

ОПТИЧНЕ РОЗПІЗНАВАННЯ ТЕКСТУ (OCR) ТА ПАРСИНГ ДОКУМЕНТІВ ЗА ДОПОМОГОЮ AI-МОДЕЛЕЙ

Вінницький національний технічний університет

Анотація. У статті розглянуто сучасні методи оптичного розпізнавання тексту (OCR) та парсингу документів із використанням AI-моделей. Описано історичний розвиток OCR-технологій, їхню інтеграцію в мобільні додатки, а також основні принципи роботи та алгоритми, що забезпечують ефективність розпізнавання. Особливу увагу приділено викликам та обмеженням, включаючи труднощі в обробці рукописного тексту, якість вхідних зображень і багатомовну підтримку. Також розглянуто етичні та правові аспекти використання OCR у сферах, що вимагають збереження конфіденційності, зокрема в медицині. Окреслено перспективи розвитку технології, такі як покращення точності, інтеграція з електронними системами документообігу та розвиток алгоритмів для розпізнавання рукописного тексту.

Ключові слова: OCR, штучний інтелект, парсинг документів, розпізнавання тексту, мобільні додатки, медичні дані, глибоке навчання.

Abstract. This article examines modern methods of Optical Character Recognition (OCR) and document parsing using AI models. It discusses the historical development of OCR technologies, their integration into mobile applications, and the core principles and algorithms that enhance recognition accuracy. Special attention is given to challenges and limitations, including difficulties in processing handwritten text, input image quality, and multilingual support. Ethical and legal considerations, particularly in data-sensitive fields like healthcare, are also analyzed. The study outlines future technological advancements, such as improved accuracy, integration with electronic document management systems, and the evolution of algorithms for handwritten text recognition.

Keywords: OCR, artificial intelligence, document parsing, text recognition, mobile applications, medical data, deep learning.

Вступ

Зростання обсягу цифрової інформації, що спостерігається в останні десятиліття, зумовило необхідність створення ефективних методів перетворення фізичних документів у цифровий формат. Одним із ключових технологічних рішень у цьому контексті є оптичне розпізнавання тексту (OCR [1]), яке дозволяє автоматизувати процес перетворення зображень, що містять текстову інформацію, у придатні для подальшої обробки дані. Проте класичні OCR-алгоритми часто мають обмеження у розпізнаванні рукописного тексту, нестандартних шрифтів та структурованих документів, що зумовило необхідність застосування сучасних методів штучного інтелекту (AI).

Використання глибоких нейронних мереж та алгоритмів природної обробки мови (NLP) дозволило суттєво покращити точність розпізнавання та аналізу змісту текстових даних. У сфері медицини, де зберігання та обробка персональних медичних записів є критично важливими завданнями, застосування цих технологій дозволяє підвищити рівень автоматизації та зменшити ризики, пов'язані з людським фактором.

Історія розвитку технології OCR

Перші спроби автоматичного розпізнавання тексту відносяться до середини XX століття, коли були розроблені спеціалізовані пристрої для зчитування друкованих символів. Найбільшого поширення ці технології набули в банківській сфері та для автоматизації обробки поштової кореспонденції. Однак перші системи OCR працювали на основі шаблонного розпізнавання символів, що обмежувало їхню гнучкість у випадках використання різних шрифтів або складних документів.

З розвитком комп'ютерних технологій та появою методів машинного навчання точність розпізнавання тексту значно зросла. Використання глибоких нейронних мереж дозволило OCR-системам навчатися на різноманітних текстах, що дало змогу ефективно обробляти багатомовні документи, рукописний текст та документи з неідеальними умовами друку. На сучасному етапі OCR-технології поєднуються з алгоритмами NLP [2], що забезпечує не лише точне розпізнавання символів, а й аналіз змісту та витяг структурованих даних.

Принцип роботи OCR та AI-моделей

Процес оптичного розпізнавання тексту складається з кількох взаємопов'язаних етапів. Початковим є попередня обробка зображення, яка передбачає корекцію геометричних спотворень, зменшення рівня шуму та підвищення контрастності текстових елементів. На цьому етапі можуть застосовуватися методи бінаризації та нормалізації зображень, що сприяє покращенню подальшого розпізнавання. Наступним кроком є сегментація тексту, що включає виявлення окремих символів, слів та рядків. Використання сучасних алгоритмів машинного навчання дозволяє ефективно здійснювати сегментацію навіть у випадках, коли текст має нерівномірні проміжки між словами або містить графічні елементи.

Розпізнавання символів здійснюється на основі глибоких нейронних мереж, які попередньо навчаються на великих обсягах текстових даних. Однією з найбільш поширених моделей є Tesseract [3], яка використовує рекурентні нейронні мережі для визначення символів та контекстного аналізу слів. Постобробка результатів включає корекцію можливих помилок та перевірку відповідності розпізнаного тексту до лексичних правил конкретної мови, що значно підвищує точність остаточного результату.

Перспективи розвитку технологій OCR та AI

Наукові дослідження у сфері OCR та AI свідчать про постійне вдосконалення методів розпізнавання та обробки тексту. Одним із ключових напрямів є покращення адаптації OCR-систем до неструктурованих та багатомовних документів, що дозволяє розширити сферу їх застосування. Значна увага приділяється інтеграції цих технологій з електронними медичними системами, що сприяє автоматизації обробки клінічної документації та підвищенню ефективності медичних послуг. Також перспективним напрямом є розпізнавання рукописного тексту, що досі залишається складним завданням для автоматичних систем. Використання гібридних AI-моделей, які поєднують технології комп'ютерного зору та обробки природної мови, дозволяє значно покращити якість розпізнавання складних текстових структур.

Подальший розвиток технологій OCR спрямований на підвищення швидкості обробки великих обсягів документів, що є критично важливим у сферах з високими вимогами до точності та оперативності. Окрім того, спостерігається зростання інтересу до створення автономних OCR-систем, здатних працювати без необхідності постійного підключення до інтернету, що особливо актуально для медичних установ та архівів.

Області застосування OCR та AI-моделей

Сучасні технології OCR знайшли застосування в різних сферах. У фінансовому секторі вони використовуються для автоматизації обробки банківських чеків, розпізнавання кредитних документів та реєстрації клієнтів. У сфері охорони здоров'я OCR допомагає у веденні електронних медичних карт, автоматичному внесенні діагнозів і результатів аналізів у бази даних. Юридична галузь активно впроваджує OCR для обробки договорів, судових рішень та архівних матеріалів. Крім того, технології OCR відіграють важливу роль у логістиці, де автоматизація введення накладних та транспортних документів значно прискорює процеси обліку та перевезень.

Виклики та обмеження OCR-систем

Попри значні досягнення, OCR-системи стикаються з низкою викликів. Розпізнавання рукописного тексту залишається складним завданням через варіативність почерку та стилів написання. Крім того, якість розпізнавання залежить від вхідних даних: низька якість зображень, нерівномірне освітлення або викривлення документа можуть значно ускладнити обробку. Іншим викликом є багатомовна підтримка та здатність OCR-систем адаптуватися до нових шрифтів і спеціалізованих символів.

Етичні та правові аспекти

Використання OCR у сфері медицини, фінансів та юридичних послуг ставить важливі питання щодо конфіденційності та безпеки даних. Автоматизоване зчитування персональної інформації потребує захисту від витоків та зловживань. Важливим аспектом є дотримання законодавчих норм, зокрема Загального регламенту захисту даних (GDPR [4]) та аналогічних законів у різних країнах, що регламентують використання персональної інформації. OCR-системи, що працюють із медичними

записами, повинні відповідати вимогам стандартів, таких як HIPAA (у США), що гарантує безпеку обробки даних пацієнтів.

Застосування OCR у мобільних додатках

Інтеграція OCR у мобільні додатки є важливим напрямком, що дозволяє користувачам швидко оцифровувати документи без необхідності використання комп'ютера. Основні виклики при впровадженні цієї технології включають оптимізацію швидкості роботи, точності розпізнавання та забезпечення коректної обробки різних форматів тексту. Такі додатки часто поєднують OCR з іншими технологіями, як-от машинне навчання для автоматичного визначення категорій документів або NLP [5] для аналізу змісту тексту.

Висновки

Застосування технологій OCR та AI-моделей для автоматичного розпізнавання та аналізу текстових документів відкриває широкі можливості для цифровізації та автоматизації різних галузей. У сфері медицини це сприяє підвищенню ефективності управління медичною документацією, покращенню якості аналізу текстових даних та спрощенню доступу до інформації. Використання глибоких нейронних мереж та алгоритмів природної обробки мови дозволяє суттєво підвищити точність розпізнавання, навіть у випадках складних та неструктурованих документів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. What is OCR [Електронний ресурс] – Режим доступу до ресурсу: <https://aws.amazon.com/what-is/ocr/>
2. From OCR to AI: The evolution of OCR technology [Електронний ресурс] – Режим доступу: <https://www.affinda.com/blog/from-ocr-to-ai-the-evolution-of-ocr-technology>
3. Tesseract Open Source OCR Engine [Електронний ресурс] – Режим доступу до ресурсу: <https://github.com/tesseract-ocr/tesseract>
4. How Optical Character Recognition (OCR) Technology Optimizes Operations in the Healthcare Industry [Електронний ресурс] – Режим доступу: <https://www.docsumo.com/blogs/ocr/healthcare/>
5. Natural Language Processing [Електронний ресурс] – Режим доступу: <https://shorturl.at/ahhKg>

Козійчук Андрій Олександрович – студент групи ЗПІ-21б, ФІТКІ, Вінницький національний технічний університет, Вінниця, e-mail: alexvk1940@gmail.com

Майданиук Володимир Павлович – д.т.н., доцент, доцент кафедри програмного забезпечення, Вінницький національний технічний університет, м. Вінниця, e-mail: maidaniuk@vntu.edu.ua.

Koziichuk Andrii Olexandrovich – student of group ЗПІ-21b, Vinnytsia National Technical University, Vinnytsia, e-mail: alexvk1940@gmail.com

Maidaniuk Volodymyr - Doctor of Technical Sciences, Docent, Docent of the Software Department, Vinnytsia National Technical University, Vinnytsia, e-mail: maidaniuk@vntu.edu.ua.