СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2018. – № 3–4 (93).** – 91–100

УДК 519.24 DOI: 10.17212/2307-6879-2018-3-4-91-100

ОБРАБОТКА СКАНИРОВАННОГО ТЕКСТА*

Ю.В. МИЛОВСКАЯ

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, студентка факультета автоматики и вычислительной техники. E-mail: milovskaya.1999@yandex.ru

В статье дается обзор методов и алгоритмов обработки сканированного текста на примере одной из самых популярных программ оптического распознавания символов ABBYY FineReader. Распознавание – процедура получения текста с картинки, которая после сканирования появляется в одном из форматов: BMP, JPG, PNG, GIF (могут быть и другие). Другими словами, это процесс перевода графического изображения символов (букв) в компьютерные текстовые символы. Сделать это можно, имея качественную цифровую копию оригинального текста и набор современных компьютерных программ для распознавания текста. Для корректного распознавания в первую очередь проводится анализ текста (сверху вниз, снизу вверх, алгоритм MDA, сочетающий в себе первые два метода). Фрагмент изображения, согласно принципу целостности, будет интерпретирован как некий объект (символ), только если на нем присутствуют все структурные элементы с соответствующими взаимосвязями. При этом система выдвигает ряд гипотез относительно того, на что похож обнаруженный объект с помощью специальных механизмов распознавания, которые называются классификаторами. После обнаружения всех фрагментов и выдвижения гипотез объекты целенаправленно проверяются с использованием принципа адаптивности, подразумевающего наличие накопленных ранее сведений о возможных начертаниях символа в распознаваемом документе. Сложность возникает с документами, содержащими в себе рисунки, таблицы, колонтитулы. Упростить работу с данными структурами позволяет бинаризация. В качестве примера приведен алгоритм Брэдли и его реализация.

Ключевые слова: сканирование, распознавание, анализ изображения, алгоритм Брэдли, бинаризация, классификаторы, гипотезы, цифровое изображение, целостность, целенаправленность, адаптивность, фоновые текстуры

* Статья получена 18 ноября 2018 г.

ВВЕДЕНИЕ

Наверное, каждый из нас сталкивался с задачей, когда нужно перевести бумажный документ в электронный вид. Особенно часто это нужно делать тем, кто учится, работает с документацией, переводит тексты при помощи электронных словарей и т. д.

Вообще сканирование и распознавание текста – довольно трудоемкий процесс, так как большинство операций приходится делать вручную. Мы попытаемся разобраться по шагам, что, как и почему.

Поможет нам разобраться в процессах сканирования и последующей обработке текста программа ABBYY FineReader.

С помощью всем известного сканера создается цифровое изображение, фрагмент которого будет интерпретирован системой как некий объект (символ).

1. БАЗОВЫЕ ПРИНЦИПЫ

Оптическое распознавание символов (OCP) относится к области ИИ (искусственный интеллект), поэтому разработчики стремятся имитировать деятельность человеческого мозга. Подобно нашей зрительной системе, программы следуют базовым принципам, таким как целостность, целенаправленность и адаптивность. Объект рассматривается как совокупность своих частей, любая его интерпретация преследует определенную цель, поэтому и выдвигаются гипотезы о целенаправленности проверки, что позволяет экономить мощность и реже ошибаться.

Чаще всего применяются два типа анализа: сверху вниз или снизу вверх, но ABBYY разработали специальный алгоритм MDA (многоуровневый анализ документа), который сочетает в себе два вышеуказанных.

2. БИНАРИЗАЦИЯ ИЗОБРАЖЕНИЙ: АЛГОРИТМ БРЭДЛИ

На этапе предварительной обработки и анализа графических данных перед любой OCR-системой стоят две основные задачи: выявление логической структуры документа и подготовка изображения к процедурам распознавания.

Подлежащий распознаванию документ часто выглядит заметно сложнее, чем белая страница с черным текстом. Таблицы, иллюстрации, фоновые изображения, колонтитулы, всё чаще применяемые для оформления, усложняют структуру страницы. Основная задача состоит в том, чтобы отделить текст от текстур и иллюстраций.

Здесь уместно будет сказать о практикуемых методах подготовки. Все современные системы распознавания начинают процесс с создания черно-белого изображения документа. При этом подлежащее анализу изображение чаще всего цветное или полутоновое (то есть состоящее из разных оттенков серого цвета).

Упростить работу с изображением позволяет бинаризация – перевод цветного (или в градациях серого) изображения в двухцветное черно-белое. Один из основных методов ее реализации – алгоритм Брэдли.

```
voidBradley_threshold(unsignedchar* src, unsignedchar* res, intwidth,
intheight) {
	constint S = width / 8;
	int s2 = S / 2;
	constfloat t = 0.15;
	unsignedlong* integral_image = 0;
	long sum = 0;
	int count = 0;
	int index;
	int x1, y1, x2, y2;

	//рассчитываем интегральное изображение
	integral_image = newunsignedlong[width*height *
sizeof(unsignedlong*)];

	for (inti = 0; i<width; i++) {
		sum = 0;
		for (int j = 0; j <height; j++) {
			index = j * width + i;
			sum += src[index];
			if (i == 0)
				integral_image[index] = sum;
			else
				integral_image[index] = integ-
ral_image[index - 1] + sum;
		}
	}

	//находим границы для локальные областей
	for (inti = 0; i<width; i++) {
		for (int j = 0; j <height; j++) {
			index = j * width + i;
			x1 = i - s2;
			x2 = i + s2;
			y1 = j - s2;
			y2 = j + s2;
```

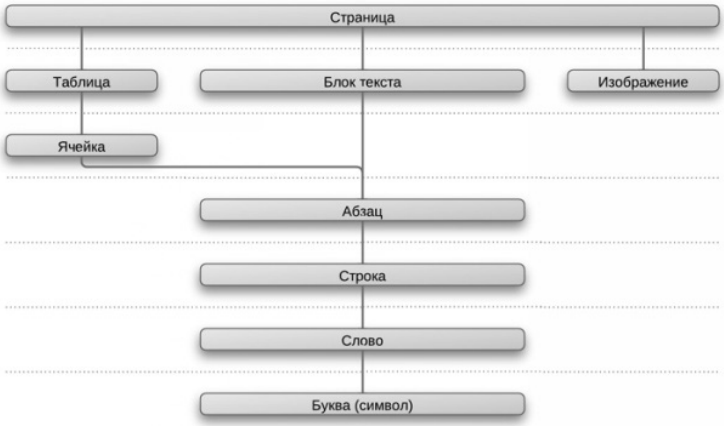
```
                    if (x1 < 0)
                          x1 = 0;
                    if (x2 >= width)
                          x2 = width - 1;
                    if (y1 < 0)
                          y1 = 0;
                    if (y2 >= height)
                          y2 = height - 1;
                    count = (x2 - x1)*(y2 - y1);
                    sum = integral_image[y2*width + x2] - inte-
gral_image[y1*width + x2] -
                          integral_image[y2*width + x1] + inte-
gral_image[y1*width + x1];
                    if ((long)(src[index] * count) < (long)(sum*(1.0 - t)))
                          res[index] = 0;
                    else
                          res[index] = 255;
             }
      }
      delete[]integral_image;

}
```

ABBYY FineReader построена на других принципах и не пытается решать задачу бинаризации напрямую. Принцип целенаправленности устанавливает другой подход к обнаружению строк в тексте или слов в строке: они должны быть в документе, надо только суметь их распознать. FineReader использует процедуры интеллектуальной фильтрации фоновых текстур и адаптивной бинаризации для повышения качества поиска. Первая позволяет уверенно отделять строки текста от фона любой сложности, вторая – гибко выбирать оптимальные для данного участка параметры бинаризации. Конечно же, к этим процедурам система прибегает лишь в тех случаях, когда предварительный анализ указывает на подобную необходимость. В каждом конкретном случае ABBYY FineReader выбирает подходящий «инструмент», опираясь на информацию, которая накопилась в процессе анализа документа.

Для того чтобы корректно воспроизводить в электронном виде такие документы, все современные OCR-программы начинают распознавание именно с анализа структуры.

3. МНОГОУРОВНЕВЫЙ АНАЛИЗ

Любой высокоуровневый объект может быть представлен как набор объектов более низкого уровня: буквы образуют слово, слова – строки, строки – абзац и т. д. Поэтому анализ всегда начинается в направлении сверху вниз. Программа делит страницу на объекты, их, в свою очередь, на объекты низших уровней, и так далее, вплоть до символов. Когда символы выделены и распознаны, начинается обратный процесс – объединение объектов высших уровней, который завершается формированием целой страницы.



Многоуровневый анализ

На всех этапах многоуровневого анализа (см. рисунок) добавлена возможность обратной связи. То есть результаты анализа на одном из нижних уровней всегда могут повлиять на действия с объектами более высоких уровней. Наличие обратной связи в процедуре MDA дает возможность резко понизить вероятность грубых ошибок, связанных с неверным распознаванием объектов более высоких уровней.

4. РАСПОЗНАВАНИЕ СИМВОЛОВ. КЛАССИФИКАТОРЫ

Как следует из общих принципов работы ABBYY FineReader, на каждом логическом уровне документа выдвигается ряд гипотез. На следующем уровне каждая из них порождает еще несколько предположений. Поэтому при распознавании букв FineReader оперирует огромным количеством гипотез, учитывающих все возможные варианты деления строки на слова, слова на

буквы и т. д. Для быстрого и точного принятия решений система объединяет гипотезы в многоуровневые структуры – модели.

В результате структурирования количество подлежащих проверке гипотез сильно сокращается, так что последующая проверка происходит максимально быстро и эффективно.

Для распознавания символов в программе FineReader используются специальные механизмы, которые называются классификаторами, порождающими список гипотез, которые затем проверяются. Входными данными для классификаторов может служить не только графическая информация, но и сформированный в ходе распознавания список гипотез. В последнем случае классификатор не выдвигает новых гипотез, а лишь изменяет веса уже имеющихся, подтверждая или опровергая их. Такой подход, в котором также четко прослеживаются принципы IPA (интеллектуальные алгоритмы обработки), обеспечивает более интеллектуальный анализ изображения и наиболее точное распознавание документа.

Для обеспечения надежной работы механизма словарной проверки лингвистами из компании ABBYY SoftwareHouse были созданы полноценные словари, позволяющие системе FineReader распознавать тексты на многих языках. Специалисты ABBYY SoftwareHouse наделили словари уникальными свойствами, тем самым сделав их морфологически структурированными. Получился гибкий и мощный инструмент, позволяющий ABBYY FineReader моделировать словоформы.

ЗАКЛЮЧЕНИЕ

Итак, предварительная обработка завершена, проведен анализ, гипотезы выдвинуты, а все слова текстового блока распознаны. Формирование документа завершено. Теперь программа обращается к самому пользователю за подтверждением, всё ли корректно распознано. Чтобы вероятность ошибки сводилась к минимуму, ABBYY FineReader предпочитает работать с цветными или полутоновыми изображениями, самостоятельно преобразуя их в черно-белые, и не пытается решить задачу бинаризации напрямую. Для повышения качества поиска использует процедуры интеллектуальной фильтрации фоновых текстур и адаптивной бинаризации.

Естественно, к этим процедурам система прибегает не всегда, а лишь в тех случаях, когда предварительный анализ указывает на подобную необходимость. В каждом конкретном случае ABBYY FineReader выбирает подходящий «инструмент», опираясь на информацию, накопленную в процессе анализа документа.

СПИСОК ЛИТЕРАТУРЫ

1. *Bradley D.*, *Roth G.* Adaptive thresholding using the integral image [Electronic resource] // Journal of Graphics Tools. – 2007. – Vol. 12 (2). – P. 13–21. – URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.420.7883 (accessed: 15.03.2019).

2. *Llamm A.* Бинаризация изображений: алгоритм Брэдли [Электронный ресурс]. – URL: https://habr.com/post/278435/ (дата обращения: 15.03.2019).

3. *Мозговой А.А.* Проблемы извлечения рукописных слов из сканированного изображения [Электронный ресурс] // Моделирование, оптимизация и информационные технологии. – 2013. – № 1. – URL: http://moit.vivt.ru/wp-content/uploads/2013/04/mozgovoy_1_13_1.pdf (дата обращения: 15.03.2019).

4. *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. – М.: Техносфера, 2005. – 1072 с.

5. Программное обеспечение системы технического зрения. Бинаризация полутоновых изображений / Д.Е. Охоцимский, И.М. Бродская, С.С. Камынин, Е.И. Кугушев. – М.: ИПМ, 1987. – 25 с.

6. *Штарьков Ю.М.* Универсальное кодирование. Теория и алгоритмы. – Москва: Физматлит, 2013. – 279 с. – ISBN 978-5-9221-1517-9.

7. *Недбайлов А.А.* Сканирование и распознавание текста: учебное пособие для студентов вузов региона. – Владивосток: Дальневост. гос. техн. ун-т, 2001. – 61 с. – ISBN 5-88871-181-0.

8. *Жадаев А.Г.* Сканирование и распознавание текстов: самоучитель по работе с ABBYY® FineReader 10. – М.: ДМК, 2010. – 247 с. – ISBN 978-5-94074-595-2.

9. *Полилова Т.А.* Технологии сканирования изображений: учебно-методическое пособие / Московский ин-т открытого образования. – М.: МИОО, 2004. – 32 с. – ISBN 5-94898-030-8.

10. *Трушин Н.Г.* Исследование передачи изображений при сканировании и получении копий фотоснимков: дис. ... канд. техн. наук: 02.00.04. – Кемерово, 2006. – 95 с.

11. *Горский Н.Д., Анисимов В., Горская Л.* Распознавание рукописного текста: от теории к практике. – СПб.: Политехника, 1997. – 126 с. ISBN 5-7325-0450-8.

12. *Литвинюк С.Б.* Разработка и исследование методов повышения достоверности информации в системах, использующих технологию оптического распознавания символов: дис. ... канд. техн. наук: 05.25.05. – М., 1999. – 161 с.

Миловская Юлия Владимировна, студентка факультета автоматики и вычислительной техники Новосибирского государственного технического университета. E-mail: milovskaya.1999@yandex.ru

DOI: 10.17212/2307-6879-2018-3-4-91-100

Processing scanned text*

Y.V. Milovskaya

Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk,630073, Russian Federation, student of the faculty of automation and computer engineering. E-mail: milovskaya.1999@yandex.ru

The article provides an overview of the methods and algorithms for processing scanned text on the example of one of the most popular optical character recognition programs – ABBYY FineReader. Recognition is the procedure for obtaining text from the image, which after scanning appears in one of the formats: BMP, JPG, PNG, GIF (there may be others). In other words, it is the process of translating graphic images of characters (letters) into computer text characters. This can be done with a high-quality digital copy of the original text and a set of modern computer programs for text recognition. For correct recognition, first of all, the text is analyzed (top-down, bottom-up, MDA algorithm, combining the first two). A fragment of an image, according to the principle of integrity, will be interpreted as a certain object (symbol) only if it contains all structural elements with corresponding interrelations. In this case, the system puts forward a number of hypotheses regarding what the detected object looks like with the help of special recognition mechanisms, which are called classifiers. After all fragments are discovered and hypotheses are put forward, objects are checked purposefully using the principle of adaptability, which implies the presence of previously accumulated information about the possible character traits in a recognizable document. The difficulty arises with documents containing drawings, tables, footers. Binarization allows to simplify work with these structures. Bradley's algorithm and his implementation is given as an example.

Keywords: scanning, recognition, analysis of the image, Bradley's algorithm, binarization, qualifiers, hypotheses, digital image, integrity, focus, adaptivity, background textures

REFERENCES

1. Bradley D., Roth G. Adaptive thresholding using the integral image. *Journal of Graphics Tools*, 2007, vol. 12 (2), pp. 13–21. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.420.7883 (accessed 15.03.2019).

* *Received 18 November 2018.*

2. Llammt A. *Binarizatsiya izobrazhenii: algoritm Bredli* [Image binarization: the Bradley algorithm]. Available at: https://habr.com/post/278435/ (accessed 15.03.2019).

3. Mozgovoy A.A. Problemy izvlecheniya rukopisnykh slov iz skanirovannogo izobrazheniya [The problem of extracting handwritten words from the scanned image]. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii – Modeling, Optimization and Information Technology*, 2013, no. 1. Available at: http://moit.vivt.ru/wp-content/uploads/2013/04/mozgovoy_1_13_1.pdf (accessed 15.03.2019).

4. Gonzalez R., Woods R. *Digital image processing*. 2nd ed. Upper Saddle River, NJ, Prentice Hall, 2002 (Russ. ed.: Gonsales R., Vuds R. *Tsifrovaya obrabotka izobrazhenii*. Moscow, Tekhnosfera Publ., 2005. 1072 p).

5. Okhotsimskii D.E., Brodskaya I.M., Kamynin S.S., Kugushev E.I. *Programmnoe obespechenie sistemy tekhnicheskogo zreniya. Binarizatsiya polutonovykh izobrazhenii* [Software of the vision system. Binarization of gray scale images]. Moscow, IPM Publ., 1987. 25 p.

6. Shtar'kov Yu.M. *Universal'noe kodirovanie. Teoriya i algoritmy* [Universal coding. Theory and algorithms]. Moscow, Fizmatlit Publ., 2013. 279 p. ISBN 978-5-9221-1517-9.

7. Nedbailov A.A. *Skanirovanie i raspoznavanie teksta* [Scanning and text recognition]. Vladivostok, Far Eastern State Technical University Publ., 2001. 61 p. ISBN 5-88871-181-0.

8. Zhadaev A.G. *Skanirovanie i raspoznavanie tekstov* [Scanning and recognition of texts]. Moscow, DMK Publ., 2010. 247 p. ISBN 978-5-94074-595-2.

9. Polilova T.A. *Tekhnologii skanirovaniya izobrazhenii* [Image scanning technologies]. Moscow, Moscow institute of open education Publ., 2004. 32 p. ISBN 5-94898-030-8.

10. Trushin N.G. *Issledovanie peredachi izobrazhenii pri skanirovanii i poluchenii kopii fotosnimkov*. Diss. kand. tekhn. nauk [Investigation of the transmission of images with scanning and obtaining copies of photos. PhD eng. sci. diss.]. Kemerovo, 2006. 95 p.

11. Gorskii N.D., Anisimov V., Gorskaya L. *Raspoznavanie rukopisnogo teksta: ot teorii k praktike* [Handwriting recognition: from theory to practice]. St. Petersburg, Politekhnika Publ., 1997. 126 p. ISBN 5-7325-0450-8.

12. Litvinyuk S.B. *Razrabotka i issledovanie metodov povysheniya dostovernosti informatsii v sistemakh, ispol'zuyushchikh tekhnologiyu opticheskogo raspoznavaniya simvolov*. Diss. kand. tekhn. nauk [Development and research of methods to improve the reliability of information in systems using optical character recognition technology. PhD eng. sci. diss.]. Moscow, 1999. 161 p.

Для цитирования:

Миловская Ю.В. Обработка сканированного текста // Сборник научных трудов НГТУ. – 2018. – № 3–4 (93). – С. 91–100. – DOI: 10.17212/2307-6879-2018-3-4-91-100.

For citation:

Milovskaya Yu.V. Obrabotka skanirovannogo teksta [Processing scanned text]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 2018, no. 3–4 (93), pp. 91–100. DOI: 10.17212/2307-6879-2018-3-4-91-100.