

Оптическое распознавание СИМВОЛОВ

Боровикова Ксения ФЛ-401

г. Ростов-на-Дону
2011г.

A hand is shown using a mouse on a laptop keyboard. The background is a blurred image of a person's hands typing on a laptop, with a blue and white color scheme. Overlaid on the image are binary code characters (0s and 1s) in a light blue color, creating a digital or data-related atmosphere.

Оптическое распознавание

- это механический или электронный перевод изображений рукописного, машинописного или печатного текста в последовательность кодов, использующихся для представления в текстовом редакторе.



Плюсы

- Перевод в текстовый вид очень сложных текстовых документов, включая таблицы, формы, диаграммы, рисунки
- Проверка орфографии
- Сохранение файла в любом текстовом формате, а также в форматах электронных таблиц и баз данных



OCR

- optical character recognition
(оптическое распознавание символов)



История OCR

- Программы появились сначала в США, а затем, и в России. OCR - optical character recognition (оптическое распознавание символов)
- 1978 год - компания «Курцвейл Компьютер Продактс» начала продажи коммерческой версии компьютерной программы оптического распознавания символов



Программы OCR

- Recognita Plus DTK фирмы Recognita Corporation □, Венгрия
- TextBridge фирмы □ Xerox Imaging Systems
- TypeReader фирмы ExperVision □, США
- CharacterEyes фирмы Ligature, Израиль
- IRIS OCR фирмы I.R.I.S., Бельгия
- Easy Reader фирмы □ Inovatic International, Франция
- WordScan Plus и OmniPage Professional □ фирмы Caera, США

A hand is shown using a mouse on a laptop keyboard. The background is a blue-tinted image with binary code (0s and 1s) overlaid, suggesting a digital or data processing theme.

Системы OCR в России

- OCR CuneiForm;
- ABBYY FineReader

A hand is shown using a mouse on a laptop keyboard. The background is a blurred image of a person working at a computer, overlaid with a pattern of binary code (0s and 1s) in a light blue color. The overall tone is professional and technological.

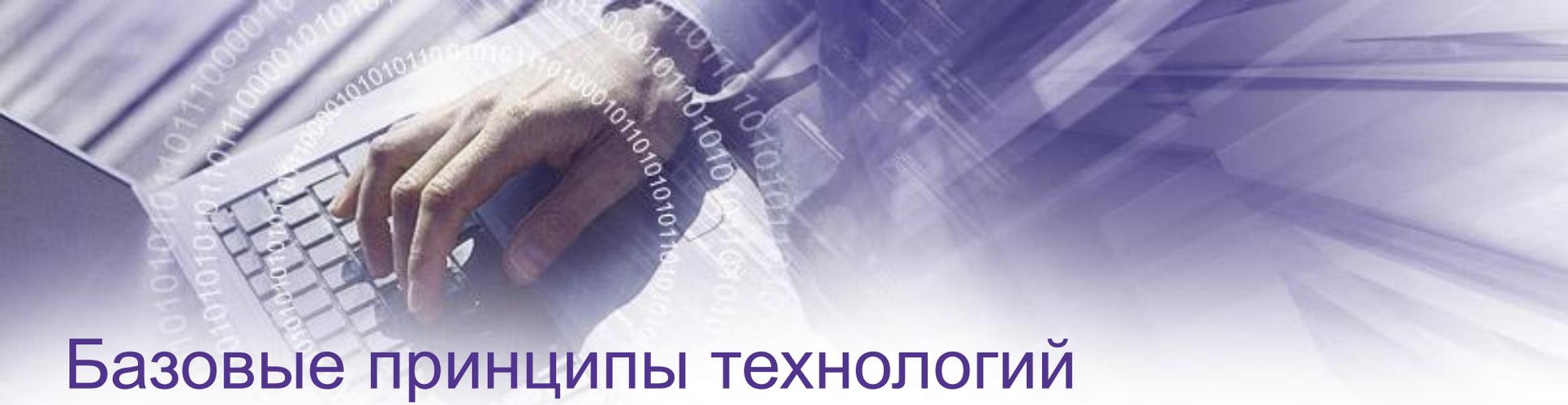
Этапы преобразования документа в электронный вид OCR-системами

- Сканирование и предварительная обработка изображения
- Анализ структуры документа
- Распознавание
- Проверка результатов
- Реконструкция документа (воссоздание его исходного вида)
- Экспорт



Описание OCR-процедуры

1. Предварительная обработка изображения
2. Распознавание объектов высших уровней.
Бинаризация
3. Распознавание символов
4. Структурирование гипотез. Словарная проверка
5. Синтез электронного документа



Базовые принципы технологий распознавания текста

- Принципы ИРА:
- Целостность (integrity)
- Целенаправленность (purposefulness)
- Адаптивность (Adaptability)
- Многоуровневый анализ документа



Принципы ИРА

Согласно принципу целостности (integrity), созерцаемый объект рассматривается как целое, состоящее из связанных частей



Принципы ИРА

Принцип целенаправленности (purposefulness) - любая интерпретация данных преследует определенную цель. Согласно этому принципу, распознавание представляет собой процесс выдвижения гипотез о целом объекте и целенаправленной их проверки



Принципы ИРА

Принцип адаптивности (adaptability) подразумевает способность системы к самообучению.

