МИНОБРНАУКИ РОССИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет *романо-германской филологии*

*Кафедра теории и методики преподавания иностранных языков и культур*

*Курсовая работа*

*Автоматическое распознавание текста. Распознавание текста с помощью ОСR-программ.*

Студент\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гладких А.К. 12.09.2019

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Донина О.В. к.н. 13.09.2019

Оглавление

[Введение 3](#_Toc19356906)

[1. Распознавание текста 4](#_Toc19356907)

[1.1 Основные понятия и определения теории распознавания текста 4](#_Toc19356908)

[1.2 Распознавание текста включает в себя следующие подзадачи и подпроцессы 5](#_Toc19356909)

[2. Технология сканирования 7](#_Toc19356910)

[2.1 Суть процесса распознавания 7](#_Toc19356911)

[3. Основные характеристики сканеров 9](#_Toc19356912)

[3.1 Принципы работы сканеров 9](#_Toc19356913)

[4. Типы пакетов OCR 11](#_Toc19356914)

[4.1 Обучаемые 11](#_Toc19356915)

[4.2 Интеллектуальные 11](#_Toc19356916)

[5. OCR CuneiForm 13](#_Toc19356917)

[5.2 Особенности 13](#_Toc19356918)

[6. Freemore OCR 15](#_Toc19356919)

[Заключение 16](#_Toc19356920)

[Литература 17](#_Toc19356921)

[Интернет-ресурсы 17](#_Toc19356922)

## Введение

В настоящее время большинство документов составляется на компьютерах. Существующие системы охватывают деятельность отдельных организаций, а обмен данными между организациями осуществляется с помощью традиционных бумажных документов. Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид.

Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл. Более предпочтительным, по сравнению с графическим, является текстовое представление информации. Этот вариант позволяет существенно сократить затраты на хранение и передачу информации, а также позволяет реализовать все возможные сценарии использования и анализа электронных документов.

Цель курсовой работы: ознакомиться с распознаванием текста с помощью OCR-программ.

1. Исследовать разновидности распознавания текста.
2. Определить типы OCR-программ.

Объект исследования: OCR-программы.

## Распознавание текста

* 1. Основные понятия и определения теории распознавания текста

Распознавание символов – это операция преобразования текстовой информации из графических форматов данных в текстовые. Применяется для ввода текстов в компьютер посредством сканирования печатного или рукописного материала.

Распознавание текста является одним из направлений распознавания образов. Распознавание образов представляет собой очень сложную задачу в теоретическом и практическом смыслах, несмотря на то, что с ней достаточно легко справляются многие живые организмы и человек. Крайне сложно создать искусственную систему и ее технически реализовать для того, чтобы эффективно выполнять данный процесс. В данном случае, под распознаванием понимается соотнесение изображения объекта, его образа, набора признаков самому объекту.

Примерами и приложениями систем распознавания образов могут являться как распознавание текста в общем, так и отдельных его символов, распознавание речи, человеческих лиц, биометрических данных человека, штрих-кодов продуктов, номеров машин и т.д.

Примерами распознавания текста являются: оцифровка изображений текста (сканированные книги, статьи, журналы) для последующей работы с его цифровым аналогом, обработка анкетных бланков, распознавание номеров машин и надписей на объектах и т.д. Задача распознавания текста остается актуальной на сегодняшний день, так как не существует стопроцентной универсальной системы по распознаванию текста.

* 1. Распознавание текста включает в себя следующие подзадачи и подпроцессы

1. Поступающее на вход системы изображение должно быть очищено от шума и приведено к виду, позволяющему эффективно выделять символы и распознавать их.
2. Система должна разбить изображение на блоки текста, основываясь на особенностях его выравнивания и распределения по нескольким колонкам.
3. Изображение с текстом должно быть разделено на изображения строк, а затем на изображения символов для того, чтобы в дальнейшем обработать каждый символ по отдельности. После данного шага разные системы распознавания работают по своим специфическим алгоритмам.
4. Изображение символа может обрабатываться целиком, для этого оно сравнивается с имеющимися шаблонами. Другим вариантом является выделение характеристик изображаемого символа: отбор характерных признаков, и классификация данных признаков по имеющимся в системе критериям. На выходе четвертого шага появляется возможный вариант буквы. Однако обычно системы на этом не останавливаются и продолжают работу на основе других методов, уточняя полученный результат.
5. Результат распознавания может быть не удовлетворительным. Для получения более хороших результатов в системе может быть встроен блок обучения. С помощью этого блока можно задать системе примеры начертания разных букв в данном шрифте. После процесса обучения предполагается лучшее качество распознавания текста.

Система распознавания текста не всегда должна следовать всем описанным шагам, но основные действия процесса распознавания являются общими для любого алгоритма.

Ввод информации с бумаги в компьютер является в настоящее время часто выполняемой задачей. До недавнего времени эта задача решалась исключительно путем кодирования с клавиатуры компьютера. Большинство документов, подлежащих обработке, представлены в оригиналах на бумажном носителе.

Поэтому неотъемлемой частью любой системы автоматизации являются компьютерные системы, включающие программы сканирования и оптического распознавания образов. Задача распознавания состоит в том, чтобы преобразовать входное (отсканированное) изображение в текст из подвижных символов. Иными словами графическое изображение каждого входного символа заменяется обозначением этого символа, используемым в компьютере (т.е. кодом, понятным компьютерной системе).

Системы оптического распознавания символов стали неотъемлемой частью интегрированных пакетов, поддерживающих ввод в компьютер, хранение и обработку бумажных и электронных документов. Система включает в свой состав сканер для ввода информации.

## Технология сканирования

Сканирование - процесс оцифровки аналогового при помощи специального устройства, называемого сканером. Сканирование производится для получения, на основе оригинала, его цифрового «портрета», пригодного для компьютерной обработки.

Сканер - оптико-электронное устройство для ввода в компьютер графических изображений. Сканер создает оцифрованное изображение документа и помещает его в память компьютера.

Для работы с архивными документами в настоящее время используются, в основном, черно-белые и полутоновые монохромные сканеры. Это связано в первую очередь, с тем, что преобладающим типом архивного документа является текстовый документ, отпечатанный на пишущей машинке или монохромном принтере, с рукописным заполнением или правкой темными чернилами и эпизодически включающий печати, рисунки, схемы или черно-белые фотографии.

* 1. Суть процесса распознавания

Чтобы реализовать автоматический или автоматизированный перевод бумажных документов в электронный вид, необходимо выполнить сканирование бумажных документов и распознать их содержимое с помощью специальных программ, называемых системами оптического распознавания символов (Optical Character Recognition - OCR). Системы оптического распознавания символов предназначены для автоматического ввода печатных документов в компьютер.

Обработка изображения OCR-системой включает в себя анализ графического изображения, переданного сканером, и распознавание каждого символа.

Процессы анализа макета страницы:

1. определение областей распознавания
2. определение таблиц
3. определение картинок
4. выделение в тексте строк и отдельных символов.

OCR-системы могут достигать наилучшей точности распознавания (процент правильно распознанных символов) - свыше 99,9% для чистых изображений, составленных из обычных шрифтов.

1. Основные характеристики сканеров

Разрешение (Resolution) - число точек или растровых ячеек, из которых формируется изображение, на единицу длины или площади. Чем больше разрешение устройства, тем более мелкие детали могут быть воспроизведены. Измеряется в «точках на квадратный дюйм» (DPI, dots per inch). Типовое разрешение промышленных сканеров - 200-300 DPI.

Разрядность цвета (глубина цвета) - количество разрядов каждого пикселя в цифровом изображении, в т.ч. выдаваемом сканером. Описывает максимальное количество цветов, воспроизводимое сканером в виде степени числа 2. Одному разряду соответствует черно-белое изображение, 8-ми - серое полутоновое, 16-ти - цветное, 24-цветное изображение, наиболее близкое к человеческому восприятию (модель RGB), 36bit и больше - полноцветное изображение с высокой достоверностью цветопередачи, предназначенное для профессиональной работы, чаще всего в издательском деле.

Время сканирования. Измеряется в страницах в минуту (иногда - в секундах на изображение). Типовые значения различны для разных типов сканеров.

Формат. Формат сканируемого документа. Как правило - A3/A4.

Интерфейс передачи данных - способ подключения сканера к компьютеру. Способы могут быть различны (к COM или USB порту, к SCSI карте и др.).

* 1. Принципы работы сканеров

1. Изображение, считываемое сканером, хранится в электронном виде;
2. Сканер не может отличать текст от рисунков и представляет их в виде отдельных
3. Графических фрагментов из черных и белых элементов.

Если созданное сканером изображение содержит текст и рисунки, то при помощи OCR можно:

1. Отделить текст от рисунков;
2. Записать этот текст в формате файла текстового процессора.

Программное обеспечение в современных системах OCR выполняет анализ форм букв и создание текстового файла, в который распознаваемый текст записывается посимвольно с последовательным формированием слов и предложений.

## Типы пакетов OCR

* 1. Обучаемые

В последнее время наблюдается тенденция к объединению этих двух типов в одном пакете, что перекликается с попытками разработать принципиально новые алгоритмы распознавания.

В случае обучаемых систем происходит поточечное сравнение оцифрованного символа с образцом в справочной таблице. Оно осуществляется путем наложения масок символов на символ из считанного документа и проверки на полное совпадение. Как только совпадение достигается, символ добавляется в файл. В процессе распознавания символа при наложении маски размеры шрифта документа и маски должны быть одинаковыми. Это означает, что необходимо иметь маски для каждого размера шрифта и что система оказывается более эффективной в случае однотипного текста. С учетом этих ограничений, при разрешающей способности сканера 300 dpi обеспечивается достаточная точность обработки для того ограниченного набора шрифтов, который система “знает”.

Число известных системе шрифтов обычно невелико -- около десятка. Однако достоинство подобных систем в том, что они предоставляют пользователю возможность обучать их новым гарнитурам. Обучение -- довольно длительный процесс, а, кроме того, впоследствии при сравнении шрифтов возможна путаница, вызванная наличием у некоторых символов дополнительных графических элементов типа умляутов, трема и седилей.

4.2 Интеллектуальные

В случае интеллектуальной системы - более мощном методе преобразования считанного документа в текстовый файл - программа самостоятельно идентифицирует каждый оцифрованный символ. Здесь маска символа заменяется на его контур. Такой эталонный контур может быть использован для любых размеров шрифтов. Стандартная система, основанная на этом методе, может распознавать несколько тысяч шрифтов. Поскольку такие системы не пользуются фиксированными масками, не возникает проблем, связанных с совпадением одной маски с несколькими разными символами. Кроме определения символа по контуру система выполняет ряд перекрестных проверок полученного текста.

Среди многих методов наиболее распространенными являются следующие два. Первый метод заключается в частотном анализе текста. Сравнивается частота появления символа в тексте с его частотой в языке оригинала. Второй метод позволяет обнаружить неправильное сочетание символов, исходя из правил орфографии. Очевидно, что для эффективной реализации этого необходимо настроить систему на конкретный язык, на котором написан документ.

## OCR CuneiForm

CuneiForm ([англ.](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%2590%25D0%25BD%25D0%25B3%25D0%25BB%25D0%25B8%25D0%25B9%25D1%2581%25D0%25BA%25D0%25B8%25D0%25B9_%25D1%258F%25D0%25B7%25D1%258B%25D0%25BA) cuneiform, кьюниформ - [клинопись](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%259A%25D0%25BB%25D0%25B8%25D0%25BD%25D0%25BE%25D0%25BF%25D0%25B8%25D1%2581%25D1%258C)), Cognitive OpenOCR –  [свободно распространяемая](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%25A1%25D0%25B2%25D0%25BE%25D0%25B1%25D0%25BE%25D0%25B4%25D0%25BD%25D0%25BE%25D0%25B5_%25D0%25BF%25D1%2580%25D0%25BE%25D0%25B3%25D1%2580%25D0%25B0%25D0%25BC%25D0%25BC%25D0%25BD%25D0%25BE%25D0%25B5_%25D0%25BE%25D0%25B1%25D0%25B5%25D1%2581%25D0%25BF%25D0%25B5%25D1%2587%25D0%25B5%25D0%25BD%25D0%25B8%25D0%25B5) [открытая](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%259E%25D1%2582%25D0%25BA%25D1%2580%25D1%258B%25D1%2582%25D0%25BE%25D0%25B5_%25D0%25BF%25D1%2580%25D0%25BE%25D0%25B3%25D1%2580%25D0%25B0%25D0%25BC%25D0%25BC%25D0%25BD%25D0%25BE%25D0%25B5_%25D0%25BE%25D0%25B1%25D0%25B5%25D1%2581%25D0%25BF%25D0%25B5%25D1%2587%25D0%25B5%25D0%25BD%25D0%25B8%25D0%25B5) система [оптического распознавания текстов](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2F%25D0%259E%25D0%25BF%25D1%2582%25D0%25B8%25D1%2587%25D0%25B5%25D1%2581%25D0%25BA%25D0%25BE%25D0%25B5_%25D1%2580%25D0%25B0%25D1%2581%25D0%25BF%25D0%25BE%25D0%25B7%25D0%25BD%25D0%25B0%25D0%25B2%25D0%25B0%25D0%25BD%25D0%25B8%25D0%25B5_%25D1%2581%25D0%25B8%25D0%25BC%25D0%25B2%25D0%25BE%25D0%25BB%25D0%25BE%25D0%25B2) российской компании [Cognitive Technologies](https://infourok.ru/go.html?href=https%3A%2F%2Fru.wikipedia.org%2Fwiki%2FCognitive_Technologies).

OCR CuneiForm была разработана компанией Cognitive Technologies как коммерческий продукт в 1993 году. Система поставлялась с наиболее популярными моделями сканеров, МФУ и ПО в России и мире: Corel Draw, Hewlet-Pachard, Epson, Xerox, Samsung, Brother, Mustek, OKI, Canon, Olivetti и др.

* 1. Особенности

CuneiForm позиционируется как система преобразования электронных копий бумажных документов и графических файлов в редактируемый вид с возможностью сохранения структуры и гарнитуры шрифтов оригинального документа в автоматическом или полуавтоматическом режиме.

Система включает в себя две программы для одиночной и пакетной обработки электронных документов. Поддерживается смесь русского и английского языка. Бесплатная программа для считывания текстовой информации с изображений. Точность распознавания на порядок ниже, чем у предыдущей рассматриваемой программы. Но как для бесплатной утилиты, функционал все-таки на высоте.

Программа может прочитать и сохранять шрифт распознаваемого текста. В базе шрифтов содержится большинство используемых печатных шрифтов.

Поддерживается даже распознавание текста вышедшего из печатной машинки. Для обеспечения точности к процессу распознавания подключаются специальные словари, которые пополняют словарный запас из сканируемых документов.

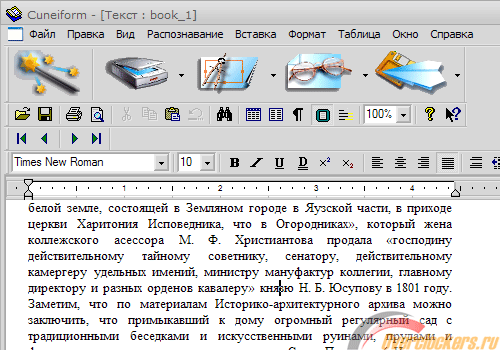


Рис.1 Пример программы OCR CuneiForm.

## Freemore OCR

Программа позволяет оперативно извлекать текст и графику с изображений. Софт поддерживает работу с несколькими сканерами без потери производительности. Извлеченный текст может быть сохранен в формате текстового документа или документа MS Office.

Кроме того предусмотрена функция многостраничного распознавания.

Распространяется Freemore OCR бесплатно, однако, интерфейс только на английском. Но это обстоятельство никак не влияет на удобство пользования, потому как организованы элементы управления интуитивно понятным образом.

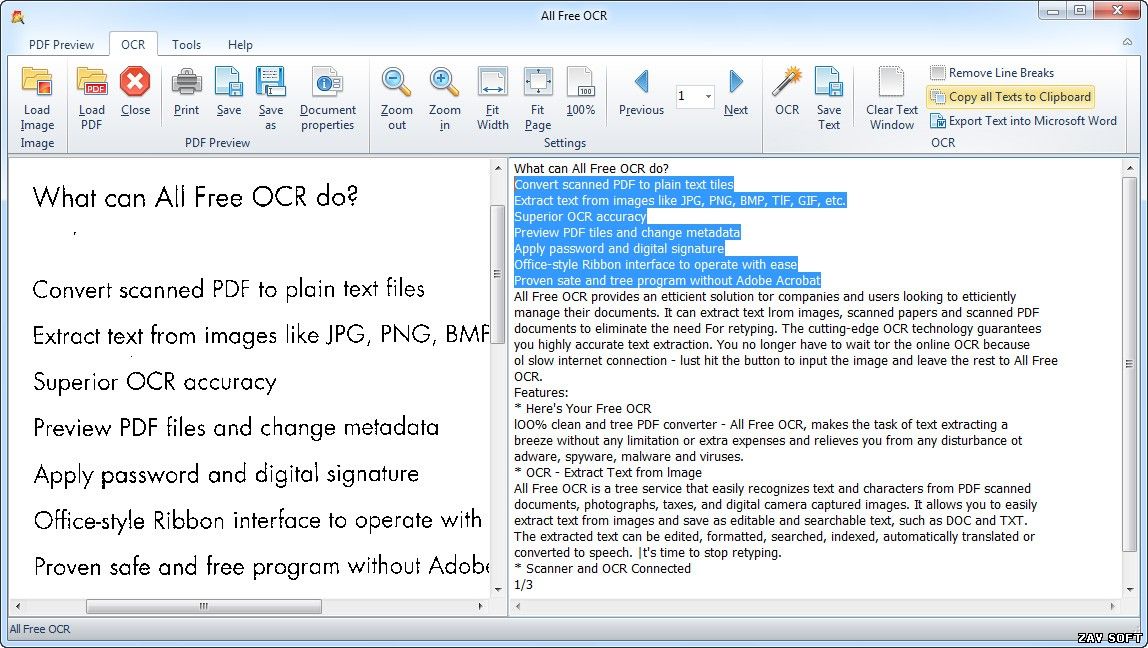


Рис.2 Пример программы Freemore OCR.

## Заключение

В процессе работы мы познакомились с распознаванием текста с помощью OCR-программ. Исследование показало, что спектр разновидностей распознавания текста очень обширный. Так же мы изучили и определили типы OCR-программ.

## Литература

Д. Блатнер, Г. Флейшман, С. РотСканирование и растрирование изображений / Пер. с англ. - М.: Издательство ЭКОМ, 1999. - 400с.: илл.

Дэвид А. Форсайт, Джин Понс Компьютерное зрение.- М.: Издательский центр «Вильямс», 2004. -- С. 928.

Информационные технологии в профессиональной деятельности: Учеб. пособие для сред. проф. образования / Елена Викторовна Михеева. - 2-е изд., стер. - М.: Издательский центр «Академия», 2005. - С. 384.

Технические средства информатизации: Учебник для сред. Проф. образования / Е.И. Гребенюк, Н.А. Гребенюк. - 2-е изд., стер. - М.: Издательский центр «Академия», 2005. - С. 272.

## Интернет-ресурсы

http://www.awella.ru/scannews17.php.htm - программное обеспечение сканеров;

http://www.compress.ru/article.aspx?id=11015&part=index11ext1 - компьютер пресс/планшетные сканеры (журнал);

http://www.gtech.ru/ - графические технологии;

http://www.xerox.ru/ru/software/s2pcdesktop/omnipage// - программы для офиса.