УДК 004.89

КЛАСТЕРИЗАЦИЯ СТАРОСЛАВЯНСКИХ РУКОПИСЕЙ ДЛЯ АНАЛИЗА СТИЛЕЙ ПИСЬМА И ОПРЕДЕЛЕНИЯ ПИСЦОВЫХ ШКОЛ

Бетеня Д.С. (ИТМО)
Научный руководитель – доцент, кандидат физико-математических наук, директор ВШЦК Михайлова Е.Г. (ИТМО)

Введение. Анализ почерка в исторических рукописях является важным инструментом для датировки документов, идентификации писцов и писцовых школ, изучения эволюции почерков и письменных практик. На данный момент существует немало исследований, посвящённых почеркам для европейских языков, например латинского или немецкого [1]. Однако для старославянского языка таких исследований практически нет. Уникальные особенности, такие как сложные надстрочные знаки и вариативность написания одной и той же буквы, усложняют обработку. Разработка методов анализа старославянских текстов, способных учитывать все эти нюансы, не только внесёт вклад в изучение истории, но и позволит использовать технологии для сохранения уникального культурного наследия.

Основная часть. Настоящая работа направлена на разработку методов кластеризации символов, позволяющих анализировать стили письма и соотносить их с определёнными писцовыми школами. Для этого были решены несколько задач, включая извлечение отдельных символов из изображений страниц рукописей, разметку датасета символов, построение информативного признакового пространства и кластеризацию изображений символов для выявления стилевых различий между писцовыми школами.

Извлечение символов из рукописного текста основано на анализе связанных областей пикселей [2]. Для анализа были выбраны два сборника рукописей XV века, происходящих из Великого княжества Литовского. Исходные изображения рукописей подвергались предварительной обработке, включая удаление шумов, выравнивание контраста и бинаризацию. После этого применялся алгоритм сегментации, выделяющий связные компоненты, соответствующие отдельным символам. Однако сложность старославянских текстов, таких как наличие надстрочных знаков и наложение букв, потребовала дополнительной коррекции сегментации. В случаях слияния символов использовались эвристические методы разрыва соединений, а для восстановления пропущенных деталей учитывался локальный контекст соседних символов.

После выделения символов формировалось их представление в виде вектора признаков, описанное в источнике [3]. Каждый символ центрировался и нормализовывался по размеру, приводясь к фиксированному разрешению 10×10 пикселей. Для кодирования формы и структуры букв использовались такие характеристики, как распределение интенсивности пикселей, соотношение сторон, доля фона относительно текста и локальные моменты изображения. Эти признаки позволяли учесть ключевые особенности почерка, такие как округлость, наклон и толщина линий.

Для анализа почерков использовались методы машинного обучения. В качестве базового алгоритма кластеризации применялся метод KMeans, дополнительно использовалось понижение размерности через PCA для визуализации распределения символов в признаковом пространстве. Полученные результаты показали высокую точность разбиения на группы, соответствующие различным писцовым школам, что подтверждает эффективность предложенного подхода.

Выводы. В ходе работы были выделены символы из двух сборников, создано компактное и информативное признаковое пространство, проведена кластеризация, выявившая различия между стилями письма с высокой точностью. Результаты подтверждают эффективность метода и открывают перспективы для автоматического анализа старославянских почерков. В дальнейшем предполагается улучшение алгоритмов

сегментации символов, расширение разметки на полный набор букв, тестирование более сложных алгоритмов кластеризации, таких как иерархическая кластеризация, а также использование дополнительных признаков, таких как динамика толщины линий и распределение кривизны, для более точного анализа почерков [4].

Список использованных источников:

1. D.S. Nascimento, F. Rayanne, S. Smith, M. Abreu. Exploring Medieval Manuscripts Writer Predictability: A Study on Scribe and Letter Identification // Digital Studies/Le champ numérique. – 2022. – № 12(1). – P. 1–41.

2. M. Dahllöf. Clustering Writing Components from Medieval Manuscripts // COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities. – 2018. – P. 11–13.

3. M. Dahllöf. Scribe Attribution for Early Medieval Handwriting by Means of Letter Extraction and Classification and a Voting Procedure for Larger Pieces // 22nd International Conference on Pattern Recognition, Stockholm, Sweden. – 2014. – P. 1910–1915.

4. H. Mohammed, M. Jampour. From Detection to Modelling: An End-to-End Paleographic System for Analysing Historical Handwriting Styles // Document Analysis Systems. DAS 2024. Lecture Notes in Computer Science. – 2024. – № 14994.