Распознавание текста

OCR-системы

Лекция №6

Лингвистическое обеспечение АСОИУ

К.т.н., доцент Филиппович Анна Юрьевна

Системы оптического распознавания текста

- Системы оптического распознавания текста – OCR-системы (optical character recognition) предназначены для ввода печатного текста для печатных и электронных изданий.

Примеры системы оптического распознавания текста

- ***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),
- ***OmniPage Professional*** и ***WordScan Plus*** фирмы Caera (США).

Системы оптического распознавания текста

Наиболее известными программами класса «Системы оптического распознавания» в России являются:

- *OCR CuneiForm;*
- *ABBYY FineReader*.

Характеристики OCR-системы ABBYY FineReader

- Интуитивно понятный интерфейс.
- Поддержка различных сканеров. Использование интерфейса TWAIN.
- Импорт различных изображений (BMP, DCX, JPEG, PCX, PNG, TIFF).
- Фрагментация изображений: полуавтоматическая, автоматическая, ручная.
- Распознавание цифровых фотографий документов.
- Преобразование PDF-файлов в том числе создание PDF-файлов с тегами.
- Языковая поддержка: 176 языков (36 с поддержкой проверки орфографии).

Характеристики OCR-системы ABBYY FineReader

- Высокое качество распознавания.
- Словарный контроль: словарь общеупотребительной лексики, возможность создания и подключения дополнительного словаря.
- Автоматическая обработка документов.
- Сохранение и экспорт результатов. Поддержка Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Lotus Word Pro, Corel WordPerfect, Sun StarWriter. Интеграция с Microsoft Word.
- ABBYY Screenshot Reader.
- Многоколоночный WYSIWYG-редактор.
- Полнотекстовый морфологический поиск.

Этапы преобразование документа в электронный вид OCR-системами

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

Базовые принципы технологий распознавания текста

- Принципы IPA:
  - Целостность (integrity);
  - Целенаправленность (purposefulness);
  - Адаптивность (adaptability).

- Многоуровневый анализ документа.

Принципы IPA

- **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.
- Преимущество системы, следующей вышеописанным правилам, выражается в способности точнее классифицировать распознаваемый объект, исключая из рассмотрения сразу множество гипотез, противоречащих хотя бы одному из положений принципа.

Принципы IPA

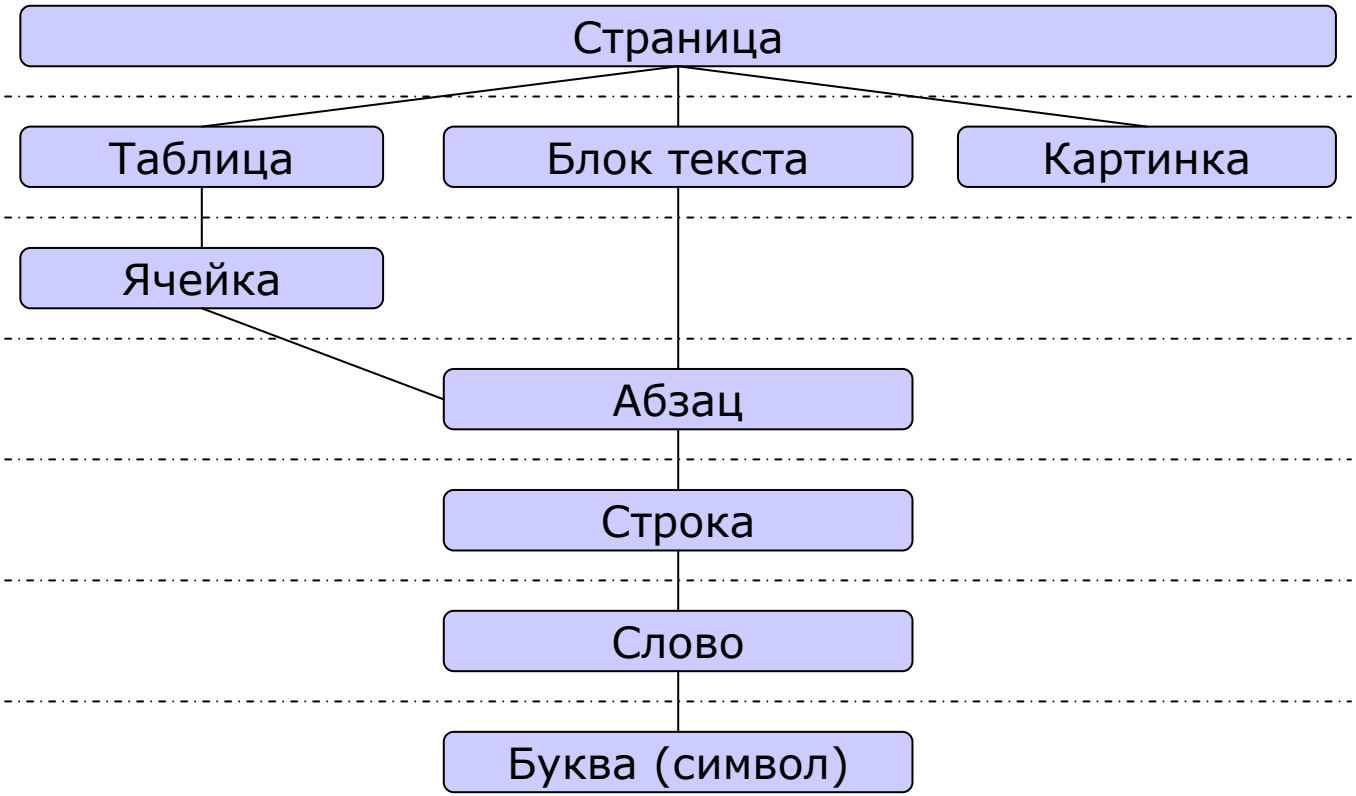
- **Принцип целенаправленности (purposefulness)** : любая интерпретация данных преследует определённую цель.
- Следовательно, распознавание должно представлять собой процесс выдвижения гипотез о целом объекте и целенаправленной их проверки.

Принципы IPA

- **Принцип адаптивности (adaptability)** подразумевает способность системы к самообучению.
- Полученная при распознавании информация упорядочивается, сохраняется и используется впоследствии при решении аналогичных задач.

Многоуровневый анализ документа

- Иерархическая структура документа:



Многоуровневый анализ документа

- Большинство современных OCR-систем ведёт анализ документа в соответствии с одним из следующих принципов: top - down («сверху вниз») либо bottom - up («снизу вверх»).

Механизм «голосования»

- Для принятия решений относительно того или иного объекта нередко используется механизм так называемого «голосования», voting.
- Суть данного метода заключается в параллельной выработке нескольких гипотез относительно объекта и передаче их «эксперту», логическому блоку, выбирающему одну из них.
- Такой метод широко практиковался на протяжении последнего десятилетия; в частности, существуют OCR-системы, буквально составленные из двух или более независимых механизмов распознавания с общим «экспертом» на выходе.
- Собственно же процедура «голосования» в наши дни используется по-прежнему широко, но не на уровне архитектуры систем, а для решения несложных, рядовых задач.

Многоуровневый анализ документа

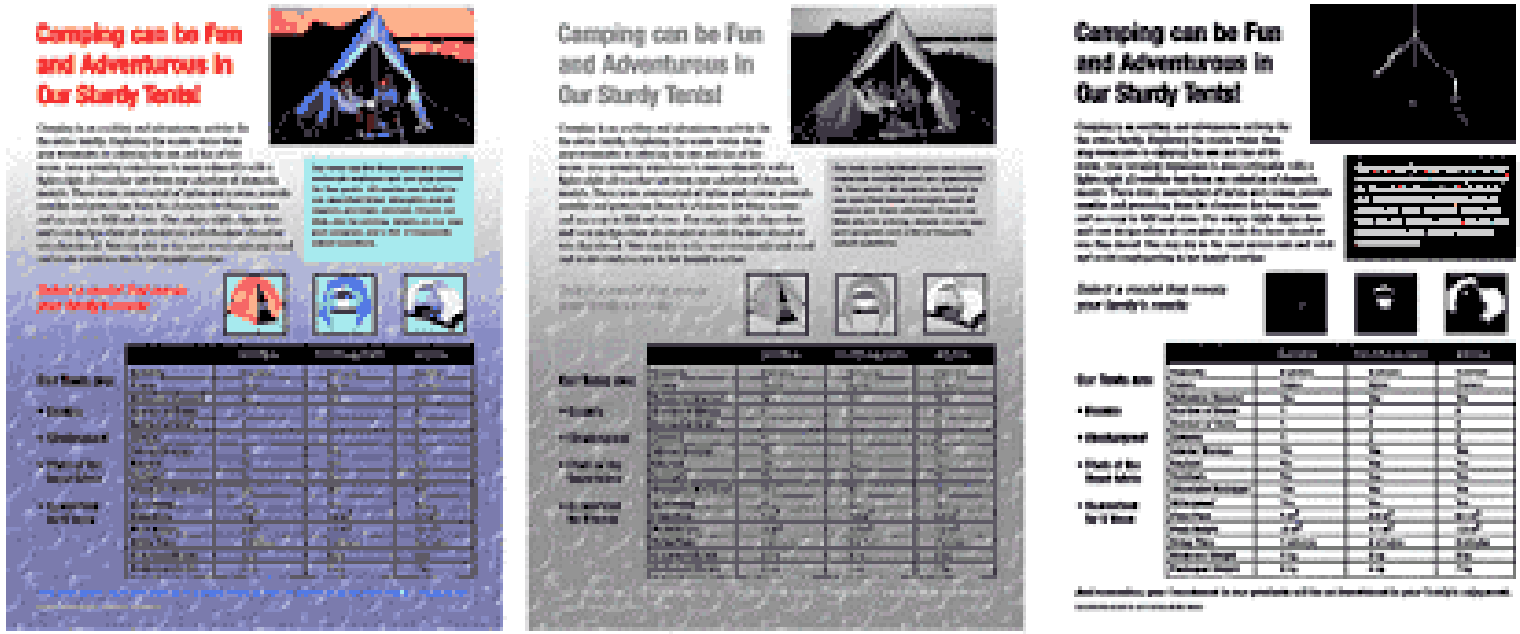
- В частности ABBY использует многоуровневый анализ документа –MDA (multilevel document analysis) – позволяющий объединить преимущества обоих вышеописанных принципов.
- В рамках MDA структура страницы рассматривается подобно тому, как это делается по методу top - down , а воссоздание документа в электронном виде по окончании распознавания ведётся «снизу вверх».
- При этом в алгоритм добавлен механизм обратной связи, охватывающей все уровни анализа.

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.

Специальные процедура фильтрации фоновых текстур

Обработка процедурой
интеллектуальной фильтрации фоновых текстур,
(intelligent background filtering , IBF).

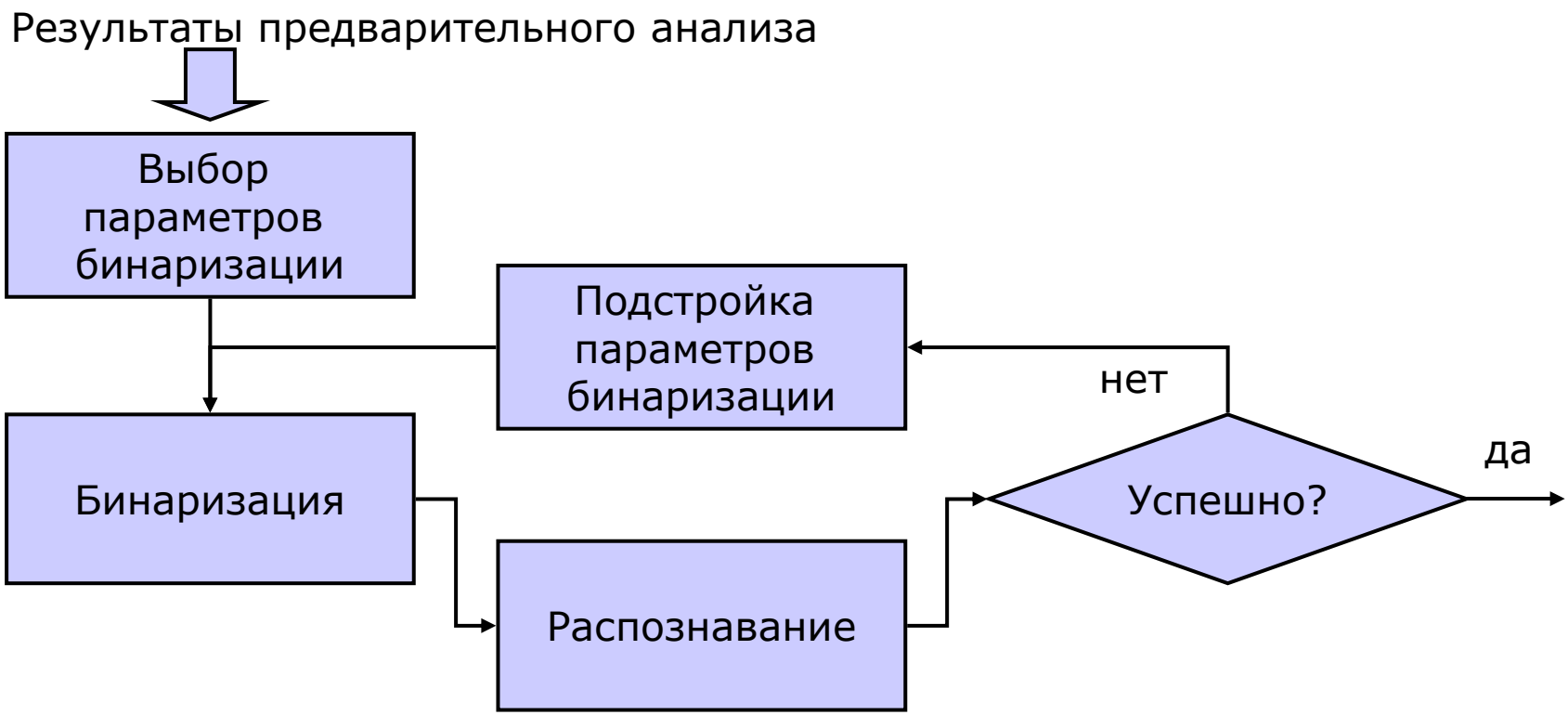


Адаптивная бинаризация

- Процедура **адаптивной бинаризации** (adaptive binarization , **AB**) способна гибко выбирать оптимальные для конкретного участка (фрагмента строки или даже слова) параметры бинаризации.

Адаптивная бинаризация

Обобщённая блок-схема
алгоритма процедуры адаптивной бинаризации



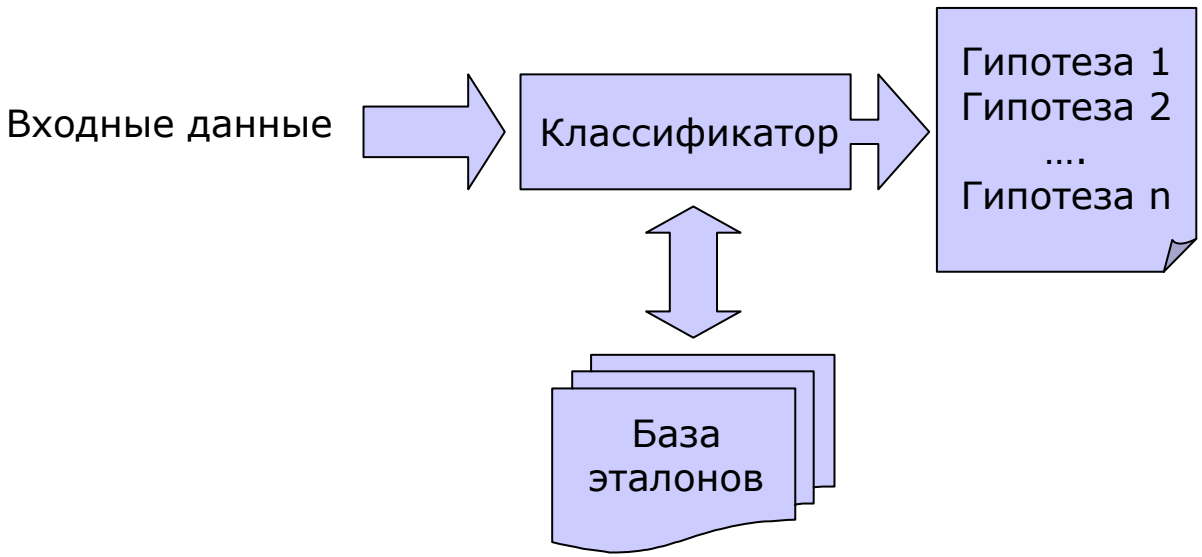
Распознавание символов

- Деление строки на слова и слов на буквы в программном ядре OCR-системы выполняется так называемой процедурой линейного деления.
- Процедура завершается по достижении конца строки и передаёт для дальнейшей обработки список гипотез, выдвинутых относительно возможных вариантов деления.
- При этом каждой гипотезе приписывается определённый вес; по смыслу эта величина соответствует численному выражению уверенности.
- Соответствующий каждой из гипотез набор графических объектов уровня «символ» поступает на вход механизма распознавания символов.

Классификатор

- Механизм распознавания символов представляет собой комбинацию ряда элементарных распознавателей, называемых классификаторами.

Упрощённая схема работы классификатора



Характеристики классификатора

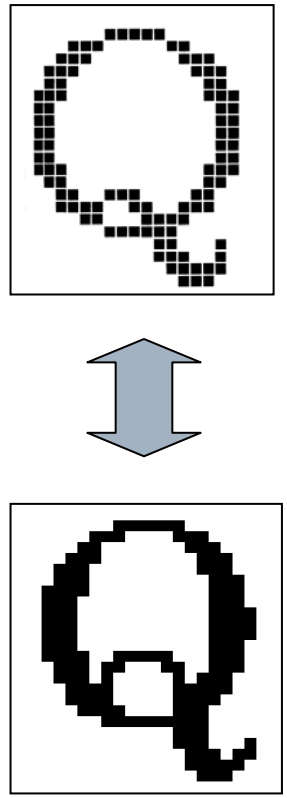
- Среднее положение правильной гипотезы;
- Точность по первому варианту распознавания;
- Быстродействие;
- Простота реализации;
- Устойчивость к различным искажениям.

Типы классификаторов

- Растровый.
- Признаковый.
- Признаковый дифференциальный.
- Контурный.
- Структурный.
- Структурный дифференциальный.

Растровый классификатор

- Принцип действия основан на прямом сравнении изображения символа с эталоном.
- Степень несходства при этом вычисляется как количество несовпадающих пикселей.
- Для обеспечения приемлемой точности растрового классификатора требуется предварительная обработка изображения: нормализация размера, наклона и толщины штриха.
- Эталон для каждого класса обычно получают, усредняя изображения символов обучающей выборки.



Растровый классификатор

+

- Простота реализации.
- Высокое быстродействие.
- Хорошая устойчивость к случайным дефектам изображения.

–

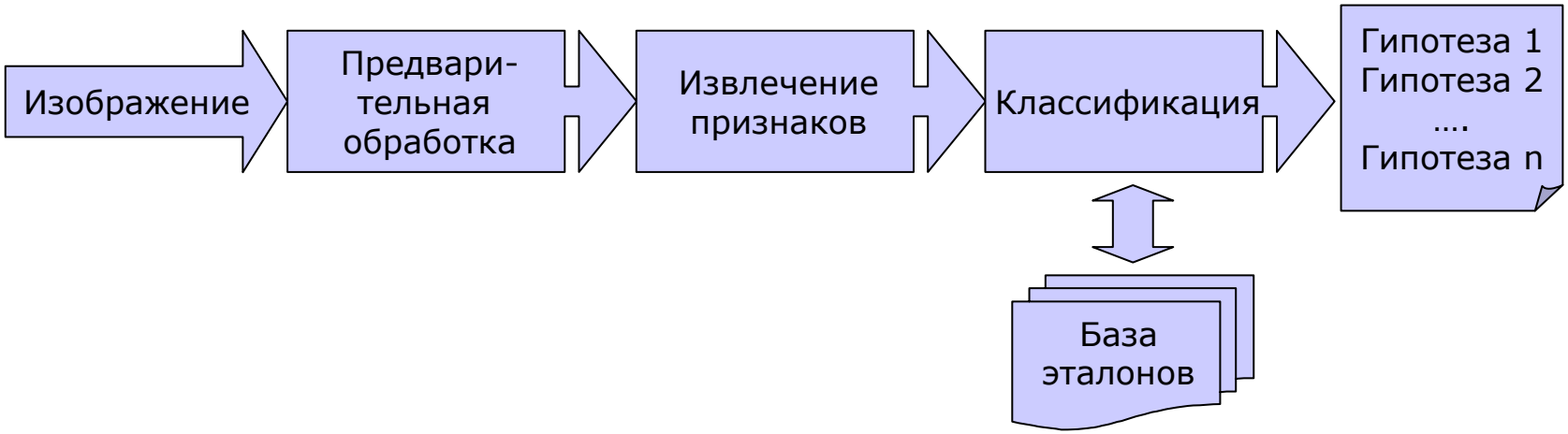
- Невысокая точность распознавания.

Признаковый классификатор

- Принцип действия: изображению ставится в соответствие N-мерный вектор признаков.
- Собственно классификация заключается в сравнении его с набором эталонных векторов той же размерности. Тип и количество признаков в немалой степени определяют качество распознавания.
- Извлечение признаков - формирование вектора (вычисление его координат в N -мерном пространстве) производится во время анализа предварительно подготовленного изображения.
- Эталон для каждого класса получают путём аналогичной обработки символов обучающей выборки.

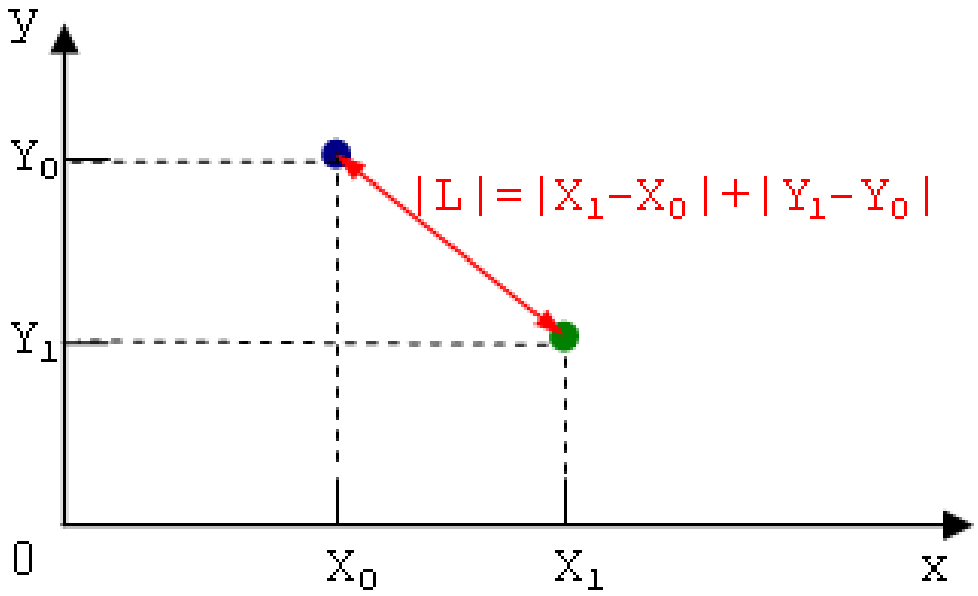
Признаковый классификатор

Блок-схема работы признакового классификатора



Признаковый классификатор

- Сравнение каждой пары векторов заключается в вычислении оценки, характеризующей расстояние между точками в N-мерном пространстве (точка – геометрическое представление такого вектора).



$$|L| = \sum_{i=1}^{N} |K_i - K_{0i}|$$

Признаковый классификатор

+

- Простота реализации.
- Хорошая обобщающая способность.
- Хорошая устойчивость к изменениям формы символов.
- Низкое число отказов от распознавания.
- Высокое быстродействие.

–

- Неустойчивость к различным дефектам изображения.
- На этапе извлечения признаков происходит необратимая потеря части информации о символе. Извлечение признаков ведётся независимо, поэтому информация о взаимном расположении элементов символа утрачивается.

Контурный классификатор

- Обособленная разновидность признакового классификатора.
- Отличается от последнего тем, что для извлечения признаков использует контуры, предварительно выделенные на изображении символа.
- Принципы функционирования, основные достоинства и недостатки такие же, как и у признакового классификатора.
- Контурный классификатор предназначен для распознавания текста, набранного декоративными шрифтами.
- Работает несколько медленнее обычного признакового классификатора.

Признаковый дифференциальный классификатор

- Предназначен для различения похожих друг на друга объектов, таких, например, как буква «m» и сочетание «rn».
- Анализирует только те области изображения, где может находиться информация, позволяющая отдать предпочтение одному из вариантов.
- Так, в случае с «m» и «rn» ключом к ответу служит наличие и ширина разрыва в месте касания предполагаемых букв.

Признаковый дифференциальный классификатор

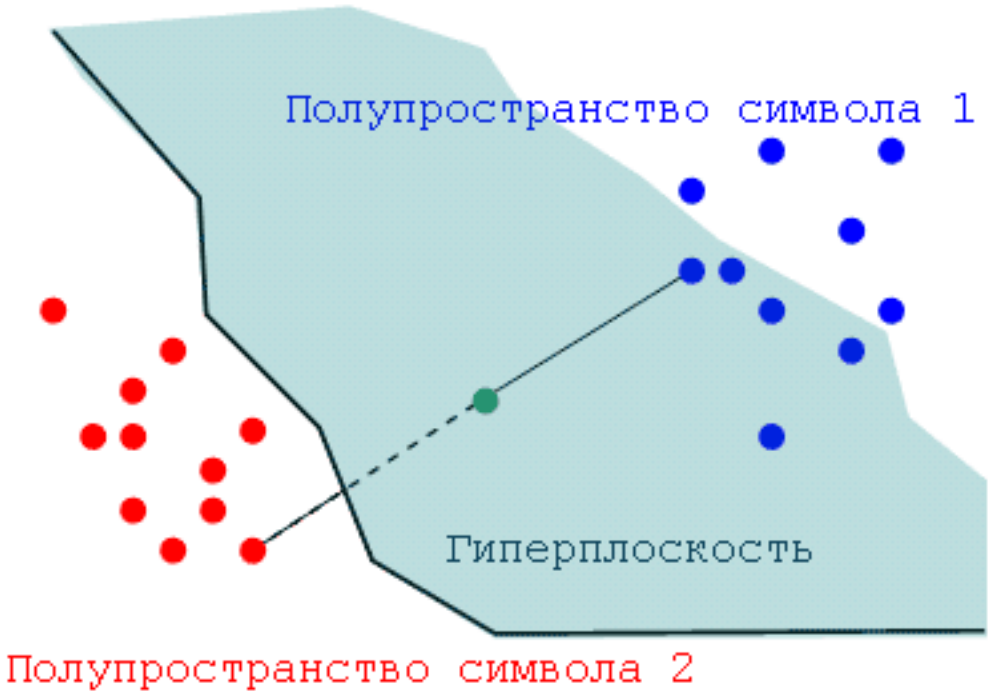
- Признаковый дифференциальный классификатор представляет собой набор признаковых классификаторов, которые оперируют эталонами, полученными для каждой пары схожих символов.
- Для всех пар используется один и тот же набор признаков, аналогичный имеющемуся у соответствующего признакового классификатора.
- В процессе обучения этого классификатора производится анализ изображений из обучающей базы. Вычисляемые при этом значения признаков интерпретируются как координаты точки в N-мерном пространстве.

Признаковый дифференциальный классификатор

- Соответственно, для двух различных символов получается два «облака» точек, расположенные на некотором удалении друг от друга.
- Когда накоплена информация о достаточном количестве точек, выполняется вычисление координат гиперплоскости. Она должна разделить пространство таким образом, чтобы «облака» оказались по разные стороны и примерно на одном расстоянии от гиперплоскости.
- Для полученных при анализе изображения значений вычисляется оценка, геометрический смысл которой – местонахождение точки относительно гиперплоскости.

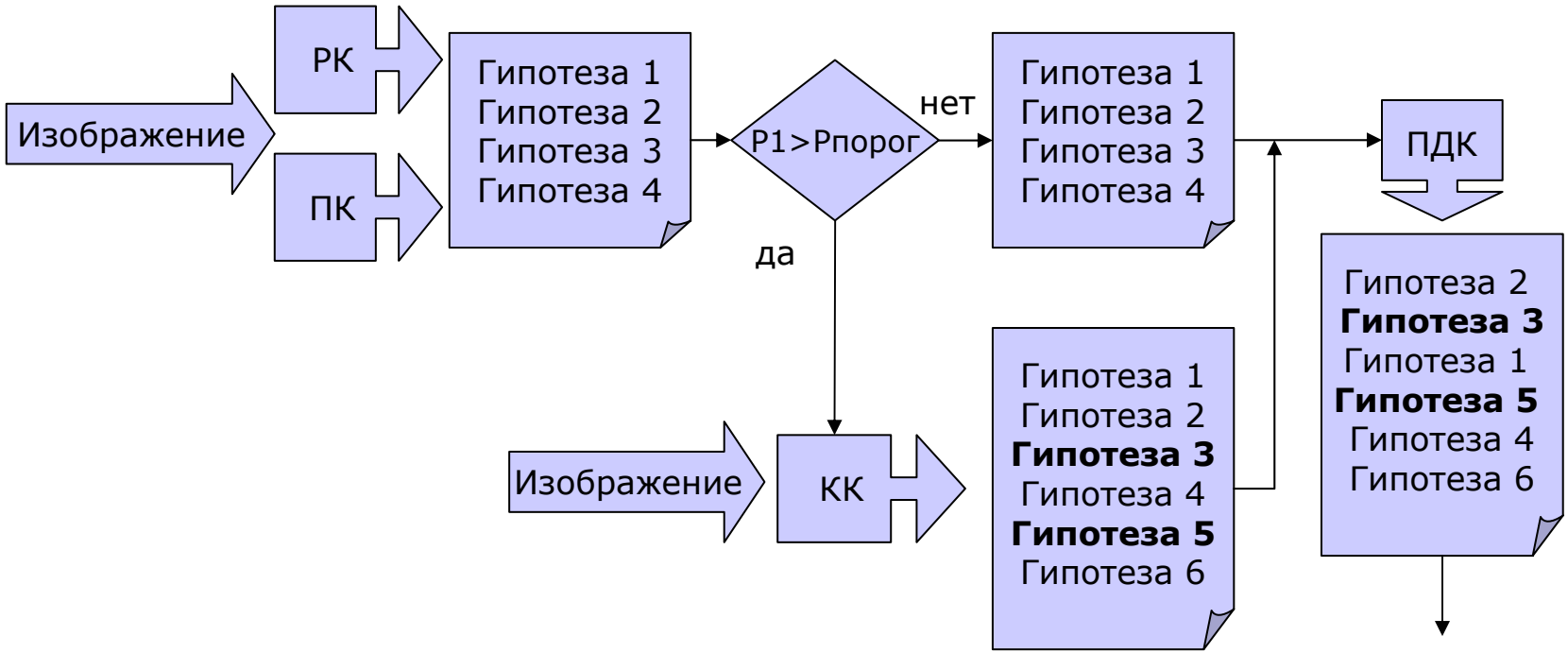
Признаковый дифференциальный классификатор

Упрощённая геометрическая модель обучения дифференциального классификатора

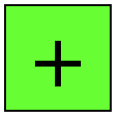


Алгоритм распознавания

Обобщённая блок-схема алгоритма распознавания
(первый уровень)



Признаковый дифференциальный классификатор



- Высокая точность распознавания.

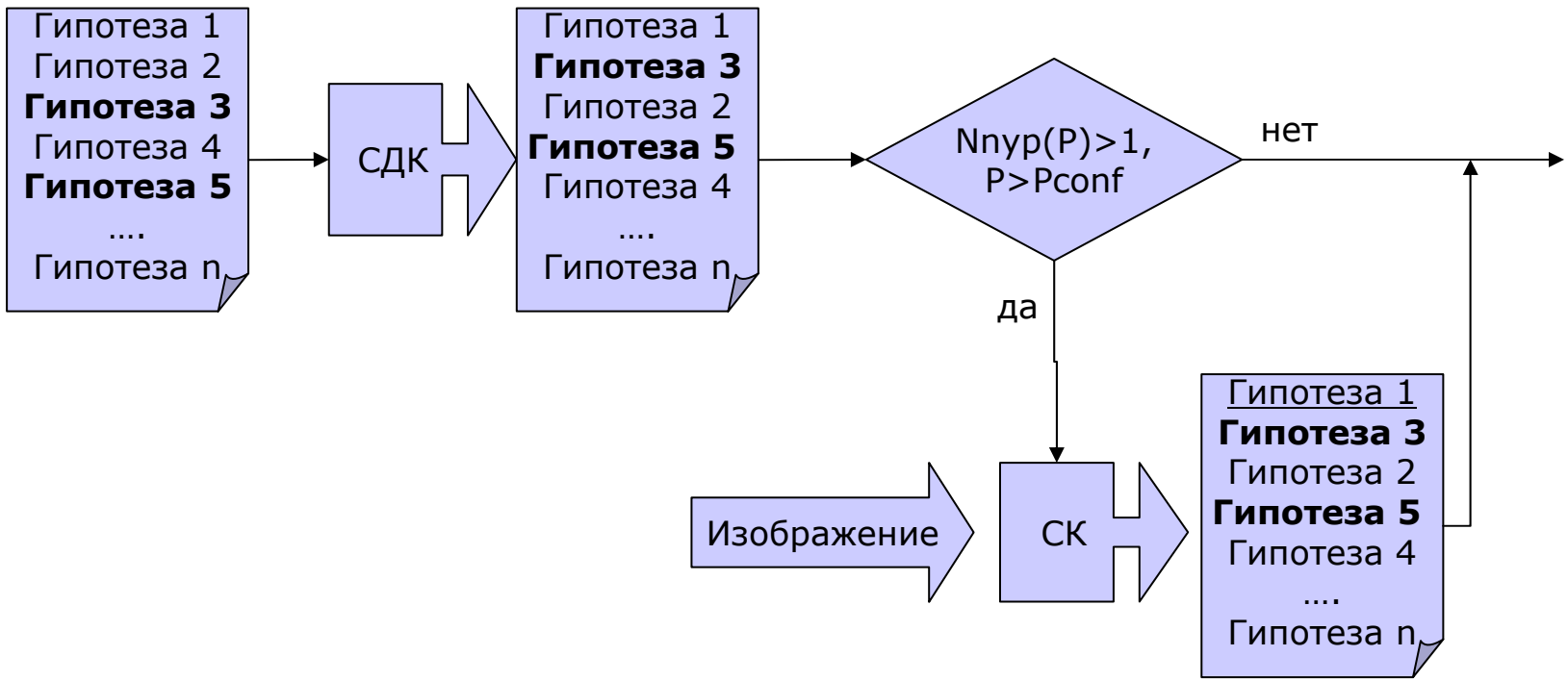
- Высокая трудоёмкость построения эталонов для классификатора.
- Низкое быстродействие.

Структурный классификатор

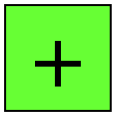
- Входными данными для структурного классификатора являются изображение символа и ранжированный список гипотез, сформированный по результатам работы остальных распознавателей.
- Собственных гипотез классификатор не выдвигает, подтверждая либо опровергая ранее выдвинутые гипотезы.
- Используется в тех случаях, когда в списке присутствуют две или более гипотез, веса которых не только превышают заданный уровень уверенности, но и сравнимы между собой.

Структурный классификатор

Обобщённая блок-схема алгоритма распознавания (структурный уровень).



Структурный классификатор



- Очень высокая точность распознавания.

- Низкое быстродействие.

Структурирование гипотез

Схема структурирования гипотез

Словарная проверка

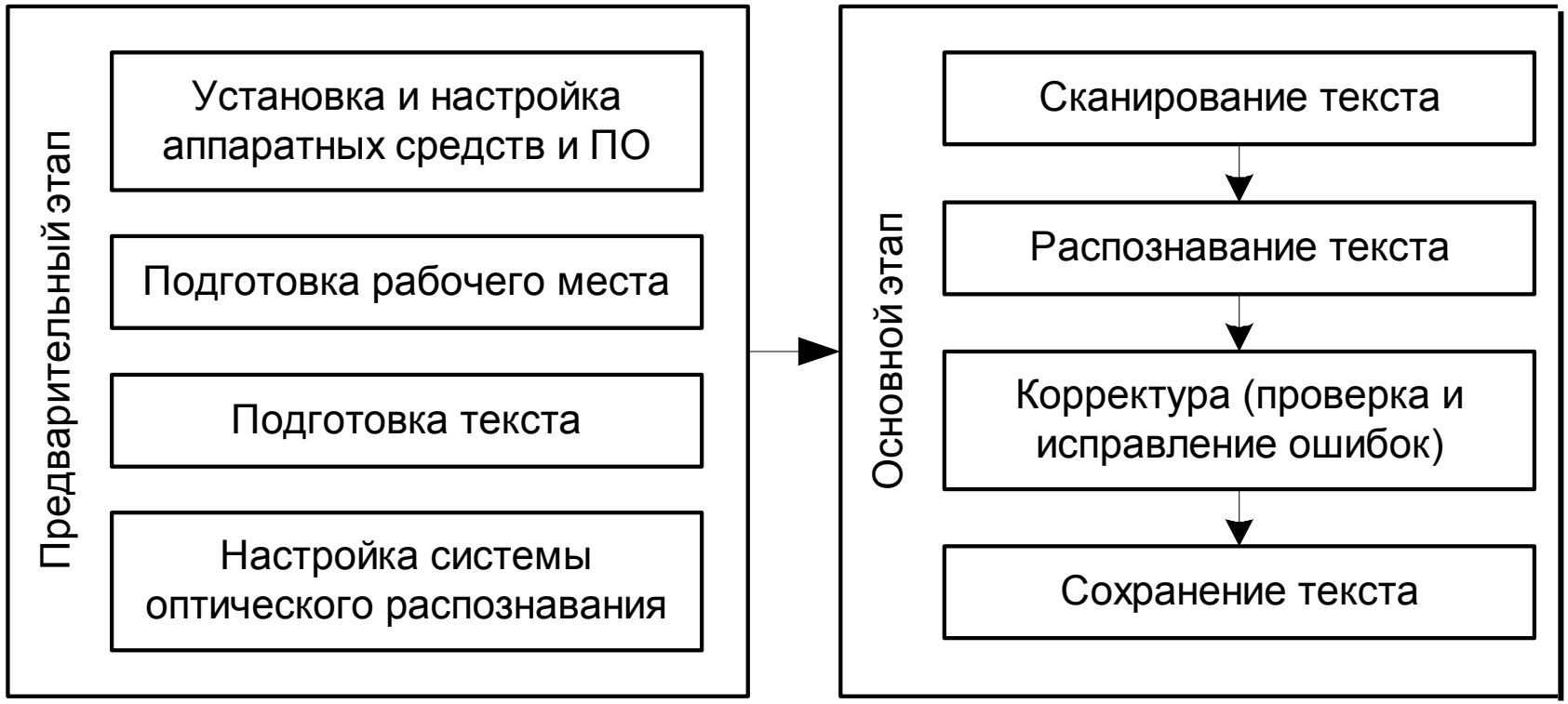
- Зачастую точность распознавания практически не зависит от полноты словарной базы проверочных словарей.
- Ведь, как известно, не существует словарей, содержащих все словоформы живого языка.
- Например в системах ABBYY для решения проблемы словарной проверки предусмотрен особый тип модели несловарное слово. При встрече с несловарным словом система распознает его в точности так, как оно было написано.
- Во-вторых для всех поддерживаемых языков распознавания были созданы морфологически структурированные словари. Каждый из них способен моделировать различные словоформы, в том числе и композиты, за счёт чего охватывает более 98% реального словарного объёма соответствующего языка.

Синтез электронного документа

- По окончании «сборки» слов, объединения их в строки, а строк – в объекты высших уровней, OCR-система, выводит на экран полученный электронный документ. Пользователь видит точную электронную копию страницы, в особом окне доступен для сравнения отсканированный образ документа; при необходимости результаты распознавания можно отредактировать средствами встроенного WYSIWYG-редактора.

Ввод текста с помощью OCR-систем

Обобщенная схема технологического процесса ввода текста с помощью OCR-систем



Оценка качества распознавания текста

Характеристики, определяющие качество ввода (распознавания) текста:

- точность распознавания;
- временные затраты;
- статистические параметры.

Исследование эффективности OCR-систем для ввода текста

Исследование эффективности ввода текста с помощью OCR-системы будет включать в себя следующие компоненты:

- исследование временных затрат;
- статистическое исследование количества ошибок;
- анализ эффективности ввода текста.

При этом рассматриваются следующие виды текстов:

- современный текст хорошего качества;
- современный текст плохого качества;
- старинный текст XVIII в.

Исследование временных затрат

Сравнение временных затрат
на этапы ввода одной страницы текста

Вид текста	**Сканирование**	**Распознавание**
современный текст хорошего качества	62 с.	29 с.
современный текст плохого качества	61 с.	30 с.
текст XVIII в.	52 с.	32 с.

Время сканирования и распознавания зависит от множества факторов: **характеристик сканера**, *производительности системы* (скорости работы процессора, объема оперативной памяти и т.д.), от **особенностей текста**, **качества оригинала**, **шрифта** и т.п.

Точность распознавания

- Одним из основных параметров качества функционирования системы распознавания является точность распознавания, обычно выражаемая процентным соотношением:

$$Ac_{расп_i} = \frac{100\% \cdot n_{верно\_расп_i}}{n_{общ_i}}$$

где $n_{верно\_расп\ i}$ и $n_{общ\ i}$ есть количество верно распознанных символов и общее количество символов на странице (в документе).

Статистическое исследование количества ошибок

Текст хорошего качества

Фрагмент по 10 страниц	**Кол-во знаков (символов)** $n_{общ}$	**Кол-во слов**	**Кол-во неуверенно распознанных символов**	**Кол-во ошибок** n_o	**Точность распознавания** $Ac_{расп}$ **(%)**
1	26377	3344	62	7	99,97 %
2	27266	3422	39	3	99,99 %
3	29809	3865	38	13	99,96 %
4	26796	3324	267	17	99,94 %
5	24361	3445	88	7	99,97 %
6	26597	3343	78	4	99,98 %
…					
10	26800	3410	63	4	99,99 %
Среднее	**27251**	**3484**	**102**	**8**	**99,97 %**

Статистическое исследование количества ошибок

Текст плохого качества

Фрагмент по 10 страниц	**Кол-во знаков (символов) $n_{общ}$**	**Кол-во слов**	**Кол-во неуверенно распознанных символов**	**Кол-во ошибок n_o**	**Точность распознавания $Ac_{расп}$ (%)**
1	19540	3085	126	75	99,62 %
2	25517	3668	138	31	99,88 %
3	33841	5290	604	427	98,74 %
6	19566	2869	548	35	99,82 %
….					
9	24953	3201	414	76	99,70 %
Среднее	22831	3394	234	79	99,61 %

Статистическое исследование количества ошибок

Текст XVIII в. (CAP)

Страница	**Кол-во знаков (символов)** $n_{общ}$	**Кол-во слов**	**Кол-во неуверенно распознанных символов**	**Кол-во ошибок** n_o	**Точность распознавания** $Ac_{расп}$ **(%)**
1 т.: 377-378	2005	328	304	220	89,03 %
2 т.: 19-20	2340	376	368	297	87,31 %
3 т.: 519-520	2097	305	366	248	88,17 %
5 т.: 43-44	2117	328	425	241	88,62 %
6 т.: 447-448	2060	351	375	277	86,55 %
1 т.: 319-320	1578	265	70	84	94,68 %
3 т.: 9-10	2343	311	458	200	91,46 %
3 т.: 137-138	2173	343	560	389	82,10 %
Среднее	**2065**	**314**	**428**	**286**	**86,00 %**

Статистическое исследование количества ошибок

Текст XVIII в. (САР)
с использованием распознавания с обучением

Стр.	**Кол-во знаков (символов)** $n_{общ}$	**Кол-во слов**	**Кол-во неуверенно распознанных символов**	**Кол-во ошибок** n_o	**Точность распознавания** $Ac_{расп}$ **(%)**
1 т.:1-2	2080		359	104	95,00%
1 т.:3-4	2186		255	102	95,33%
1 т.:9-10	2238		280	140	93,74%
1 т.:15-16	2286	357	376	90	96,06%
1 т.:31-32	2251	377	653	114	94,94%
1 т.:33-34	2239	285	597	107	95,22%
1 т.: 377-378	2005	328	175	96	95,21%
…	…	…	…	…	…
Среднее	**2247**	**349**	**388**	**103**	**95,16 %**

Статистическое исследование количества ошибок

Текст XVIII в. (CAP)
с использованием распознавания с обучением, словаря, с системой замен

Страница	**Кол-во знаков (символов)** $n_{общ}$	**Кол-во слов**	**Кол-во неуверенно распознанных символов**	**Кол-во ошибок** n_{oi}	**Точность распознавания** $Ac_{расп}$ **(%)**
1 т.:1-2	2080		180	54	99,47%
1 т.:3-4	2186		128	52	99,25%
1 т.:9-10	2238		140	90	99,33%
1 т.: 11-12	2308	370	216	61	99,56%
1 т.: 31-32	2251	377	327	64	99,71%
1 т.: 377-378	2005	328	109	45	97,76%
…	…	…	…	…	…
Среднее	**2247**	**349**	**197**	**53**	**98,93%**

Анализ типов ошибок, обнаруженный при вводе текста CAP

- Ошибки, связанные с качеством оригинала.
- Ошибки, связанные со сходством графем символов.
- Ошибки, связанные с некорректным выделением символов.