Метод графемного описания и распознавания букв на основе медиального представления

Липкина Анна Львовна
Местецкий Леонид Моисеевич

19-я Всероссийская конференция с международным участием «Математические методы распознавания образов»

Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Введение

- **Графема** — *единица письменности; универсальное изображение символа языка.*
- Нет общепринятого формального определения графемы.
- **Свойства графемы:**
  - Любые две графемы хорошо отличимы между собой.
  - Пусть изображения I_1 и I_2 представляют одну и ту же графему. Тогда различие между I_1 и I_2 несущественное.

Рис.: Различные графемы строчной буквы «т»

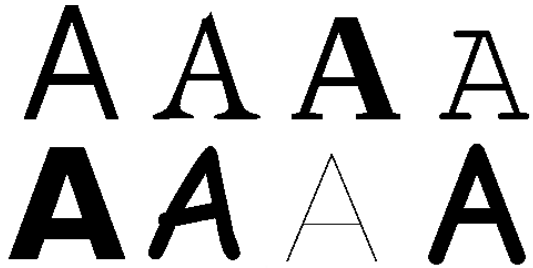


Рис.: Варианты написания букв, соответствующих графеме прописной буквы «А»

Постановка задачи

- **Данные:** цифровые изображения кириллических букв в различных шрифтах
- **Цель:**
  - Составление математической модели графемы (для любого шрифта); разработка алгоритма ее построения.
  - Проверка гипотезы о том, что графемы достаточно для распознавания символов в других шрифтах.
  - Решение задачи классификации изображений букв в кириллических шрифтах.

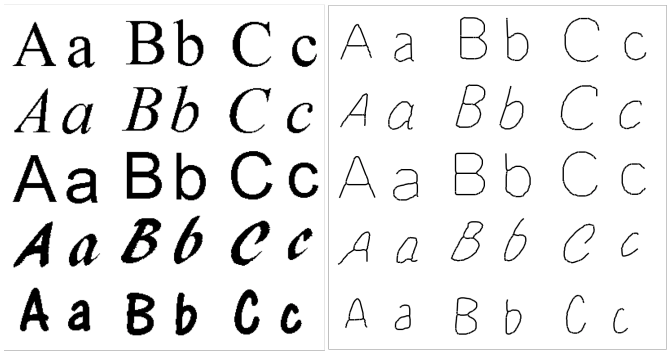


Рис.: Пример букв и их графем в различных шрифтах

Математическая модель графемы

- *Скелетное представление фигуры* — множество центров всех вписанных пустых кругов фигуры.
- *Операция агрегирования скелетного графа* — «склеивание» в одну цепь всех таких последовательных рёбер, инцидентные вершины которых имеют степень 2.

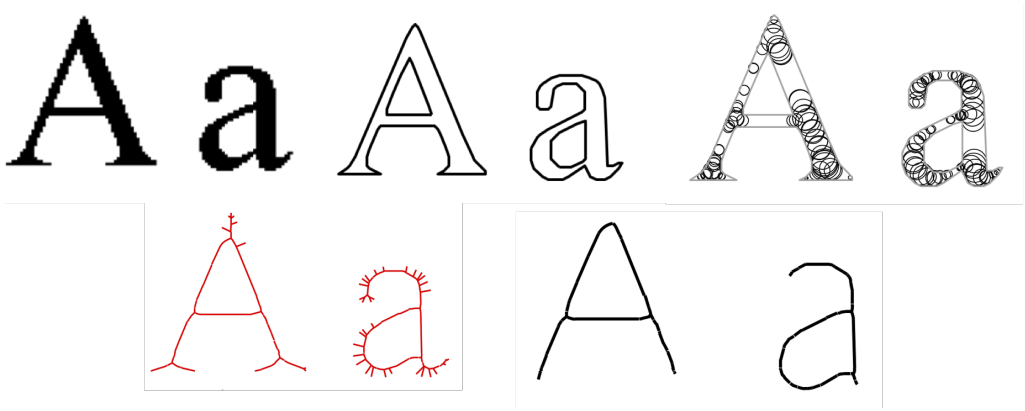


Рис.: Процесс построения графемы, слева направо: бинарное изображение символа; аппроксимирующий многоугольник; вписанные пустые круги фигуры; скелетный граф; метаграф — математическая модель графемы.

Математическая модель графемы

- *Основа модели* — скелетный граф бинарного изображения.
- *Обобщение модели* — отсечение всех элементов скелета, которые различаются в символах в разных шрифтах (например, засечки), и сохранение тех элементов, которые универсальны.

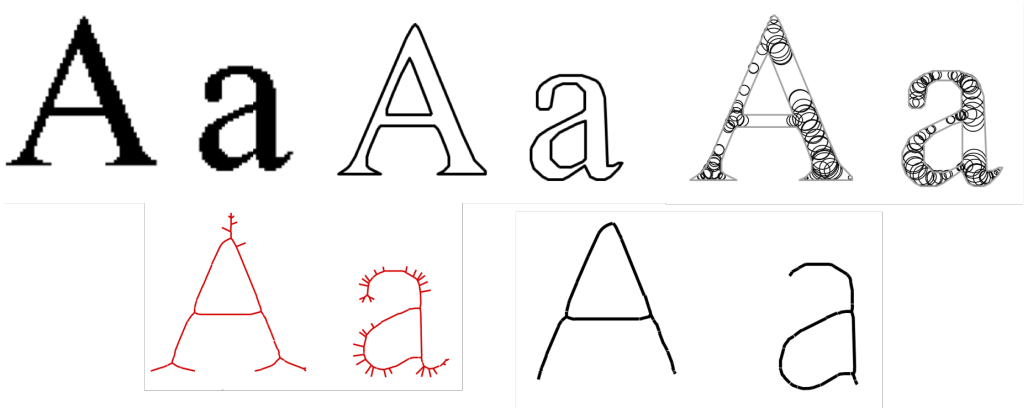
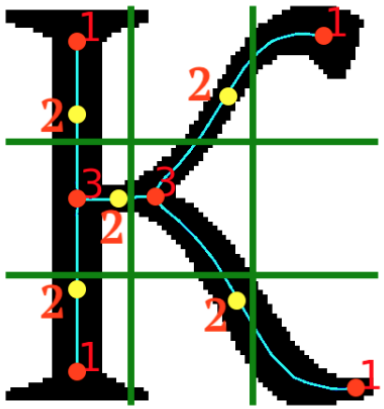


Рис.: Процесс построения графемы, слева направо: бинарное изображение символа; аппроксимирующий многоугольник; вписанные пустые круги фигуры; скелетный граф; метаграф — математическая модель графемы.

Генерация признаков

- Признаки выделяются на основе положения ключевых вершин и связности скелетного графа.



1	0	1
0	0	0
1	0	1

КОЛИЧЕСТВО ВЕРШИН СТЕПЕНИ 1

0	0	0
1	1	0
0	0	0

КОЛИЧЕСТВО ВЕРШИН СТЕПЕНИ 3

1	1	0
1	0	0
1	1	0

КОЛИЧЕСТВО ВЕРШИН СТЕПЕНИ 2

Генерация признаков

- Выделение вершин степени 1 и 3
- Выделение вершин степени 2 — середин всех существующих гиперребер.
- Фиксация набора масок разбиения, по которым будут считаться статистики.
- Подсчет количества каждого типа вершин в областях в каждой маске.
- Добавление числа компонент связности в скелетном графе буквы.

Обучение

- Пусть X^{tr} — обучающая выборка. Она строится генерацией 68 букв из 116 различных шрифтов (включают в себя различные стили), и, возможно, с последующим зашумлением.
- Пусть X^{te} — тестовая выборка, $|X^{te}| = n_{te}$, x^i_{te} — i-й объект тестовой выборки, y^{te}_i — его класс. В качестве метрики качества алгоритма a используется *точность классификации* (accuracy):

$$Q(a, X^{te}) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \mathbb{I}[y^{te}_i = a(x^{te}_i)].$$

- На этапе тестирования используется *расстояние Левенштейна* и *относительное расстояние Левенштейна.*
- Выбранный классификатор — *Случайный лес (Random Forest)*, использование *кросс-валидации.*

Эксперименты

- **Эксперимент 1**
  - Исследуется зависимость качества распознавания от размера и качества обучающей выборки.
  - Обучающие выборки состоит из 10, 20 или 30 шрифтов. Генерируются варианты с зашумлением и без. Размеры шрифта – 50pt.
  - Тестовые выборки — либо «дублеры» тренировочных в другом размере шрифта (30pt), либо в других шрифтах (50 других шрифтов). Также генерируются варианты с зашумлением или без.

Эксперименты

Train \ Test	duplicate orig	duplicate noisy	other orig	other noisy
fonts:10 orig	0.936	0.88	0.9	0.82
fonts:20 orig	**0.96**	0.913	0.925	0.855
fonts:30 orig	0.953	**0.9**	**0.93**	**0.859**
fonts:10 noisy	0.93	0.9	0.895	0.84
fonts:20 noisy	**0.966**	**0.92**	0.92	0.86
fonts:30 noisy	0.954	**0.92**	**0.94**	**0.87**

Результаты эксперимента, точность классификации. Размер шрифта на обучении — 50pt, на тесте — 30pt.

Эксперименты

- **Эксперимент 2**
  - Проводится анализ ошибок обученных выше классификаторов. Исследуется, для каких букв классификаторы чаще всего ошибаются.
  - Формат: `p(предсказанная буква|правильная буква)= процент случаев, в котором классификатор ошибся для данной правильной буквы.`

Эксперименты

Test / Train	duplicate orig	duplicate noisy	other orig	other noisy
fonts	p(э|з)=0.2	p(з|э)=0.28	p(а|д)=0.44	p(а|д)=0.5
10	p(л|п)=0.15	p(о|д)=0.2	p(з|э)=0.3	p(з|э)=0.49
orig	p(і|т)=0.15	p(а|д)=0.18	p(ч|ц)=0.27	p(ч|ц)=0.33
fonts	p(л|п)=0.1	p(з|э)=0.31	p(з|э)=0.29	p(з|э)=0.51
20	p(з|э)=0.1	p(о|д)=0.14	p(а|д)=0.17	p(а|д)=0.24
orig	p(о|д)=0.1	p(п|л)=0.13	p(п|л)=0.16	p(п|л)=0.21
fonts	p(з|э)=0.2	p(з|э)=0.35	p(з|э)=0.29	p(з|э)=0.51
30	p(л|п)=0.1	p(л|п)=0.14	p(а|д)=0.21	p(а|д)=0.29
orig	p(п|л)=0.1	p(а|д)=0.13	p(п|л)=0.14	p(п|л)=0.19
fonts	p(і|т)=0.45	p(з|э)=0.28	p(і|т)=0.44	p(з|э)=0.49
10	p(л|п)=0.25	p(л|п)=0.22	p(а|д)=0.36	p(а|д)=0.48
noisy	p(з|э)=0.2	p(о|д)=0.2	p(з|э)=0.3	p(ч|ц)=0.32
fonts	p(з|э)=0.25	p(з|э)=0.39	p(з|э)=0.36	p(з|э)=0.57
20	p(л|п)=0.125	p(л|п)=0.15	p(а|д)=0.19	p(а|д)=0.25
noisy	p(о|д)=0.1	p(п|л)=0.125	p(п|л)=0.16	p(п|л)=0.22
fonts	p(з|э)=0.26	p(з|э)=0.36	p(з|э)=0.34	p(з|э)=0.5
30	p(п|л)=0.12	p(л|п)=0.14	p(а|д)=0.17	p(а|д)=0.2
noisy	p(л|п)=0.1	p(о|д)=0.12	p(п|л)=0.14	p(п|л)=0.16

Результаты эксперимента, топ-3 ошибок.
Размер шрифта на обучении — 50pt, на тесте — 30pt.

Эксперименты

- **Эксперимент 3**
  - Исследуется зависимость качества распознавания от расширения обучающего множества при фиксированном выборе количества шрифтов на тесте.
  - Обученные модели — те же модели, которые были обучены в эксперименте 1 Также обучается модель на всем доступном корпусе из 116 шрифтов в зашумленном варианте.
  - Тестовые выборки — 2 датасета: orig и noisy, являющихся дубликатами тренировочного множества, составленного из 10 шрифтов размера 30pt в оригинальном и зашумленных вариантах.

Эксперименты

Train \ Test	orig	noisy
fonts:10 orig	0.936	0.88
fonts:20 orig	0.936	**0.89**
fonts:30 orig	**0.94**	**0.89**
fonts:10 noisy	0.93	**0.9**
fonts:20 noisy	**0.95**	**0.9**
fonts:30 noisy	0.946	**0.9**
fonts:116 noisy	**0.954**	**0.94**

Результаты эксперимента, точность классификации. Количество шрифтов в тестовой выборке — 10 размера 30pt.

Эксперименты

- **Эксперимент 4**
  - Эксперимент аналогичен эксперименту 2 в применении для моделей и тестовых выборок из эксперимента 3.

Эксперименты

Train \ Test	orig	noisy
fonts 10 orig	р(э|з)=0.2 р(л|п)=0.15 р(і|т)=0.15	р(з|э)=0.28 р(о|д)=0.2 р(а|д)=0.18
fonts 20 orig	р(з|э)=0.2 р(д|а)=0.15 р(л|п)=0.15	р(з|э)=0.35 р(о|д)=0.2 р(л|п)=0.15
fonts 30 orig	р(з|э)=0.3 р(д|а)=0.15 р(л|п)=0.15	р(з|э)=0.38 р(н|м)=0.18 р(о|д)=0.16
fonts 10 noisy	р(і|т)=0.45 р(л|п)=0.25 р(з|э)=0.2	р(з|э)=0.28 р(л|п)=0.22 р(о|д)=0.2
fonts 20 noisy	р(л|п)=0.25 р(з|э)=0.2 р(у|ё)=0.15	р(з|э)=0.36 р(л|п)=0.22 р(о|д)=0.16
fonts 30 noisy	р(з|э)=0.25 р(л|п)=0.25 р(і|т)=0.15	р(з|э)=0.36 р(л|п)=0.22 р(о|д)=0.17
fonts 116 noisy	р(д|а)=0.15 р(л|п)=0.15 р(і|т)=0.15	р(о|д)=0.16 р(з|э)=0.15 р(л|п)=0.15

Результаты эксперимента, топ-3 ошибок. Количество шрифтов в тестовой выборке — 10 размера 30pt.

Эксперименты

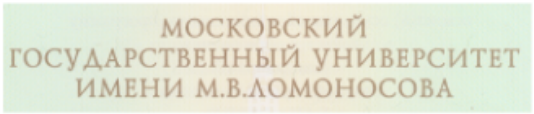
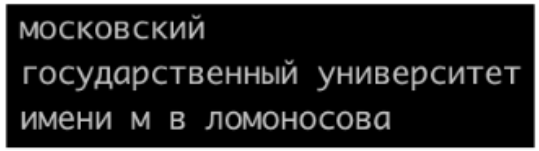
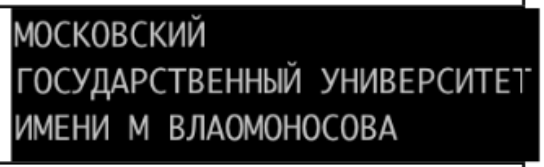
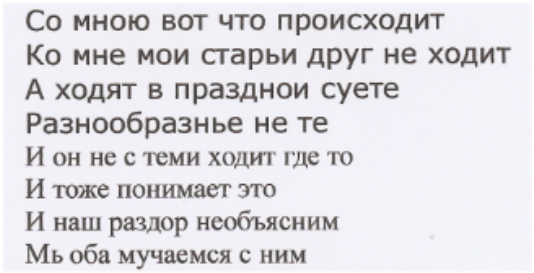
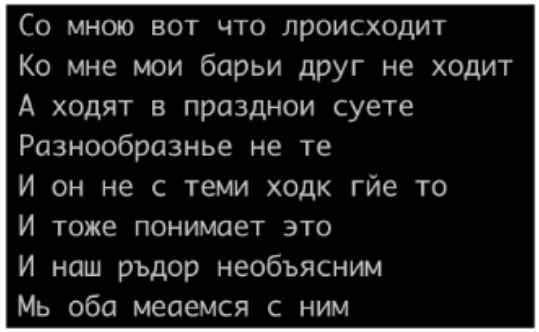
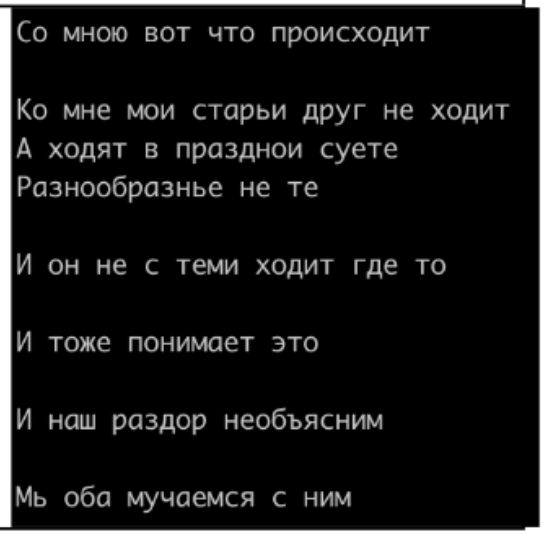
- **Эксперимент 5**
  - Исследуется качество предложенного метода в сравнении с базовым алгоритмом tesseract на реальных данных.
  - Для тренировочной выборки выбирается 116 шрифтов. Выборка генерируется в зашумленном варианте.
  - Тестовая выборка состоит из размеченных скриншотов, сканов и фотографий реальных текстов. Её объем — около 40 изображений.
  - Метрики качества — расстояние Левенштейна L и относительное расстояние Левенштейна L'.

Эксперименты

	L	L'
Train fonts:116 noisy	7.625	0.0753
Tesseract	2.125	0.015

Результаты эксперимента, средние расстояния Левенштейна и относительного Левенштейна

Эксперименты

Исходное изображение	Train fonts:116 noisy	Tesseract
		
		
		

Результаты распознавания текста двумя методами

Результаты экспериментов

- Больше шрифтов на обучении — качество распознавания лучше.
- Добавление шумов на обучении улучшает качество распознавания.
- Необязательно обучаться на всем множестве шрифтов — можно предсказывать буквы и новых шрифтах с хорошей точностью.
- Алгоритм путает буквы, схожие по скелетному строению.
- Алгоритм сильно зависит от качества бинаризации и сегментации — из-за этого увеличивается редакторское расстояние.

Заключение

Преимущества предлагаемого метода:

- Независимость от размера, типа шрифта и типа начертания буквы.
- Выделение общей структуры (математической модели графемы) букв, которой достаточно для распознавания букв в новых шрифтах.
- Отсутствие отказов от классификации.
- Интуитивно понятное построение признакового пространства, интерпретируемый алгоритм классификации.
- Отсутствие необходимости обучения на большой выборке — приемлемое качество достигается при малых размерах обучающего множества.

Заключение

Получены следующие результаты:

- Построение математической модели графемы для кириллических шрифтов на основе выделения подграфа непрерывного скелета изображения символа.
- Подтверждается гипотеза о возможности распознавания символов по графеме.
- Подход к генерации структурных признаков из предложенной модели.
- Метод классификации изображений букв в кириллических шрифтах и, как следствие, распознавание текста на изображении.