# Системы распознавания символов

- Системы оптического распознавания символов (Optical Character Recognition, ли OCR-системы) предназначены для автоматического ввода документов в память компьютера.
- Сейчас OCR-системы успешно справляются с обработкой печатных документов. Задача распознавания рукописных символов решается только в нескольких частных случаях.
- Распознавание символов - это сложная проблема, которая требует для своего решения привлечения новейших методов дискретной математики и искусственного интеллекта. Она не решается простыми переборными алгоритмами.
- На рынке программных продуктов предлагается несколько систем автоматического распознавания примерно равного класса, обладающих похожими функциональными возможностями. В нашей стране самой популярной является программа FineReader, разработанная фирмой ABBYY.
- Очевидно, что конечная цель сканирования текста —получение текстового файла со всеми возможностями редактирования, которые дает текстовый редактор.
- Компьютер наделяется некоторым интеллектом в виде программного обеспечения специализированного назначения, чтобы он смог опознать символы алфавита. Решаемые задачи при разработке таких программ относятся к области создания искусственного интеллекта, а именно — к области распознавания образов.

- Первоначально задача распознавания символов решалась путем сравнения участков текста-изображения с заранее заданным начертанием символов алфавита. Отсюда и название **OCR** — **O**ptical **C**haracter **R**ecognition, *оптическое распознавание символов*.
- Программы, реализующие алгоритм сравнения, предполагали использование в документах определенного, специально сконструированного шрифта. Применение любого другого шрифта приводило к непредсказуемым результатам.
- В последнее десятилетие появилось много новых для распознавания символов. Ранее созданные программы, получившие признание пользователей, также постоянно развиваются. Появляются новые функциональные возможности, повышаются основные показатели таких программ — *точность и скорость распознавания*.

- **Преобразование документа в электронный вид выполняется OCR-системами поэтапно: сканирование и предварительная обработка изображения, анализ структуры документа, распознавание, проверка результатов, затем производится реконструкция (воссоздание исходного вида) документа, и экспорт.**
- **Методы, применяемые при распознавании, весьма разнообразны.**

# 2. Базовые принципы функционирования OCR-систем

- **Принцип целостности (integrity)**, согласно которому объект рассматривается как целое, состоящее из связанных частей. Связь частей выражается в пространственных отношениях между ними, и сами части получают толкование только в составе предполагаемого целого, то есть в рамках гипотезы об объекте.
- **Принцип целенаправленности (purposefulness)**: любая интерпретация данных преследует определённую цель. Следовательно, распознавание должно представлять собой процесс выдвижения гипотез о целом объекте и целенаправленной их проверки. Такая система не только экономнее расходует вычислительные мощности, но и существенно реже ошибается.
- **Принцип адаптивности (adaptability)** подразумевает способность системы к самообучению. Полученная при распознавании информация упорядочивается, сохраняется и используется впоследствии при решении аналогичных задач. Преимущество самообучающихся систем заключается в способности «спрямлять» путь логических рассуждений, опираясь на ранее накопленные знания.

# 3. Типовые проблемы, связанные с распознаванием символов

Существует ряд существенных проблем, связанных с распознаванием рукописных и печатных символов:

- разнообразие форм начертания символов;
- искажения изображений;
- вариации размеров и масштаба символов.

Каждый отдельный символ может быть написан различными стандартными шрифтами, например, специальными шрифтами, использующимися в системах OCR , а также множеством нестандартных шрифтов. Кроме того, различные символы могут обладать сходными очертаниями. Например, **'U'** и **'V'** , **' S '** и **'5'** , **' Z '** и **'2'** , **' G '** и **'6'** .

Искажения цифровых изображений символов могут быть следующих видов:

- Искажения формы;
- вращения с изменением наклона символов;
- грубым дискретом оцифровки изображений;

# 4. Разрешение сканирования для систем оптического распознавания символов (OCR)

- Если отсканировать печатную страницу с текстом, то только человек может воспринять полученное изображение как совокупность символов, имеющую смысловое текстовое содержание и обладающую возможностью быть отредактированной.
- **Для преобразования графических символов в буквы и представления их совокупности в текстовом формате существуют системы оптического распознавания символов (OCR).** В основном, имеющиеся программы распознавания символов могут автоматически устанавливать режим и разрешение сканирования, без участия оператора сканирования.
- В режиме line art стандартна установка разрешения в 300 dpi. (текст хорошо выделяется на фоне и имеет размер (кегль) 10–14 пунктов). При размерах символов меньше 8 пунктов рекомендуется разрешение увеличить до 600 dpi.
- Символы больших размеров могут отсутствовать в библиотеке символов наиболее распространенных начертаний и кеглей, которые используются программами в процессе распознавания.

# 5. Система FineReader. Основные возможности

- Система распознавания текста FineReader 8.0 была разработана российской компанией ABBYY, основанной в 1989 г . студентом 4-го курса МФТИ Давидом Яном. На сегодняшний день эта компания является одной из ведущих в области распознавания документов и обработки естественного языка.
- Для быстрого порождения предварительного списка гипотез используются признаковые классификаторы. Эти же классификаторы используются для повышения точности распознавания на изображениях с дефектами. Путем их комбинации выдвигаем гипотезу о том, что может быть на изображении. Каждый классификатор дает не один результат, а несколько лучших, которые объединяются в общий список. Получаем некий набор гипотез о том, что может быть на изображении. Далее гипотезы последовательно проверяются структурным классификатором, который целенаправленно анализирует имеющийся символ, исходя из знаний о его структуре. То есть, когда мы предполагаем, что на изображении может быть буква "а", мы можем целенаправленно проверить те свойства, которые должны быть именно у буквы "а", а не у какой-то другой буквы, сравнивая имеющийся у нас символ со структурным эталоном.
- В FineReader анализ документа проводится как до, так и после непосредственно распознавания, что позволяет гораздо лучше сохранять внешний вид документа при его экспорте в другие приложения из FineReader. В результате использования совмещенной процедуры значительно улучшилось выделение таблиц и отделение текста от графики.

- Дает возможность ввести документ в компьютер посредством нажатия всего на одну кнопку.
- Имеется возможность экспортировать распознанный текст в текстовый редактор или электронную таблицу, а также сохранить его в формате PDF или HTML.
- Имеется возможность сохранять цвета распознанного текста в форматах RTF, PDF и HTML.
- Встроенная технология «адаптивного распознавания»: Необычайно высокая точность распознанных текстов и малая чувствительность к дефектам печати.
- Распознанные страницы представляются миниатюрными изображениями.
- Имеется возможность сканировать разворот книги и распознавать ее каждую страницу по отдельности, при этом, изображение, содержащее сдвоенные страницы, сохраняется в две различные страницы пакета.
- Встроенный алгоритм автоматического поиска блоков (участков изображения, выделенных в рамку) распознаваемого текста: Анализ отсканированного материала и его распознавание происходит одновременно.
- Программа «видит» изображения в распознаваемом макете.
- 176 языков распознавания.
- Распознавание языков программирования (Basic, Cobol, Fortran, Java, C++, Pascal).
- Распознавание подстрочных символов и вертикального текста.
- Поддержка кодировки Unicode при сохранении распознанного текста в форматах RTF, DOC, XLS, HTML, TXT и CSV.