Научная статья
Статья в открытом доступе
УДК 331.101.1: 004.9
doi: 10.30987/2658-4026-2024-4-456-462

Методы анализа и визуализации этнокультурных данных при разработке дизайна обучающего ресурса

Игорь Юрьевич Коцюба[1]✉, Анна Мнацакановна Петросян[2]
[1,2.] Университет ИТМО; Санкт-Петербург, Россия
[1]ikotciuba@itmo.ru;
[2]anyapet6020@gmail.com

Аннотация.
Предложен цифровой подход в области анализа культурных данных в рамках образовательной деятельности. Автоматизация анализа данных позволяет дополнить исследования, выполненные учёными вручную, особенно в условиях больших выборок, а область его применения, то есть культурный анализ данных в обучающих целях способствует сохранению исторической памяти и упрощает образовательных процесс с точки зрения представления знаний. В качестве исследуемых данных выбраны цифровые копии армянских рукописей. Проблематику непосредственно распознавания можно определить, как культурную стратификацию на основе древних рукописей по таким направлениям как: выявление различных типов материальной и нематериальной культуры, типологизация культурной деятельности людей (например, выявление различных типов профессиональной, религиозной, языковой, региональной культуры или культуры, связанной с историческими этапами развития общества). Рассмотрено распознавание текста и оценка точности, поставлены гипотезы для анализа, позволяющие изучать аспекты исторических источников. Исходя из этих гипотез сделаны выводы по языковой и культурной сегментации и требования к эргономическому отображению результатов анализа культурных данных.

Ключевые слова: цифровизация образования, цифровое образование, условия успешности цифровизации, когнитивные компетенции, когнитивная компетентность, модель цифровизации образования

Для цитирования: Коцюба И.Ю., Петросян А.М. Методы анализа и визуализации этнокультурных данных при разработке дизайна обучающего ресурса // Эргодизайн. 2024. №4 (26). С. 456-462. http://dx.doi.org/10.30987/2658-4026-2024-4-456-462.

Original article
Open access article

Methods of Analysis and Visualization of Ethnocultural Data When Designing a Learning Resource

Igor Yu. Kotsyuba[1]✉, Anna M. Petrosyan[2]
[1.2.] ITMO University; Saint Petersburg, Russia
[1]ikotciuba@itmo.ru;
[2]anyapet6020@gmail.com

Abstract.
The paper proposes a digital approach to analysing cultural data in the context of educational activities. Automation of data analysis allows supplementing research performed manually by scientists, especially in conditions of large samples. The area of its application, that is, cultural data analysis for educational purposes, contributes to preserving historical memory and simplifies the learning process in terms of knowledge presentation. The authors select digital copies of Armenian manuscripts as the data under study. The problem of direct recognition can be defined as cultural stratification based on ancient manuscripts in such areas as identification of various kinds of material and intangible culture, typology of people's cultural activities (for instance, identification of various types of professional, religious, linguistic, regional culture or culture associated with historical stages of society development). The work considers text recognition and

accuracy assessment, formulates hypotheses for analysis that allow studying aspects of historical sources. Based on these hypotheses, the authors make conclusions on linguistic and cultural segmentation and requirements for the ergonomic display of the cultural data analysis.

Key words: education digitalization, digital education, conditions for successful digitalization, cognitive competencies, cognitive competence, model of education digitalization

For citation: Kotsyuba I.Yu., Petrosyan A.M. Methods of Analysis and Visualization of Ethnocultural Data When Designing a Learning Resource. Ergodizayn [Ergodesign]. 2024;4(26):456-462. Doi: 10.30987/2658-4026-2024-4-456-462.

Введение

Особое внимание в современном уделяется вопросам индивидуализации образования, учет особенностей восприятия обучающимся информации. Данный аспект особенно зависит от контента информационной системы, специфики изучаемой дисциплины, эргономики ее интерфейса. Цифровая гуманитаристика занимается исследованием социальных, исторических, лингвистических, философских, искусствоведческих вопросов с помощью математических и компьютерных наук. Одно из основных назначений этой междисциплинарной области - обеспечение сохранности культурного наследия. В задачи входит и трансформация данных в цифровой вариант, и сбор уже оцифрованных данных, но наиболее популярное направление на данный момент - анализ данных.

Таким образом, исследования цифровой гуманитаристики позволяют расширить использование технологий, которые обычно применяются на очевидно поддающихся математической интерпретации данных (например, экономических показателях или любых других легко категоризируемых данных). Культурное наследие является показателем духовного и материального достояния человечества. Как в генетике есть понятие о наследственности и изменчивости, так и культура сегодняшнего дня, несмотря на перманентные трансформации, базируется на культурном опыте предыдущих поколений.

Именно по причине своего влияния на формирование поколений культурное наследие является предметом регулирования во многих государствах. Соответственно, встаёт вопрос о его сохранении в различных формах - от охранительно-запретительных до созидательно-репродуктивных, в том числе на основе анализа данных как инструмента репродуктивной формы. В статье предложена реализация анализа и моделирования текстовых данных для создания обучающего сервиса с учетом эргономики представления культурных данных на образовательном ИТ-ресурсе.

1. Обзор подходов к анализу культурных данных

1.1. Связь со смежными исследованиями.

Анализ памятников культурного наследия в образовательных целях уже проводился разными мировыми институциями. Университет Индианы [1], чьи учёные придали песеннику Франческо Петрарки цифровой формат и добавили различные метаданные, благодаря которым студенты могут быстрее изучать творчество итальянского поэта. Исследователи руководствовались принципами простоты и удобства использования, ведь проект изначально реализован для студентов. Команда подчёркивает, что ею создан инновационный подход в цифровой работе с историческими документами, который четко проводит различие между историей и повествованием, между материальными ресурсами и ненадежными анекдотическими рассказами.

Факультет вычислительных наук и данных Парижского университета (Сорбонны) [2], где занимаются цифровым анализом иконографических корпусов, а именно разработкой инструментов анализа и новых инструментов компьютеризированной обработки и интеграцией этих инструментов в средах цифровых изданий и платформах, которые могут служить, например, историческим словарём, потому как реализован поиск по слову по оцифрованным рукописным материалам. Основная миссия проекта - сформулировать два взаимодополняющих исследовательских подхода к данным из корпусов текстов и изображений: локальный анализ и детальное описание этих корпусов (например, критических изданий) и использовать цифровые исследовательские инструменты, более ориентированные на количественный анализ, а также разработать методы обучения, применимые к более крупным наборам корпусов

1.2. Распознавание текстов и оценка точности.

Существуют сервисы по распознаванию рукописных текстов с собственной (не латинской) системой письменности [3]. Однако в данной работе предполагается применение библиотек, а не использование готовых сервисов. Наибольшее признание

[4], [5] имеют две библиотеки, а именно Tesseract OCR и Google Cloud Vision. Они регулярно упоминаются в прикладных исследованиях в рамках Международной конференция по анализу и распознаванию документов [6], [7].

На основании этих статей и документации инструментов выявлены особенности библиотек, которые прежде всего позволяет оценить применимость технологий в нашем случае, когда инструментов, поддерживающих армянский язык не так много, а именно: – языковая поддержка:

– Google Cloud Vision и Tesseract OCR поддерживают указанный язык [8], [9],

– методы распознавания: Google Cloud Vision применяет свёрточные нейронные сети для классификации объектов на изображениях. В Tesseract OCR используется Optical Character Recognition, которое возможно комбинировать с LSTM-моделями и скрытыми марковскими моделями.

Однако для работы с Google Cloud Vision требуется платное переменное окружение. К тому же Tesseract OCR показывается лучшие результаты [10], [11], поэтому для распознавания будет применена эта библиотека.

Предобработка является важным этапом, так как упрощает работу алгоритма и способствует его эффективности [12]. Шагами предобработки являются: изменение размера изображения с помощью функции “cv2.resize()”, конвертация цветового пространства в оттенки серого с помощью функции “cv2.cvtColor()”, уменьшение шума с помощью функции “cv2.fastNlMeansDenoising()”, улучшение контрастности с помощью адаптивной гистограммной эквализации с ограничением контраста с использованием функции “cv2.createCLAHE()” и метода “CLAHE.apply()” [13].

Учтена специфика исторических документов и удалены изображения с низкой степенью информативности. Таких страниц оказалось меньше 5% и их удаление не повлияло на репрезентативность выборки по географическому признаку. Для этого выведены размеры “потерь” и выбрано пороговое значение с помощью экспертного метода (просмотра страниц с потерями) выбрано значение в 60 пикселей, что составляет около 2 сантиметров.

Что касается сегментации, то она является частным случаем задач компьютерного зрения [14]. Основной функцией в сегментации строк служит “cv2.findContours()”, которая ищет области, ограниченные кривой, которая образует границу объекта, в нашем случае, строки.

В общей сложности распознано около 3000 страниц. В хранилище рукописей Матенадаран есть расшифрованные специалистами издания (назовём их эталонными), поэтому появилась возможность оценить точность распознавания. С помощью поиска пересечений слов эталонного и распознанного текста рассчитывается отношение количества совпадающих слов на общее количество слов в эталонном тексте и умножает результат на 100 для получения процентного значения. Такая тестовая выборка репрезентативна и составляет 1024 страницы (выбирались из рукописей, которые распознавались и имелись в коллекции Матенадарана), или чуть больше 30%.

Таким образом, точность при этом составляла около 84%, что считается хорошим результатом, так как с точки зрения развития письменности следует учитывать некоторые изменения в части начертания символов.

Поэтому было принято решение проверить, действительно ли это связано с различиями в наборе символов рукописи и инструмента распознавания, и рассчитать метрику Character Error Rate (CER), которая измеряет долю символов, которые были неправильно распознаны моделью, по сравнению с эталонным текстом. Это можно реализовать с помощью расстояния Левенштейна, которое сравнивает распознанную и эталонную строки, а далее делится на общее количество символов в эталонном тексте, чтобы получить нормализованное значение CER в процентах [15].

Действительно, CER приняло значение, близкое к 16%. Такие символы можно наблюдать на страницах рукописей, а лингвисты отмечают, что они были заменены на другие буквы. В следующей главе будет приведена статистика буквенных замен (заменённых символьных употреблений).

2. Результаты

2.1. Проверка языковых гипотез.

Что касается первой гипотезы (потери в точности распознавания связаны с заменой букв), то она заключается в проверке несоответствия символов в связи с их появлением. Данная гипотеза является

отсылкой к оценке качества распознавания. Здесь речь идёт о метрике CER (Character Error Rate), которая была рассчитана с помощью нахождения пересечения эталонного и распознанных текстов.

Армянский алфавит дошёл до наших дней почти в неизменном виде, однако было установлено, что некоторые символы в определённый рубеж времени стали употребляться реже. Таких найдено 2, и динамика их употребления ухудшается, а динамика употребления тех, на которые они были заменены, улучшается (рис. 1), что подтверждает и саму гипотезу, и вывод, сделанный при анализе причин результатов оценки точности.

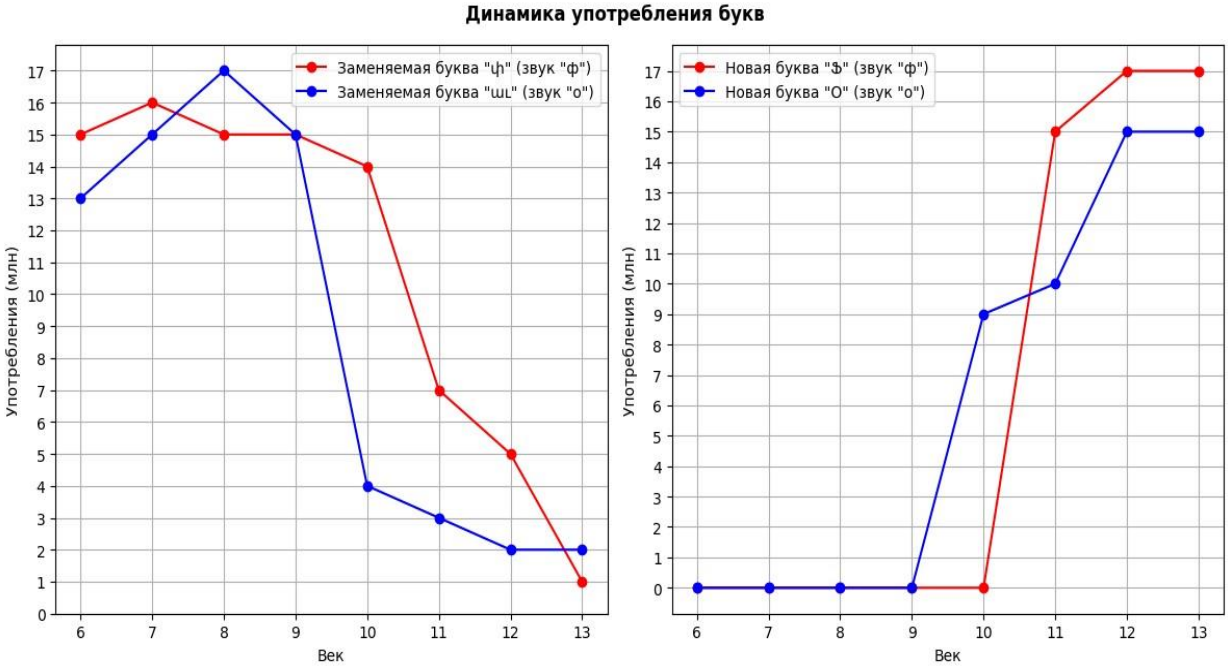

Рис. 1. Динамика использования букв
Fig. 1. Dynamics of letter use

Ещё одна гипотеза для символьного анализа - это проверка того, что можно выделить разные диалекты, которые отличаются написанием. Гипотеза подтверждается: в результате получились две группы и их главное отличие - это оглушение и озвончение согласных для восточного и западного диалектов, соответственно. Особенно это показательно и понятно на именах собственных (транслитерация приведена в таблице 1).

Таблица 1.

Примеры оглушения и озвончения в восточном и западном диалектах

Table 1.

Examples of deafening and voicing in Eastern and Western dialects

Восточный диалект	**Западный диалект**
*К*рик*к*ор	*Г*риг*г*ор
Сур*п*	Сур*б*
Ге*х*ард	Га*г*рад
Эр*с*рум	Эр*з*рум
Ар*т*вин	Ар*д*вин
Ашо*т*	Ашо*д*

Третья гипотеза подтверждается: определённые виды языковых конструкций появляются в контексте определённого вида деятельности, то есть социально стратифицированы (рис. 2): для неё был составлен корпус и проведена классификация с помощью метода опорных векторов.

2.2. Гео-тематическая сегментация.

Визуализация является необходимой в данном случае не только для удобного

представления материалов широкой публике, но и структурирования данных, потому что как в коллекции Бодлианской библиотеки, так и в электронном каталоге Матенадаран рукописи выложены просто сборниками.

Географическо-тематическая сегментация покрывает миссию Матенадарана по изучению центров книгописания, что необходимо для литературоведения и анализа самих рукописей и истории.

Поставленные гипотезы:

– в регионах создания, которые не относятся к современным границам Армении, доминирует религиозная тематика,

– виды занятий были сугубо сельскохозяйственного или промышленного характера в соответствии с этапом исторических обществ,

– на протяжении исследуемого периода была широко представлена фольклорная тематика.

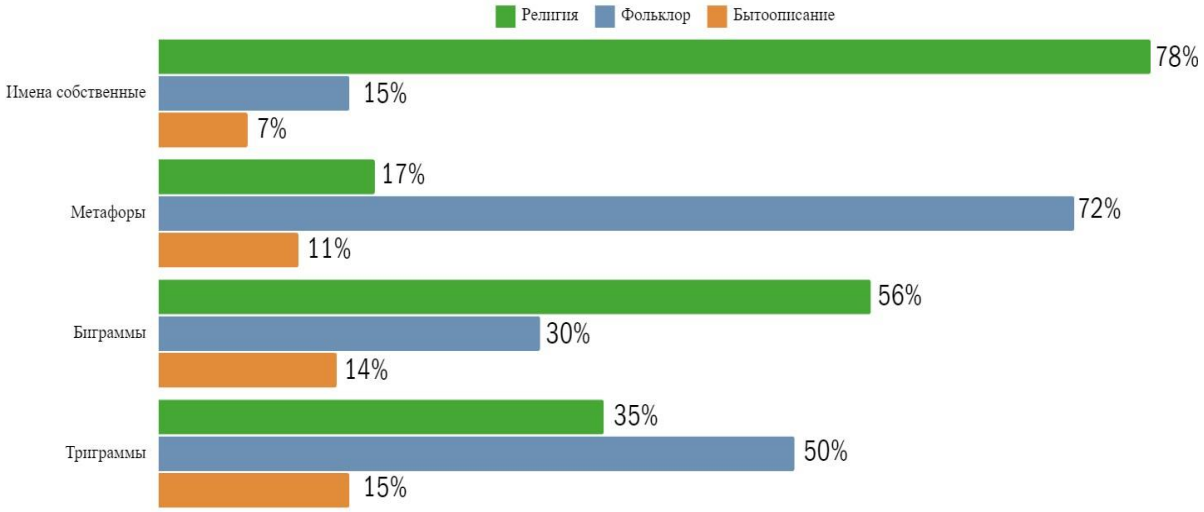

Рис. 2. Языковые конструкции и виды деятельности
Fig. 2. Language constructions and activities

Применив этот метод для имеющихся в данных регионов, можно моделировать карту с указанием наиболее типичного для определённого региона вида культурного наследия (рис. 3). Из поставленных гипотез не подтверждается только вторая.

Рис. 3. Визуализация для проверки гипотез по географическо-тематическому анализу
Fig. 3. Visualization for testing hypotheses for geographic-thematic analysis

Обсуждение/Заключение

Таким образом, проведенное исследование позволяет сделать такие выводы, как:

1. Положительной стороной цифрового подхода к анализу данных рукописей в образовательных целях является упрощение

представления информации в условиях больших выборок.

2. При этом следует разделять анализ на отдельные направления, разрабатывая гипотезы для каждого.

3. Разработанные визуализации также отвечают миссиям мировых организаций, выполняющих проекты в области цифровой гуманитаристики.

В результате, применив интеллектуальные технологии работы с данными, удалось определить историко-культурную природу исследуемых явлений, выявить основные виды культурного наследия по регионам и выявить некоторые языковые особенности, что особенно важно с учётом региона и его специфики, заключающейся в древности истории. Работа представляет интерес, так как развивает проекты по оцифровке рукописей, которые пока ещё не полностью отсканированы. К тому же такой подход может быть полезным и для уже готовых проектов, которые, в основном, представляют из себя скан страницы из рукописи и расшифрованный текст рядом. Также задачи позволяют сохранять память о культурном наследии, что благоприятно влияет на межкультурную коммуникацию, особенно посредством разработки электронных образовательных ресурсов с особыми возможностями визуальной эргономики.

СПИСОК ИСТОЧНИКОВ

1. **Kuhry E.** Medieval Glosses as a Test Subject for the Building of Tools for Digital Critical Editions. Journal of the Text Encoding Initiative. 2020;13. DOI 10.4000/jtei. 3544.

2. **Waters D.J.** The emerging digital infrastructure for research in the humanities. Международный журнал цифровых библиотек. International Journal on Digital Libraries. 2023;24:87-102. DOI 10.1007/s00799-022-00332-3.

3. **Бобров К.А., Шульман В.Д., Власов К.П.** Анализ технологий распознавания текста из изображения // Международный журнал гуманитарных и естественных наук. 2022. № 3-2(66). С. 124-128. DOI 10.24412/2500-1000-2022-3-2-124-128. EDN QPCEMR.

4. **Марков А.В.** Проведение сравнительного анализа attention ocr и tesseract в задаче распознавания символов на изображениях прейскурантов // Евразийский союз ученых. 2020. № 5-3(74). С. 65-67. EDN OOLAXF.

5. **Золотарев О.В., Юрчак В.А.** Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения // Инженерный вестник Дона. 2023. № 2(98). С. 156-164. EDN VOEMBZ.

6. **Kun J., Nussbaumer A., Pirker J., Conlan O., Memmel M., Steiner Ch.M.** Advancing Physics Learning Through Traversing a Multi-Modal Experimentation Space. 11th Int'l Conference of the Intelligent Environments, IOS Press. 2015, p. 373-380. DOI 10.3233/978-1-61499-530-2-373.

7. **Karatsaz D.** ICDAR 2015 competition on Robust Reading. 2015. 13th International Conference on Document Analysis and Recognition (ICDAR). 2015, p. 360-365. DOI 10.1109/icdar.2015.7333942.

8. **Gribomont I.** OCR with Google Vision API and Tesseract. Programming Historian. 2023;12(12). DOI 10.46430/phen0109.

9. **Omena J.J., Pilipets E., Gobbo B., Chao J.** The Potentials of Google Vision API-based Networks to Study Natively Digital Images. Diseña. 2021;19:1. DOI 10.7764/disena.19.

10. **Valentino J., Susetio Y.A.** Analisis Perbandingan Optical Character Recognition Google Vision dengan Microsoft Computer Vision pada Pembacaan KTP-el. *Jurnal JTIK.* 2023;7(4):552-561. DOI 10.35870/jtik.v7i4.1046.

11. **Kirana C.K., Ningrum J.D.K., Soraya D.U., Alqodri F., Mubarak A.A., Wibewanto W.**

REFERENCES

1. **Kuhry E.** Medieval Glosses as a Test Subject for the Building of Tools for Digital Critical Editions. Journal of the Text Encoding Initiative. 2020;13. DOI 10.4000/jtei. 3544.

2. **Waters D.J.** The Emerging Digital Infrastructure for Research in the Humanities. International Journal on Digital Libraries. 2023;24:87-102. DOI 10.1007/s00799-022-00332-3.

3. **Bobrov K.A., Shulman V.D., Vlasov K.P.** Analysis of Text Recognition Technologies from Image. International Journal of Humanities and Natural Sciences. 2022;3-2(66):124-128. DOI 10.24412/2500-1000-2022-3-2-124-128.

4. **Markov A.V.** Comparative Analysis of Attention OCR and Tesseract in the Problem of Character Recognition in Price List Images. Eurasian Union of Scientists. 2020;5-3(74):65-67.

5. **Zolotarev O.V., Yurchak V.A.** Tools for Solving Problems of Recognition and Clustering of Data From Documents Using Machine Learning Methods. Engineering Journal of Don. 2023;2(98):156-164.

6. **Kun J, Nussbaumer A, Pirker J, Conlan O, Memmel M, Steiner Ch.M.** Advancing Physics Learning Through Traversing a Multi-Modal Experimentation Space. In: Proceedings of the 11th Int'l Conference of the Intelligent Environments: IOS Press: 2015, p. 373-380. DOI 10.3233/978-1-61499-530-2-373.

7. **Karatsaz D.** ICDAR 2015 Competition on Robust Reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR): 2015, p. 360-365. DOI 10.1109/icdar.2015.7333942.

8. **Gribomont I.** OCR With Google Vision API and Tesseract. Programming Historian. 2023;12(12). DOI 10.46430/phen0109.

9. **Omena J.J., Pilipets E., Gobbo B., Chao J.** The Potentials of Google Vision API-based Networks to Study Natively Digital Images. Diseña. 2021;19:1. DOI 10.7764/disena.19.

10. **Valentino J., Susetio Y.A.** Analisis Perbandingan Optical Character Recognition Google Vision dengan Microsoft Computer Vision pada Pembacaan KTP-el. Jurnal JTIK. 2023;7(4):552-561. DOI 10.35870/jtik.v7i4.1046.

11. **Kirana C.K., Ningrum J.D.K, Soraya D.U., Alqodri F., Mubarak A.A., Wibewanto W.**

Synchronization of the Internship Registration System using a Study Plan and Acceptance Letter Based on Tesseract-OCR and Template Matching. 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE). 2023. DOI ICEEIE59078.2023.10334686.

12. **Murcia-Gomes D., Rojas-Valenzuela I., Valenzuela O.** Impact of Image Preprocessing Methods and Deep Learning Models for Classifying Histopathological Breast Cancer Images. Applied Sciences. 2022;12(22):11375. DOI 10.3390/app122211375.

13. **Gayatri G., Mitesh I., Chaudhari N., Chaware M.** Hand Gesture-based Virtual Mouse using Open CV. 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). 2023, p. 820-825. DOI 10.1109/IDCIoT56793.2023.10053488.

14. **Chen Q., Hao Ya., Jiuyn L.** FISMI-DRL: A Framework for Interactive Segmentation of Medical Image Based On Deep Reinforcement Learning. 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2023. DOI 10.1109/SMC53992.2023.10393973.

15. **Kumar R.P., Keya D.C., Kumar M.Ch.** Optical Character Recognition Systems for Accurate Interpretation of Handwritten Telugu Scripts. 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT). 2024. DOI 10.1109/IC2PCT60090.2024.10486323.

Synchronization of the Internship Registration System Using a Study Plan and Acceptance Letter Based on Tesseract-OCR and Template Matching. In: Proceedings of the 8th International Conference on Electrical, Electronics and Information Engineering (ICEEIE): 2023. DOI ICEEIE59078.2023.10334686.

12. **Murcia-Gomes D., Rojas-Valenzuela I., Valenzuela O.** Impact of Image Preprocessing Methods and Deep Learning Models for Classifying Histopathological Breast Cancer Images. Applied Sciences. 2022;12(22):11375. DOI 10.3390/app122211375.

13. **Gayatri G, Mitesh I, Chaudhari N, Chaware M.** Hand Gesture-based Virtual Mouse Using Open CV. In: Proceedings of 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT): 2023, p. 820-825. DOI 10.1109/IDCIoT56793.2023.10053488.

14. **Chen Q, Hao Ya, Jiuyn L.** FISMI-DRL: A Framework for Interactive Segmentation of Medical Image Based on Deep Reinforcement Learning. In: Proceedings of 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC): 2023. DOI 10.1109/SMC53992.2023.10393973.

15. **Kumar R.P., Keya D.C., Kumar M.Ch.** Optical Character Recognition Systems for Accurate Interpretation of Handwritten Telugu Scripts. In: Proceedings of 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT): 2024. DOI 10.1109/IC2PCT60090.2024.10486323.

Информация об авторах:

Коцюба Игорь Юрьевич (Санкт-Петербург, Россия) – кандидат технических наук, доцент, факультет инфокоммуникационных технологий, Университет ИТМО (Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, 49, литер А, e-mail: ikotciuba@itmo.ru) SPIN-код: 5296-3099, AuthorID: 757740.

Петросян Анна Мнацакановна (Санкт-Петербург, Россия) – студент бакалавриата, факультет инфокоммуникационных технологий, Университет ИТМО (Россия, 197101, г. Санкт-Петербург, Кронверкский проспект, 49, литер А, e-mail: anaypet6020@gmail.com).

Information about the authors:

Kotsyuba Igor Yuryevich (Saint Petersburg, Russia) – Candidate of Technical Sciences, Associate Professor of the Faculty of Telecommunication Systems and Technologies, ITMO University (bldg. A, 49 Kronverkskiy Prospekt, Saint Petersburg, 197101, Russia, e-mail: ikotciuba@itmo.ru) SPIN-code: 5296-3099, AuthorID: 757740.

Petrosyan Anna Mnatsakanovna (Saint Petersburg, Russia) – Bachelor's student of the Faculty of Telecommunication Systems and Technologies, ITMO University (bldg. A, 49 Kronverkskiy Prospekt, Saint Petersburg, 197101, Russia, e-mail: anaypet6020@gmail.com).

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.
Contribution of the authors: the authors contributed equally to this article.
Авторы заявляют об отсутствии конфликта интересов.
The authors declare no conflicts of interests.
Статья поступила в редакцию 09.08.2024; одобрена после рецензирования 15.08.2024; принята к публикации 22.08.2024. Рецензент –Кухта М.С.., доктор философских наук., профессор Томского политехнического университета, член редакционного совета журнала «Эргодизайн
The paper was submitted for publication on the 09th of August 2024; approved after the peer review on the 15th of August 2024; accepted for publication on the 22nd of August 2024. Reviewer – Kukhta M.S., Doctor of Sciences (Philosophy), Professor of Tomsk Polytechnic University, member of the editorial board of the journal "Ergodesign".