Распознавание букв и элементов письма в русской каллиграфии

Постановка задачи и подход к решению (work in progress)

Местецкий Л.М.

МГУ имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Challenge

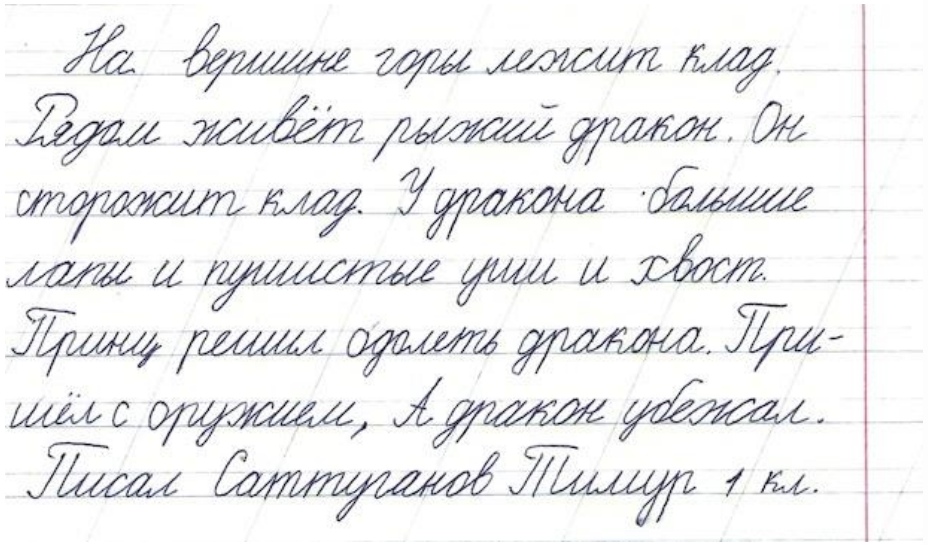
- Все, кто учился в школе, научились читать печатные тексты и научились писать рукописные тексты – есть отработанные **методики обучения**
- Дети не могут прочесть рукописные тексты, а взрослые могут – **как они научились**?
- Считается, что тот, кто умеет писать, тот может и читать. Методики обучения чтению рукописного текста на русском языке **не существует**!
- Задача автоматического распознавания рукописного текста – **обучение робота навыкам, которым людей обучать не умеем**!

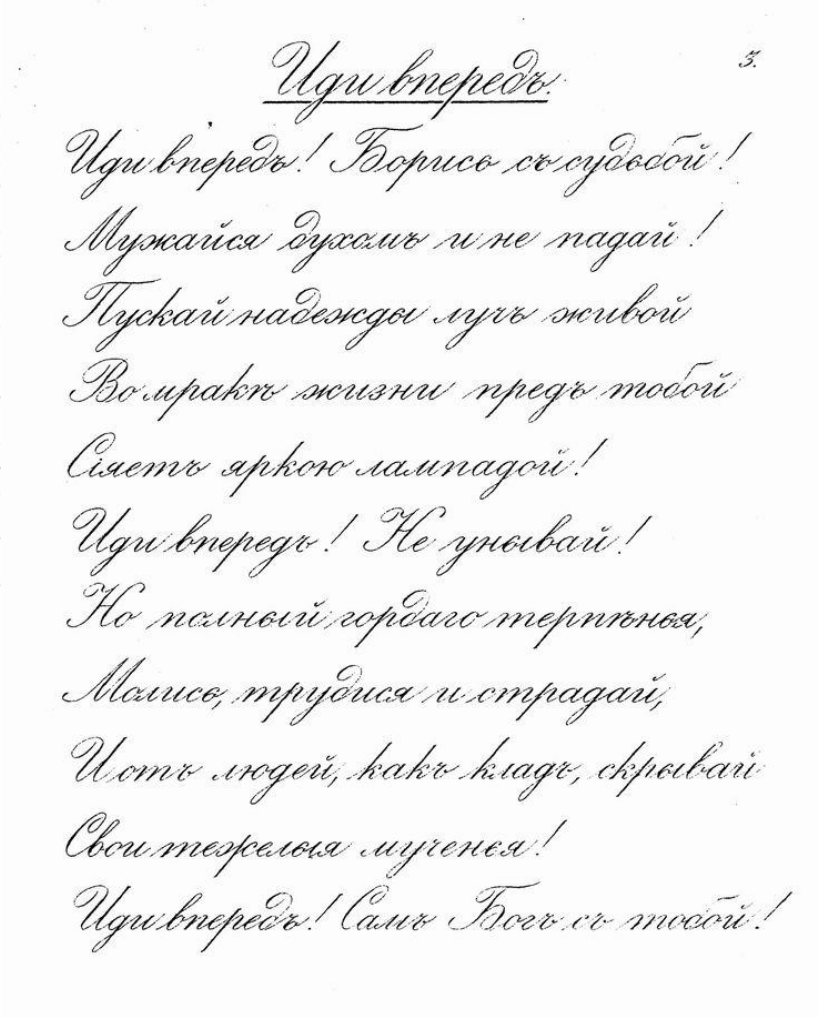
Каллиграфия

- Каллигра́фия — одна из отраслей изобразительного искусства, это искусство красивого письма.
- Мы употребляем это слово как набор навыков письма всех русскоязычных грамотных людей, полученных в процессе обучения в школе на уроках чистописания, с использований эталонных прописей. Эти навыки мало менялись за последние сто лет, как мало менялись сами прописи.
- Мы предполагаем, что все пишущие на русском языке делают это по тому стандарту, который представлен в прописях и на основе тех навыков, которые получили в школе, а отличия образуются вследствие индивидуальных свойств людей.

Русская каллиграфия

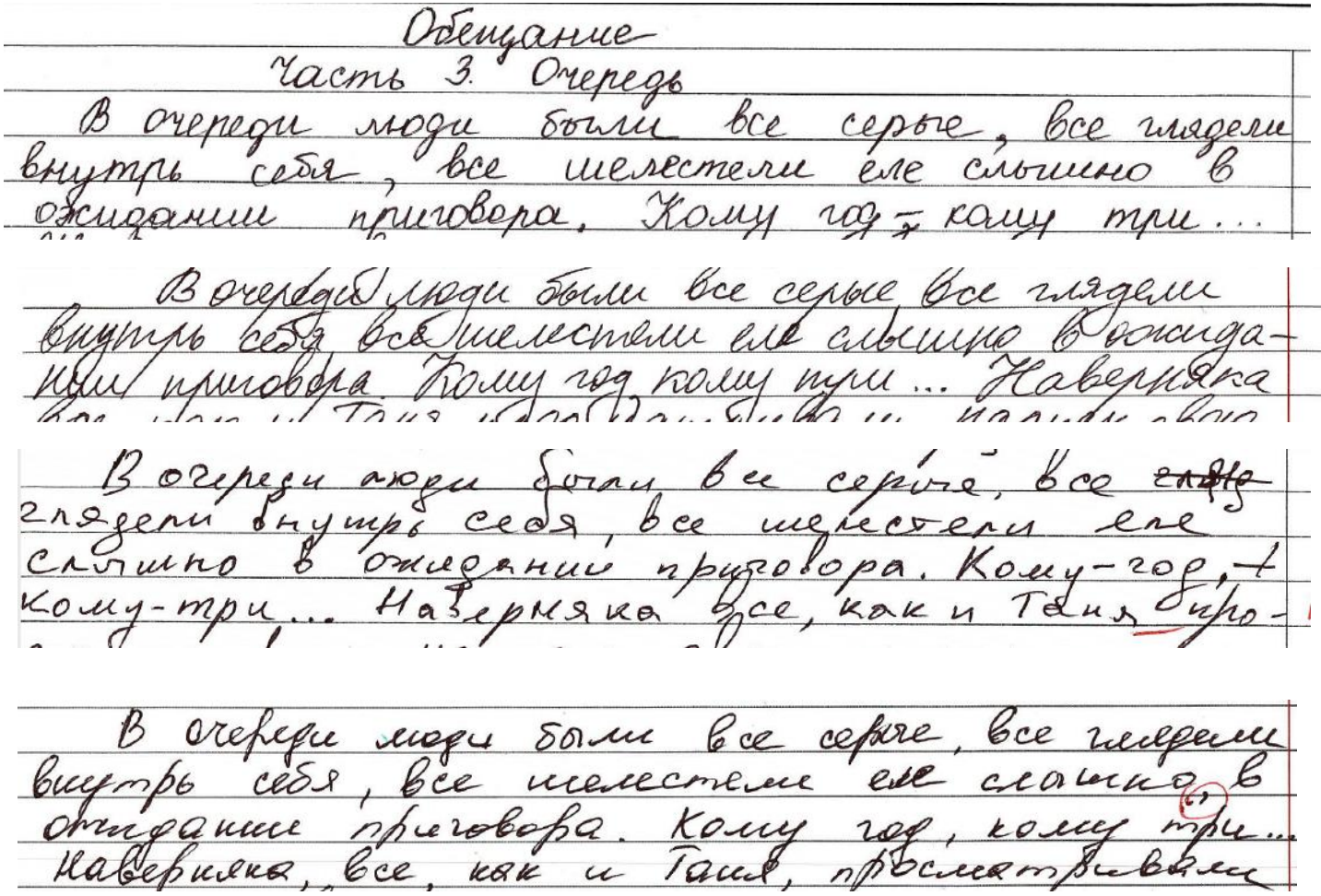
Каллиграфическое письмо


На вершине горы лежит клад.
Рядом живёт рыжий дракон. Он
сторожит клад. У дракона большие
лапы и пушистые уши и хвост.
Принц решил одолеть дракона. При-
шёл с оружием, А дракон убежал.
Писал Саттуганов Тимур 1 кл.


Иди вперёд. 3.

Иди вперёд! Борись со судьбой!
Мужайся духомъ и не падай!
Пускай надежды лучъ живой
Во мракъ жизни предъ тобой
Сіяетъ яркою лампадой!
Иди впередъ! Не унывай!
Но полный гордаго терпѣнья,
Молись, трудися и страдай,
Исть людей, какъ кладъ, скрывай
Свои тяжелыя мученья!
Иди впередъ! Самъ Богъ съ тобой!

Обычное письмо



Обещание

Часть 3. Очередь

В очереди люди были все серые, все глядели внутрь себя, все шелестели еле слышно в ожидании приговора. Кому год – кому три...

В очереди люди были все серые все глядели внутрь себя все шелестели еле слышно в ожидании приговора. Кому год кому три... Наверняка

В очереди люди были все серые, все ~~глядя~~ глядели внутрь себя, все шелестели еле слышно в ожидании приговора. Кому-год, кому-три... Наверняка все, как и Таня, про-

В очереди люди были все серые, все глядели внутрь себя, все шелестели еле слышно в ожидании приговора. Кому год, кому три... Наверняка, все, как и Таня, пребывали

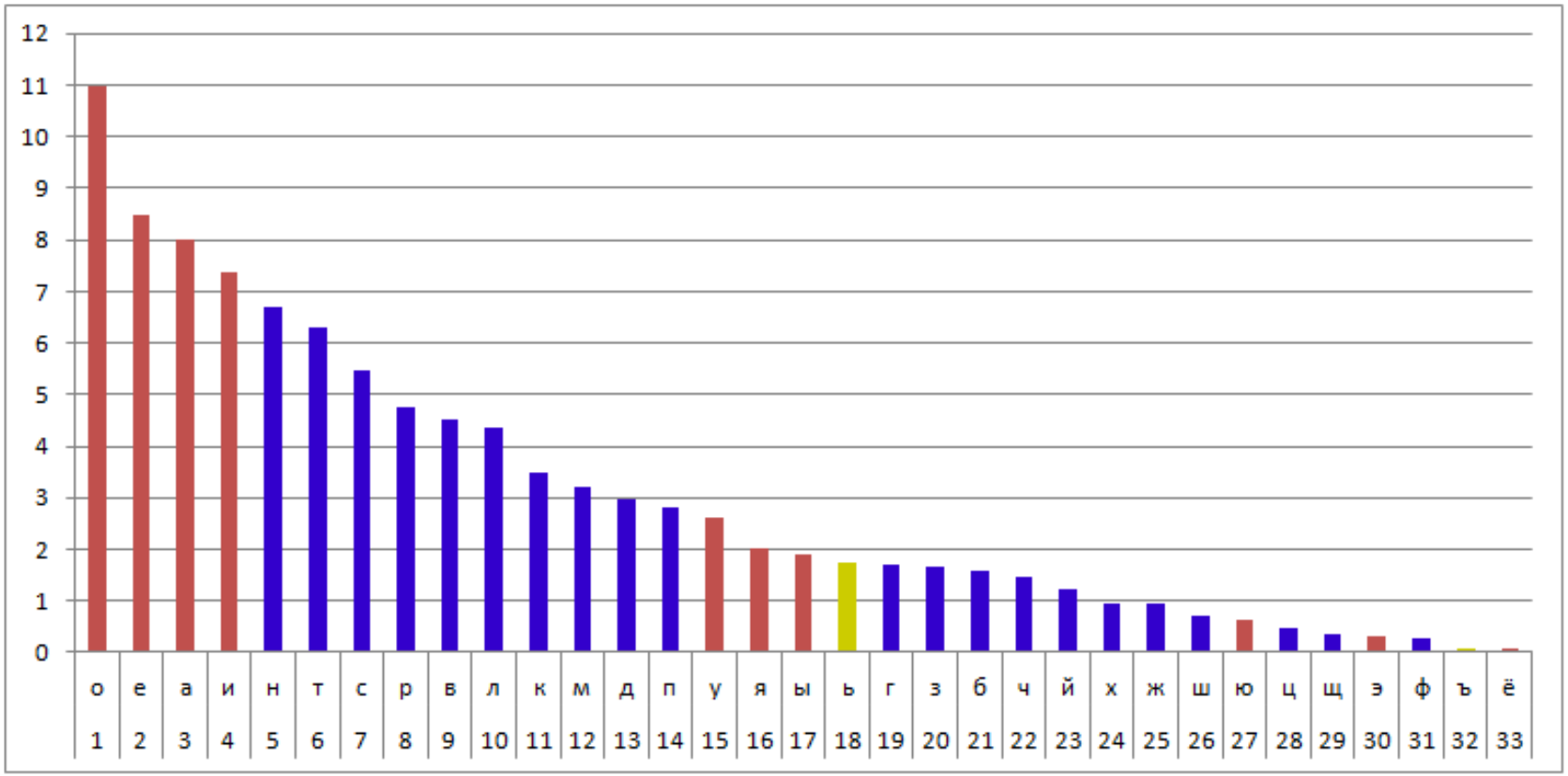
Задача

- Автоматизация оцифровки рукописных текстовых документов
- Поиск ключевых слов в рукописных документах

Подход

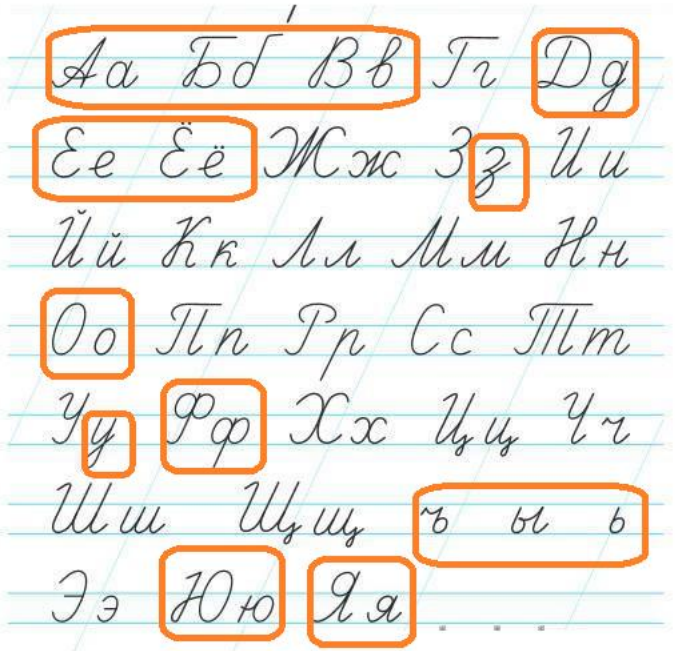
- Штриховое медиальное представление цифровых изображений текстовых документов
- Распознавание базовых рукописных штрихов – овалов, крючков, палочек и их сочетаний
- Машинное обучение на основе штрихового представления

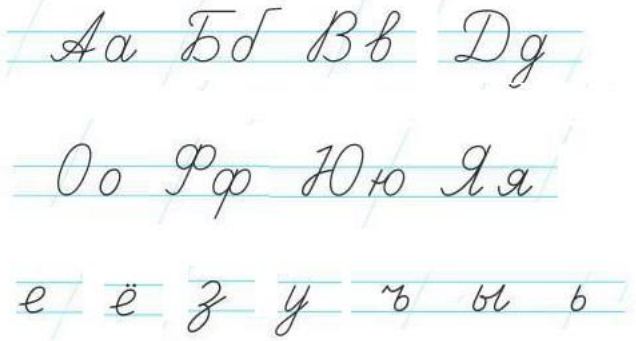
Частотное распределение букв



место	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
буква	о	е	а	и	н	т	с	р	в	л	к	м	д	п	у	я	ы	ь	г	з	б	ч	й	х	ж	ш	ю	ц	щ	э	ф	ъ	ё
%	10,98	8,48	8,00	7,37	6,70	6,32	5,47	4,75	4,53	4,34	3,49	3,20	2,98	2,80	2,62	2,00	1,90	1,74	1,69	1,64	1,59	1,45	1,21	0,97	0,94	0,72	0,64	0,49	0,36	0,33	0,27	0,04	0,01

Буквы с овальными элементами

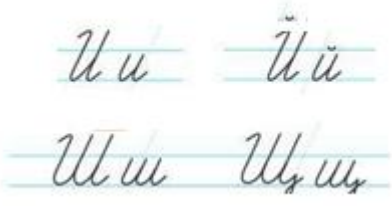




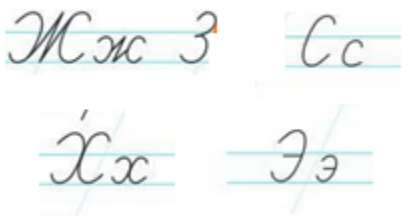
Овалы – 52.17%

Буквы с другими элементами «палочка», крючок, полуовал

Палочка – 19.31%



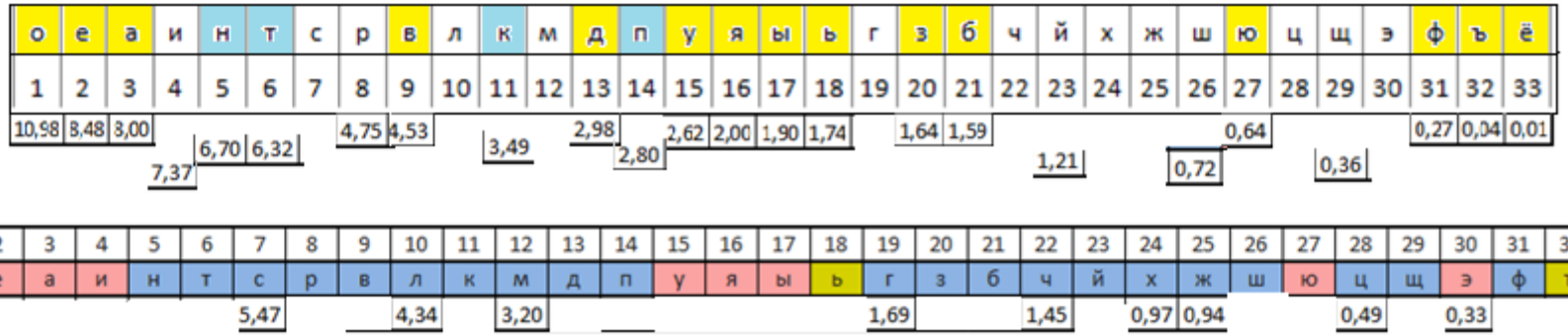
Крючок – 9.66 %



Полуовал – 7.71%

Распределение букв по каллиграфическим элементам

место	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
буква	о	е	а	и	н	т	с	р	в	л	к	м	д	п	у	я	ы	ь	г	з	б	ч	й	х	ж	ш	ю	ц	щ	э	ф	ъ	ё
%	10,98	8,48	8,00	7,37	6,70	6,32	5,47	4,75	4,53	4,34	3,49	3,20	2,98	2,80	2,62	2,00	1,90	1,74	1,69	1,64	1,59	1,45	1,21	0,97	0,94	0,72	0,64	0,49	0,36	0,33	0,27	0,04	0,01



Частота букв с овалами 52.17%

Частота букв с палочками 19.31% (всего 71.48%)

Частота букв с крючками 9.66% (всего 81.14%)

Частота букв с полуовалами 7.71% (всего 88.85%)

Восстановление текста по буквам

Буря ___ою _ебо _рое_,
В__р_ __е__ые _ру_я;
_о, _а_ зверь, о_а завое_,
_о за__а_е_, _а_ д__я,
_о _о _ров_е обве__а_ой
Вдру_ _о_о_ой за_у___,
_о, _а_ _у____ за_озда_ый,
_ _а_ в о_о__о за__у___.
_а_а ве__ая _а_у__а
_ _е_а_ь_а _ _е__а.
__о _е _ы, _оя __ару__а,
_р_у_о___а у о__а?
___ бур_ завыва_ье_
_ы, _ой дру_, у_о__е_а,
___ дре__е_ь _од _у__а_ье_
_вое_о вере_е_а?
Вы_ье_, добрая _одру__а
Бед_ой ю_о___ _оей,
Вы_ье_ _ _оря; _де _е _ру__а?
_ерд_у буде_ ве_е_ей.
__ой __е _е__ю, _а_ _____а
___о за _оре_ ___а;
__ой __е _е__ю, _а_ дев__а
За водой _оу_ру __а.

Буквы с овалами

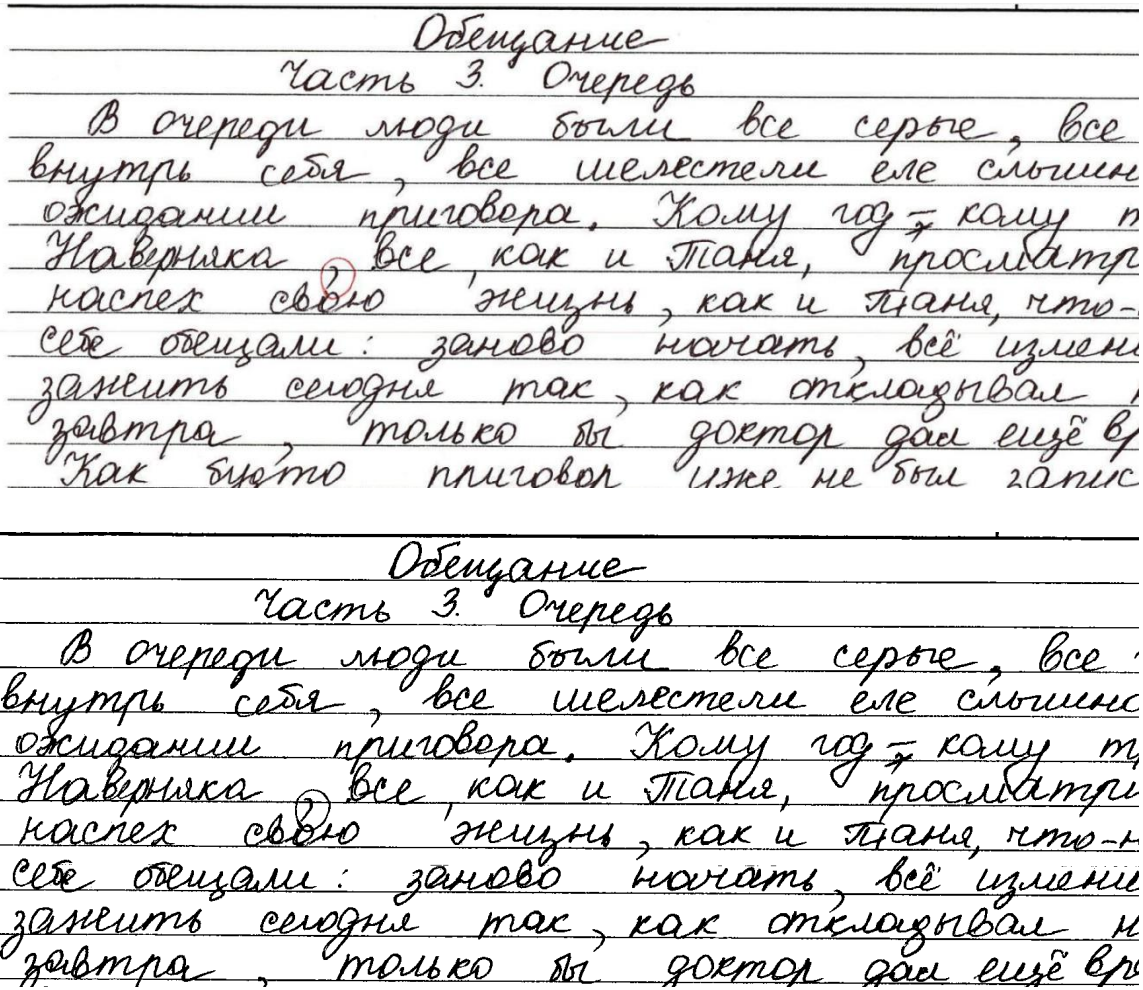
Буря ___ою небо кроет,
В__р_ _не_ные крутя;
То, как зверь, она завоет,
То зап_а_ет, как д_тя,
То по кров_е обвет_а_ой
Вдру_ _о_о_ой за_у__т,
То, как путн_к запозда_ый,
К на_ в око_ко за_ту__т.
На_а вет_ая _а_у_ка
_ пе_а_ьна _ те_на.
_то _е ты, _оя _тару_ка,
Пр_у_о_к_а у окна?
___ бур_ завыванье_
Ты, _ой дру_, уто__ена,
___ дре__е_ь под _у__анье_
_вое_о веретена?
Выпье_, добрая подру_ка
Бедной юно_т_ _оей,
Выпье_ _ _оря; _де _е кру_ка?
_ердцу будет ве_е_ей.
_пой _не пе_ню, как __н_ца
Т__о за _оре_ ___а;
_пой _не пе_ню, как дев_ца
За водой поутру __а.

Буквы с овалами и палочками

Буря ___ою небо кроет,
Вихри снежные крутя;
То, как зверь, она завоет,
То зап_а_ет, как дитя,
То по кров_е обветша_ой
Вдру_ со_о_ой зашу_ит,
То, как путник запозда_ый,
К на_ в окошко засту_ит.
Наша ветхая _а_ужка
И пе_а_ьна и те_на.
_то же ты, _оя старушка,
Приу_о_к_а у окна?
И_и бури завыванье_
Ты, _ой дру_, уто__ена,
И_и дре__ешь под жужжанье_
Свое_о веретена?
Выпье_, добрая подружка
Бедной юности _оей,
Выпье_ с _оря; _де же кружка?
Сердцу будет весе_ей.
Спой _не песню, как синица
Тихо за _оре_ жи_а;
Спой _не песню, как девица
За водой поутру ш_а.

Буквы с овалами, палочками и полуовалами

Исходный документ



Обещание

Часть 3. Очередь

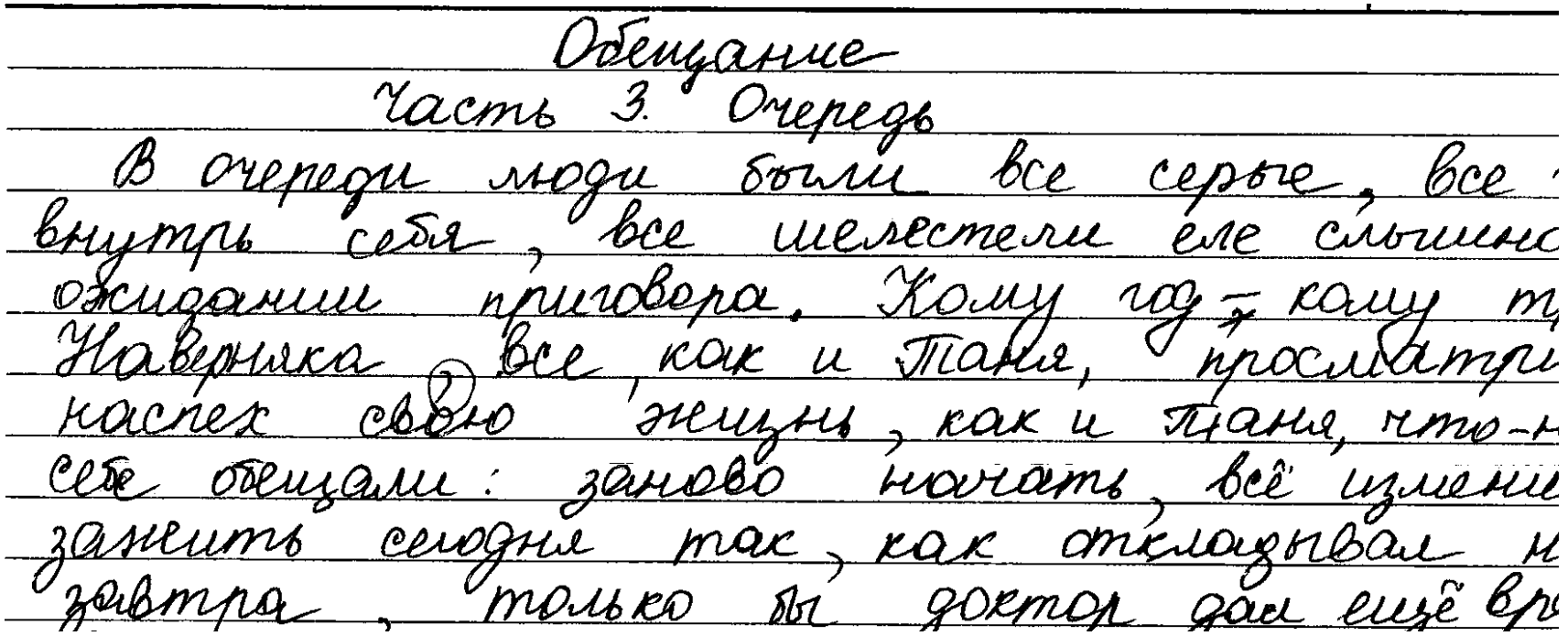
В очереди люди были все серые, все
внутрь себя, все шелестели еле слышн
ожидании приговора. Кому год – кому п
Наверняка все, как и Таня, прокрутр
каспех свою жизнь, как и Таня, что-
себе обещали: заново начать, всё измен
занести сегодня так, как откладывал
завтра, только бы доктор дал ещё вр
Как будто приговор уже не был запис

Обещание

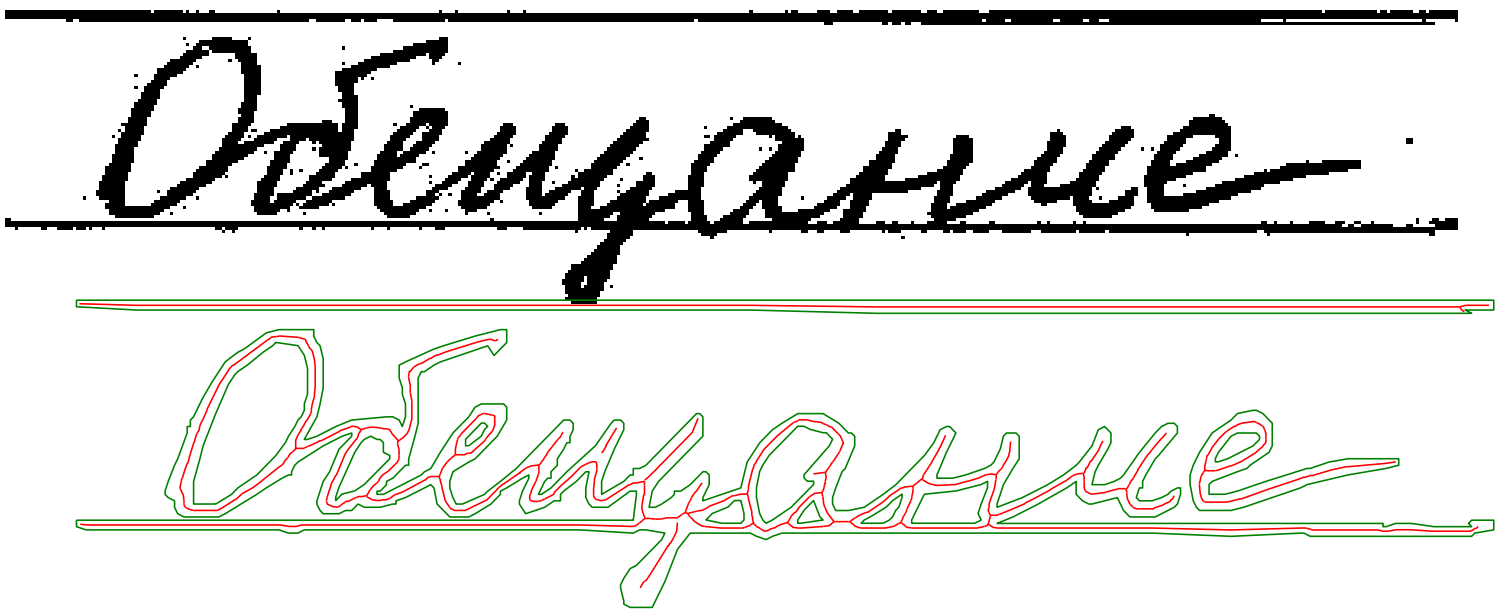
Часть 3. Очередь

В очереди люди были все серые, все
внутрь себя, все шелестели еле слышн
ожидании приговора. Кому год – кому т
Наверняка все, как и Таня, прокрутр
каспех свою жизнь, как и Таня, что-
себе обещали: заново начать, всё измен
занести сегодня так, как откладывал
завтра, только бы доктор дал ещё вр

Сегментация - бинаризация


Обещание
Часть 3. Очередь
В очереди люди были все серые, все
внутрь себя, все шелестели еле слышно
ожиданием приговора. Кому год — кому т
Наверняка все, как и Таня, прослезятся
каспех свою жизнь, как и Таня, что-н
себе обещали: заново начать, всё измени
заменить сегодня так, как откладывал н
завтра, только бы доктор дал ещё вр

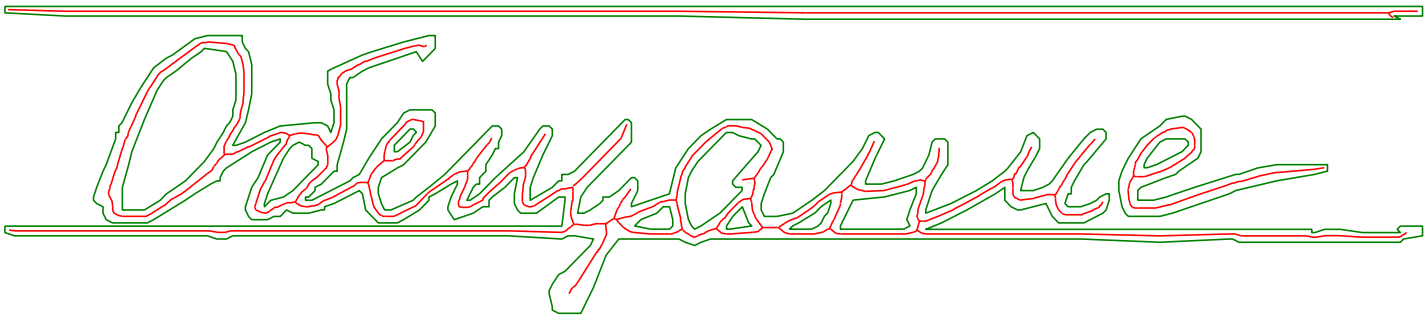
Гранично-скелетное представление



Граница – многоугольная фигура, аппроксимирующая растровое изображение

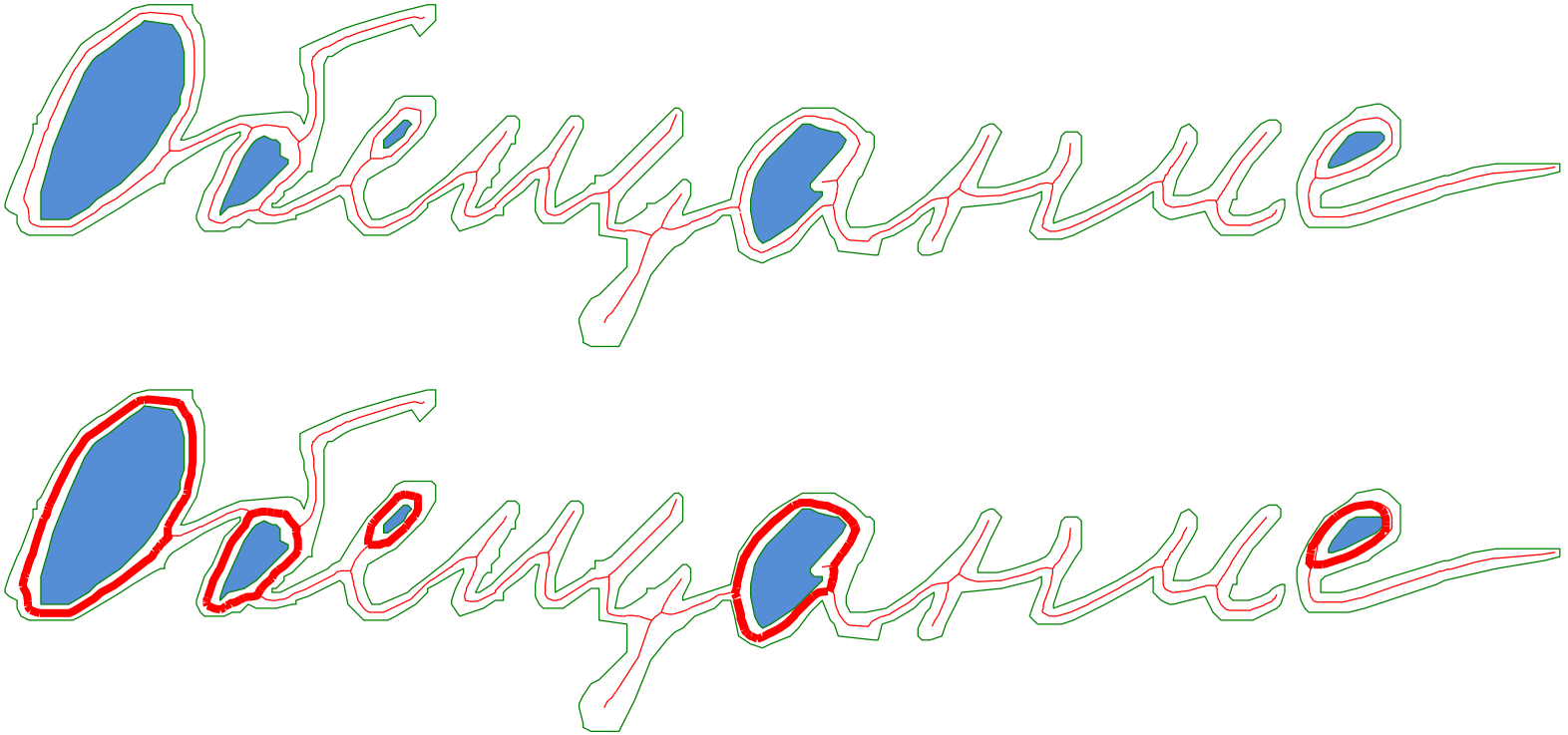
Скелет – срединная ось состоит из точек-центров вписанных в фигуру окружностей

Выделение текста



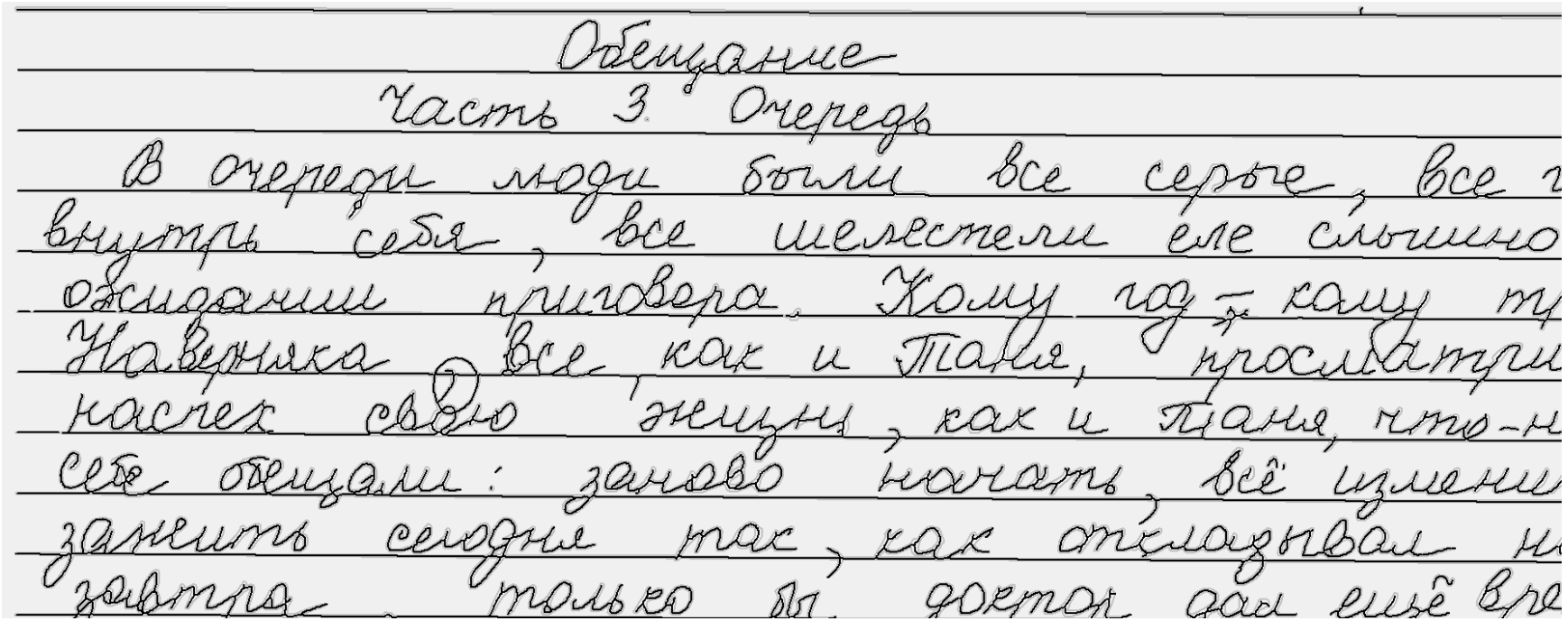
Прямолинейная разметка удаляется

Выделение овалов



Каждый овал определяется связной компонентой разбиения изображения границей многоугольной фигуры (дыра внутри фигуры)

Построение скелета

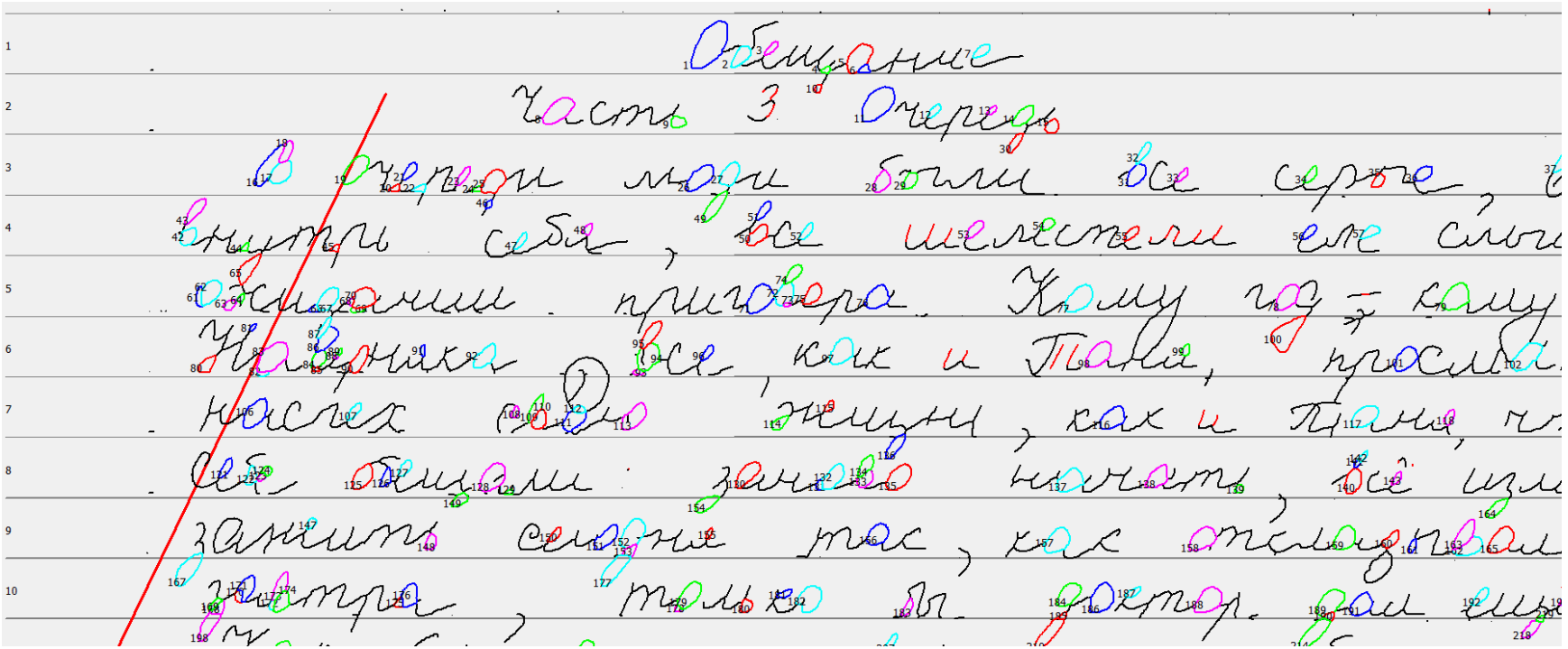


Обещание

Часть 3. Очередь

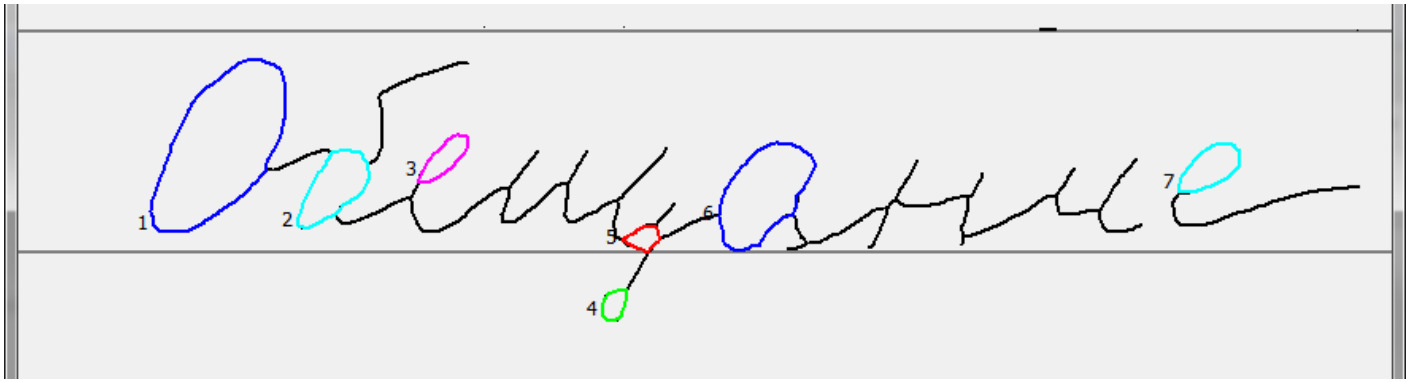
В очереди люди были все серые, все
внутрь себя, все шелестели еле слышно
ожидании приговора. Кому год — кому т
Наверняка все, как и Таня, проклинали
каждая свою жизнь, как и Таня, что-н
себе обещали: заново начать, всё измени
завтра только бы доктор дал ещё вре

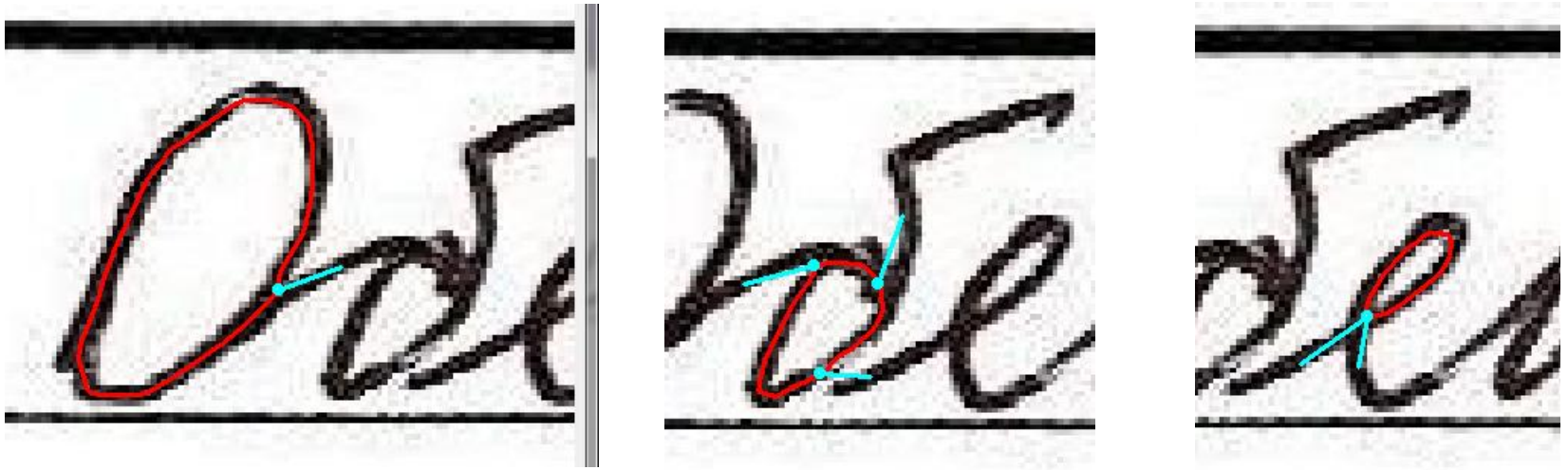
Выделение овалов



Скелетное представление позволяет вычислить наклон письма (красная линия)

Штрихи-ростки на овалах





Набор ростков является индивидуальным для каждой буквы и используется для классификации

Задача обучения

Распознавание букв

- Датасет – 35 диктантов на 2 страницах каждый
- 30 строк на странице по 30-35 букв в строке – это 1000 букв на странице
- 600 овалов на странице
- Требуемое распознавание – ранжированный список возможных вариантов для каждой буквы

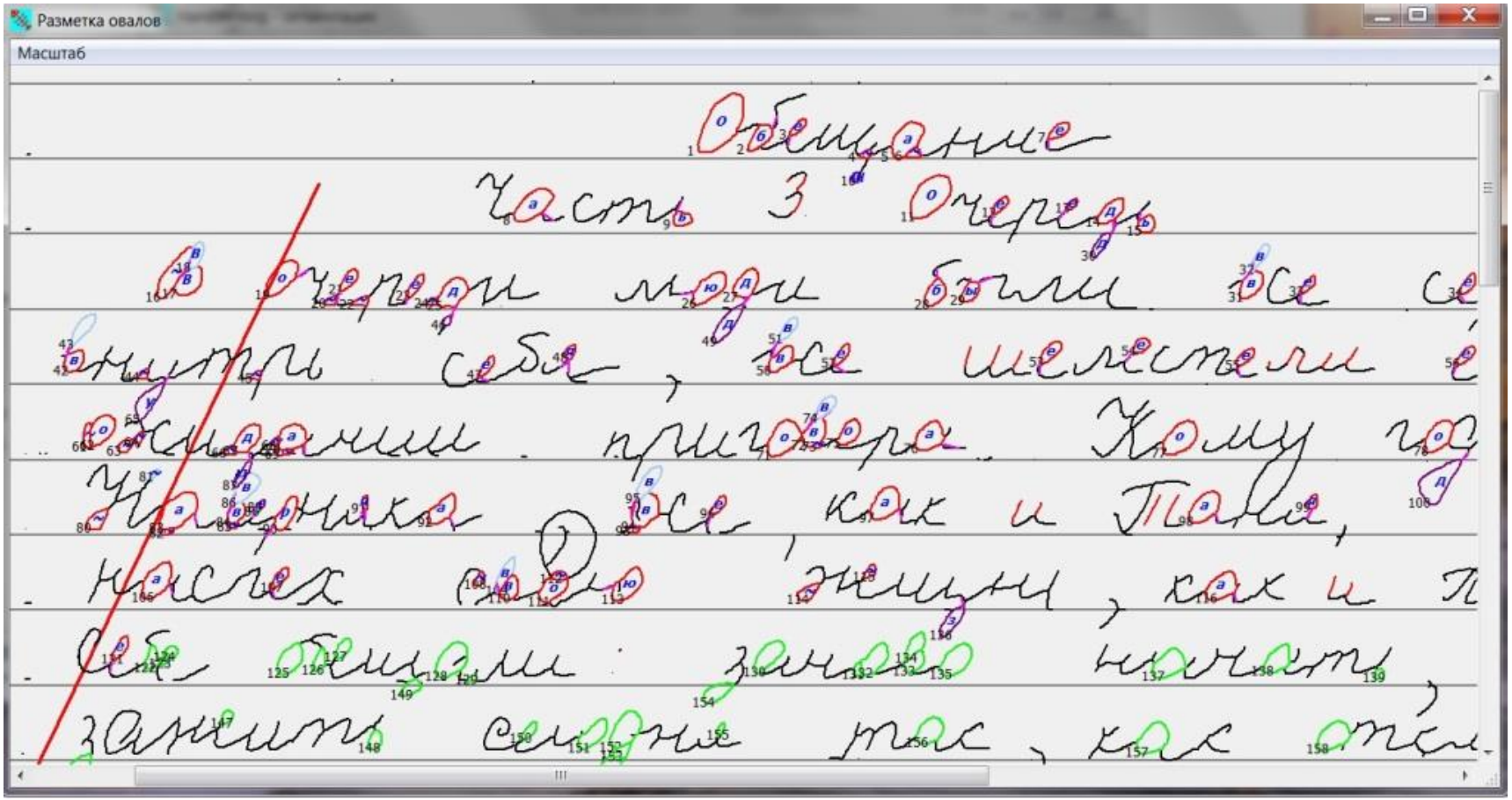
Распознавание слов

- Синтаксические правила, работа со словарём

Распознавание текста

- Семантический анализ , тематические модели

Разметка рукописных текстов



Разметка овалов и ростков

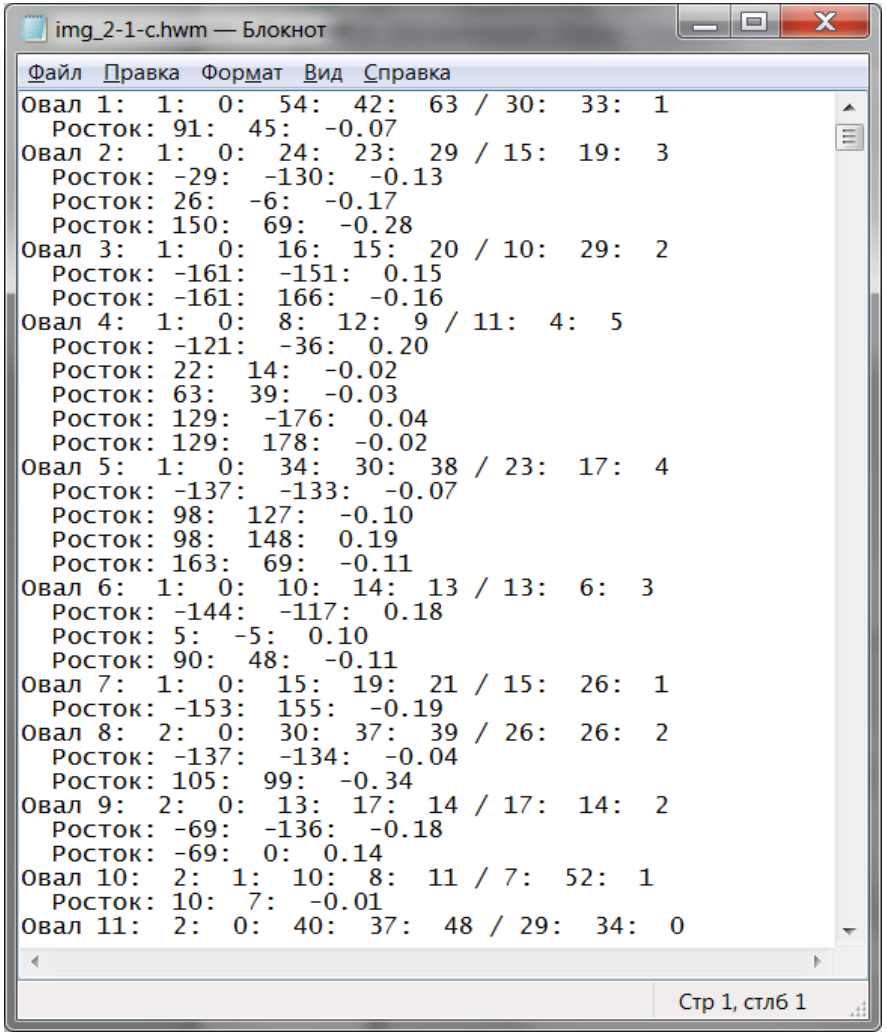
Овал:

- номер овала
- буква
- рамка (высота, ширина)
- координаты центра над базовой линией

Росток:

- азимут точки начала,
- направление
- кривизна,
- значимость

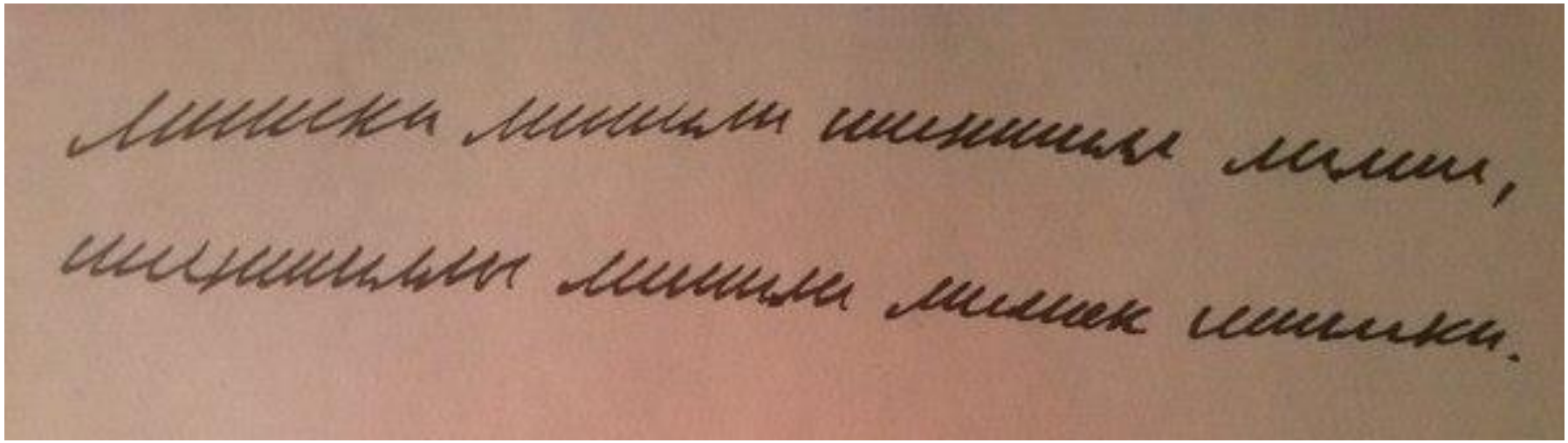
Признаковое описание овалов



Выводы

- Предлагается подход к распознаванию рукописного текста на основании выделения и распознавания каллиграфических элементов письма и букв
- Выделение элементов осуществляется на основе скелетного представления изображения текста и геометрического анализа срединных осей
- Разработан метод выделения и классификации овальных каллиграфических элементов письма
- Классификация осуществляется на основе машинного обучения по размеченным образцам текста
- Основой эффект от предлагаемого подхода ожидается за счет высокой информативности и малой размерности признакового описания геометрических характеристик базовых каллиграфических элементов письма

Спасибо за внимание!



Мишки лишили шиншилл лилии, шиншиллы лишили мишек шишки