and Avoiding Character Segmentation / Khaoula, G. Elagouni, M. Christophe et al. // International Journal on Document Analysis and Recognition (IJDA). — 2013. — № 17. — С. 1–13.
4. Ravina, Mithe Optical Character Recognition / Mithe Ravina, Indalkar Supriya, Divekar Nilam. // International Journal of Recent Technology and Engineering (IJRTE). — 2013. — № 2. — С. 45–49.
5. Bhaskar, S. Implementing Optical Character Recognition on the Android Operating System for Business Cards / S. Bhaskar, N. Lavassar, Green Scott. // Digital Image Processing. — 2010. — № . — С. 139–146.
6. Qashlim Smartphone Technology Applications for Milkfish Image Segmentation Using OpenCV Library / Qashlim, B. Akhmad, H. Basri. — T // International Journal of Interactive Mobile Technologies (IJIM). — 2020. — № 14. — С. 150–155.
7. Adrian, Rosebrock OpenCV OCR and text recognition with Tesseract / Rosebrock Adrian. // pyimagesearch: [Internet]. — URL: <https://pyimagesearch.com/2018/09/17/opencv-ocr-and-text-recognition-with-tesseract/> (Accessed: Oct 05.2022).
8. Smith, R. An overview of the Tesseract OCR Engine / R. Smith. // Proc 9th Int. Conf. on Document Analysis and Recognition. — Curitiba: IEEE, 2007. — С. 629–633.
9. Adrian, Rosebrock OpenCV OCR and text recognition with Tesseract / Rosebrock Adrian. // pyimagesearch [Internet]. — URL: <https://pyimagesearch.com/2018/09/17/opencv-ocr-and-text-recognition-with-tesseract/> (Accessed: Oct 05.2022).
10. Hagos Learn Android Studio 3: Efficient Android App Development / Hagos, T. — 4th ed. — Berkeley: 2018. — 260 c.

© Максимов Юрий Максимович (yuramaximo@mail.ru), Касымов Алексей Алексеевич (kasimlele@live.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»