ОПТИЧНЕ РОЗПІЗНАВАВАННЯ ТЕКСТУ В ІСТОРИЧНИХ ДОКУМЕНТАХ: КЛЮЧОВІ ПРОБЛЕМИ ТА ІНОВАЦІНІ РІШЕННЯ

***Абросімов Є.О.**, аспірант, кафедра МСТ, ХНУРЕ*
***Зелений О.П.**, доцент, кафедра МСТ, ХНУРЕ*
***Дейнеко А.О.**, доцент, кафедра ШІ, ХНУРЕ*

Abstract. *Optical Character Recognition (OCR) of historical documents is critically important not only for preserving cultural heritage and providing access to archival materials, but also for the modern printing industry, which increasingly turns to digitized classical texts and the reprinting of rare editions. However, this task is complicated by a number of specific challenges: from physical degradation and non-standard fonts in old prints to the diversity of handwriting in manuscripts. This paper analyzes the key difficulties of OCR applied to historical materials, examines current innovative approaches (specialized neural networks, page layout segmentation, language models for post-processing), and outlines future research directions relevant to both archival science and publishing. It is shown that combining modern machine learning technologies with an understanding of the particular features of historical sources significantly improves recognition quality and opens new possibilities for reprinting and editorial work with classical texts.*

Основною сучасною проблемою архівів та бібліотек є низький рівень цифровізації. Багато документів зберігаються у вигляді фотографій, що ускладнює їх пошук і аналіз. Використання OCR-технологій (оптичного розпізнавання тексту) дозволяє перетворювати ці зображення на цифровий текст, придатний для пошуку за ключовими словами та інтелектуальної обробки. Проте, стандартні OCR-системи стикаються з труднощами під час роботи з історичними документами [1]. Серед таких проблем – фізична деградація документів, нестандартні шрифти (наприклад, готичний шрифт фрактура чи дореформений український правопис із літерою Ѣ), варіативність почерків і скорочень у рукописах, а також нестандартні макети сторінок. Таким чином, існує потреба в адаптації технологій OCR до специфіки історичних джерел.

Задача OCR для обробки історичних документів привертає увагу дослідників. Результати нещодавніх робіт свідчать про суттєвий прогрес у використані методів глибинного навчання. Al-Kendi та ін. [1, 2] відзначають, що попри успіхи в розпізнаванні сучасних рукописів, якість історичних даних залишається обмеженою через ряд проблем. Автори підкреслюють необхідність подальших досліджень, спеціально орієнтованих на історичні тексти.

Активно досліджуються підходи, що поєднують аналіз структури документа з OCR. Fleischhacker та ін. [3] запропонували систему, що спершу використовує методи машинного навчання для розпізнавання макету сторінки, а потім застосовує адаптовану OCR-модель. Інший підхід зосереджений на поліпшенні якості розпізнавання через пост-обробку тексту. Beshirov у [2] показали на прикладі історичних документів болгарською мовою, що застосування великої мовної моделі (LLM) та алгоритмів виправлення помилок після OCR систем може підвищити точність на 25% порівняно з результатами без постобробки [2].

Метою даної роботи є узагальнення проблем оптичного розпізнавання тексту в історичних документах, аналіз сучасних рішень для їх подолання, а також окреслення перспективних напрямів для подальших досліджень.

Однією з ключових проблем OCR для історичних документів є низька якість сканованих документів: вицвілий текст, плями, потертості, розриви паперу. Для подолання цієї проблеми застосовують попередню обробку зображень: фільтрацію шуму, підвищення контрастності, бінаризацію. Нові підходи використовують нейронні мережі для поліпшення якості – генеративні моделі можуть «очищувати» зображення або реконструювати вицвілий текст.

Не менш актуально проблемою є те що друковані видання минулих століть часто виконані шрифтами, відмінними від сучасних. Наприклад, у староукраїнських текстах вживалися літери ѣ («ять»), ѳ, ѵ, які відсутні у сучасній абетці. Один з можливих підходів – донавчання відкритих OCR-систем на історичних документах. Іншим варіантом є розробка моделей що здатні швидко адаптуватися до історичної орфографії. В умовах обмеженої кількості доступних даних застосовують синтетичне генерування текстів із старими шрифтами.

Розпізнавання рукописів історичного періоду є окремою задачею: почерк різниться від автора до автора, архаїчні скорочення, декоративні елементи. Сучасні системи розпізнавання рукописного тексту (handwritten text recognition, HTR) використовують нейронні мережі з архітектурами LSTM/GRU або трансформери, що здатні моделювати послідовності штрихів і літер. Наприклад, платформа Transkribus пропонує моделі, навчені на значних датасетах історичних рукописів. Для збільшення робастності системи також впроваджують автоматичну сегментацію рядків та слів і мовні моделі для перевірки осмисленості отриманого тексту. Комбінація візуального розпізнавання і лінгвістичного аналізу здатна суттєво знизити рівень помилок [1, 4].

Багато історичних джерел мають нестандартний формат: багатоколонні газети, таблиці, паралельні тексти кількома мовами, глоси на полях. Саме тому ще однією темою дослідження є аналізу макету історичних документів. Широку популярність мають детектори об'єктів, які виділяють заголовки, колонки, ілюстрації, основний текст тощо з використання нейронних мереж. Впровадження такого рішення дозволяє сфокусувати модель розпізнавання на правильні області. Більше того, з’являється можливість застосовувати різні специфічні моделі до різних зон: одну модель для основного тексту, а іншу для рукописних нотаток.

Найперспективнішим напрямком є розвиток методів штучного інтелекту, адаптованих до специфіки історичних даних. Глибинне навчання вже забезпечило прорив: нейронні мережі успішно навчаються розпізнавати складні шрифти та почерки. У майбутньому слід очікувати появи досконаліших архітектур, зокрема на основі трансформерів. Значущим напрямом є розширення корпусів навчальних даних. Відкриті ініціативи зі збору та анотації історичних документів забезпечать матеріал для навчання універсальніших OCR-систем – моделі зможуть імітувати старовинні шрифти та почерки.

Важливою інновацією стало використання великих мовних моделей для задач, пов’язаних з OCR. Мовні моделі типу GPT або BERT можна застосувати не лише для посткорекції, але й для контекстуального розпізнавання – коли алгоритм враховує зміст речення, щоб розрізнити схожі літери чи розплутати нечитабельний фрагмент. Подальші дослідження можуть призвести до появи систем, що розпізнають текст одночасно на рівні зображення і змісту.

Перспективним напрямом є мультимодальні та багатоетапні системи. Вони передбачають комбінування різних джерел інформації: контекст документу, порівняння кількох копій одного документа для взаємного усунення помилок. Актуальною лишається ідея людино-машинного співробітництва: інтерфейси, де попередньо розпізнаний текст перевіряється архівістами, можуть прискорити створення якісних цифрових колекцій.

Для України розвиток OCR для стародруків і архівів відкриває можливості широкого доступу до джерел, які донедавна були доступні лише вузькому колу дослідників. Інноваційні рішення вже застосовуються до розпізнавання староукраїнських рукописів чи документів доби УНР, і їхнє вдосконалення сприятиме збереженню національної пам'яті.

Оптичне розпізнавання тексту історичних документів є складною, але важливою технологією для збереження документів культурної спадщини. Основні труднощі пов'язані з низькою якістю старих матеріалів, нестандартними шрифтами та почерками, застарілою орфографією і складними макетами сторінок. Сучасні дослідження пропонують низку інновацій: від глибинних нейронних мереж до інтелектуальних методів постобробки із залученням мовних моделей. Наведені приклади демонструють, що спеціалізовані рішення дозволяють суттєво підвищити точність розпізнавання навіть на дуже складних історичних текстах. У майбутньому слід очікувати подальшого зближення технологій комп'ютерного зору та обробки мови для створення більш «розумних» OCR-систем, а також розширення відкритих датасетів і інструментів. Вирішення поставлених завдань сприятиме не лише автоматизації роботи з архівами, але й появі нових дослідницьких можливостей – від повнотекстового пошуку до статистичного аналізу еволюції мови. Розвиток OCR для історичних документів є інвестицією в збереження минулого та його інтеграцію в цифрове майбутнє.

Список літератури

1. AlKendi, W., Gechter, F., Heyberger, L., & Guyeux, C. (2024). Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey. Journal of Imaging, 10(1), 18. https://doi.org/10.3390/jimaging10010018.

2. Beshirov, A., Dobreva, M., Dimitrov, D., Hardalov, M., Koychev, I., & Nakov, P. (2025). Post-ocr text correction for Bulgarian historical documents. International Journal on Digital Libraries, 26(1). https://doi.org/10.1007/s00799-025-00415-x.

3. Fleischhacker, D., Kern, R., & Göderle, W. (2025). Enhancing OCR in historical documents with complex layouts through machine learning. International Journal on Digital Libraries, 26(1). https://doi.org/10.1007/s00799-025-00413-z.

4. Mousavi, S.M.H., & Lyashenko, V. (2017). Extracting old persian cuneiform font out of noisy images (handwritten or inscription). Machine Vision and Image Processing (MVIP). (p. 241-246). IEEE.