Scientific Collection «InterConf», (44): with the Proceedings of the 8th International Scientific and Practical Conference «Scientific Research in XXI Century» (March 6-8, 2021). Ottawa, Canada: Methuen Publishing House, 2021. P. 741-749.

АНАЛІЗ ПІДХОДІВ ДО РОЗПІЗНАВАННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ У ТЕХНОЛОГІЇ OCR

Переяславська Світлана Олександрівна,
кандидат педагогічних наук, доцент, доцент кафедри інформаційних технологій та систем Луганського національного університету імені Тараса Шевченка, pereyaslav9@gmail.com
Шевченко Віталій Миколайович
здобувач другого (магістерського) рівня вищої освіти Луганського національного університету імені Тараса Шевченка, mynamevetal@gmail.com
Смагіна Ольга Олександрівна
кандидат педагогічних наук, доцент кафедри інформаційних технологій та систем Луганського національного університету імені Тараса Шевченка, smagina1804@gmail.com

Анотація. Стаття присвячена аналізу та дослідженню сучасних підходів до обробки зображення та розпізнавання текстової інформації у технології

OCR (Optical Character Recognition). У ході дослідження виявлено та проаналізовано найбільш популярні методи класифікації (шаблонний, структурний, ознаковий, статистичний, із застосуванням штучних нейронних мереж) та зроблено висновки стосовно ефективних напрямів їх застосування.

Ключові слова: алгоритм розпізнавання, оптичне розпізнавання символів, метод розпізнавання, OCR

Сьогодні існує великий попит на зберігання у цифровому вигляді текстової інформації, яка розміщена у друкованих, графічних або рукописних документах, з метою подальшої її обробки, редагування та аналізу. З іншого боку, широке поширення засобів оцифровування і сканування призвело до активного розвитку методів детектування і розпізнавання об’єктів на зображеннях. Це обумовлює розвиток систем оптичного розпізнавання символів (OCR), які дозволяють автоматично аналізувати надруковані або рукописні документи і готувати текстові дані в редагованих форматах для їх обробки.

Популярність таких OCR-систем викликає потребу у подальшому дослідженні технологій оптичного розпізнавання символів, в тому числі й аналізу, узагальненню сучасних підходів та алгоритмів обробки зображення та розпізнавання текстової інформації.

Технології оптичного розпізнавання символів приділяється багато уваги в вітчизняних та закордонних дослідженнях. Так, в працях В. Спіцина, Ю. Болотової досліджуються етапи та методи розпізнавання тексту. T. Zdebska, V. Andrunyk, R. Kempnyk, V. Chyhura проводили аналіз сучасних систем оптичного розпізнавання символів та технологій, які в них застосовуються. Проблеми розпізнавання рукописного тексту висвітлено в наукових студіях H. Beigi, V. Kumar, застосування штучних нейронних мереж у розпізнанні символьної інформації вивчали А. Друки, М. Милешин, О. Солдатова та інші. Аналіз робіт доводить актуальність технології оптичного розпізнавання символів та доцільність теми дослідження.

Мета дослідження полягає в узагальненні сучасних підходів та алгоритмів розпізнавання тексту та аналізі методів розпізнавання в технології OCR.

Основною сучасною технологією перетворення друкованого тексту в електронний формат є оптичне розпізнавання символів OCR (Optical Character Recognition). Аналіз літератури [1, 2, 3, 4] дозволив узагальнити сутність OCR, яка полягає у переведенні різноманітного типу зображення (рукописного, машинописного, друкованого та ін.) до текстового електронного вигляду з метою проведення подальшого редагування, обробки та аналізу інформації.

На теперішній час ця технологія застосовується у великій кількості програмних рішень, пов’язаних з розпізнаванням тексту. Основним завданням таких OCR-систем є призначення фрагменту зображення тексту відповідної символьної інформації.

Останні дослідження технології оптичного розпізнавання тексту базуються на застосуванні принципів роботи зорової системи людини (IPA), таких як [1]:

– принцип цілісності (integrity), згідно з яким об’єкт представлено як сукупність частин і просторових взаємозв’язків між ними;

– принцип цілеспрямованості (purposefulness): будь-яка інтерпретація даних переслідує певну мету. Згідно з цим принципом, розпізнавання являє собою процес висунення гіпотез про цілий об’єкт і цілеспрямованої їх перевірки;

– принцип адаптивності (adaptability) передбачає здатність системи до самонавчання шляхом збереження інформації в процесі її обробки.

Переваги системи розпізнавання, що працює відповідно до принципів IPA, очевидні: саме вони здатні забезпечити максимально гнучку і осмислену поведінку системи. Так, саме за цими принципами на всіх етапах обробки документу діє FineReader компанії ABBY (https://www.abbyy.com/) – одна з найпоширеніших OCR-систем.

Аналіз досліджень [2, 3, 4, 5, 6] дозволив визначити базові етапи алгоритму розпізнавання тексту на графічному зображенні у технології OCR. Розглянемо детальніше.

Попередня обробка зображення, метою якої є зниження рівня шумів (перешкод) і поліпшення якості зображень. Як основні методи попередньої обробки даних можна визначити бінарізацію, зменшення шуму, корекцію перекосу символів тощо [3]. Для придушення різноманітних видів шумів і перешкод можуть використовуватися такі методи попередньої обробки зображень, як лінійне усереднення пікселів по сусідах, медіанна фільтрація, математична морфологія, гаусовське розмиття, методи на основі вейвлет-перетворення, метод головних компонент, анізотропна дифузія, фільтри Вінера тощо [6]. Можуть застосовуватися спеціальні фільтри відновлення пошкоджених зображень.

В деяких дослідженнях цей блок знаходиться у іншій послідовності [4], деталізований на декілька етапів [5] або має інші цілі [2], але застосування цього етапу саме на початку процесу дозволить зменшити суперечливість даних та збільшити ефективність наступних етапів, особливо якщо зображення з текстовою інформацією є кольоровим, нечітким, має півтони тощо.

Сегментація тексту. На цьому етапі відбувається декомпозиція зображення, що містить послідовність символів, на фрагменти – окремі символи. Сегментація здійснюється в кілька етапів: відокремлення рядків, а потім слів та окремих символів. Пошук текстових рядків, як правило, ґрунтується на періодичності та регулярності текстових областей і здійснюється на основі методу Хафа, методу зв’язкових компонент, аналізу горизонтальних, вертикальних і діагональних гістограм [5].

Серед проблем, що найчастіше зустрічаються на цьому етапі – це розпізнавання тексту, накладеного на зображеннях зі складним фоном, символів, що мають різні текстові шрифти і розміри тощо.

Треба зазначити, що останнім часом популярними стали методи (ієрархічні приховані моделі Маркова, згорткові нейронні мережі), які не потребують попередньої сегментації [7].

На етапі *виділення ознак* здійснюється пошук і фіксування характерних структурних ознак кожного символу перед розпізнаванням. Для цього можуть використовуватися різні системи ознак. Існує безліч методів, спрямованих на виділення ознак зображень символів (статистичні, із застосуванням формального евристичного підходу до виділення наборів ознак тощо). Складність полягає в тому, щоб виділити найбільш ефективні ознаки, які дозволять досить добре відрізняти один клас символів від усіх інших.

Розпізнавання (класифікація). Це один з ключових та найбільш складних етапів, під час якого реалізується алгоритм, який розбиває простір ознак на частини, що відповідають заданим класам C1, ..., Cq. [8]. Тобто здійснюється прийняття рішення про відповідність певним визначеним класам.

Подальша обробка отриманого результату полягає у приведенні розпізнаних символів до текстового вигляду з метою подальшого редагування тощо. До основних засобів пост-обробки відносять [2, 5]:

– групування. Окремі символи, які отримані в результаті розпізнавання та знаходяться в документі поруч, зв’язуються один з одним. Тим самим здійснюється формування слова та текстового рядка. Зазвичай проблеми з групуванням символів виникають для рукописних шрифтів, оскільки можливе помилкове визначення роздільника між літерами та словами;

– пошук та виправлення помилок. Існує два основних підходи. Перший перевіряє можливість послідовного знаходження деяких символів. Наприклад, необхідність великої букви після точки або неможливість знаходження послідовності конкретних символів. Другий полягає у використанні словників. Підхід виявився найбільш ефективним для пошуку і виправленні помилок. В словнику відбувається пошук слова, в якому ймовірна помилка. Якщо слово не знайдене, то помилка

підтверджена і слово замінюється на найбільш схоже. Недоліком підходу є необхідність деякого часу для пошуку в словнику та порівняння.

Етап розпізнавання тексту є найбільш важливим в технології OCR, і в залежності від того, які методи будуть застосовуватися на цьому етапі, залежить якість отриманих результатів. Зазвичай методи розпізнавання тексту поділяють на шаблонний, структурний та ознаковий.

Шаблонний метод. Основна ідея полягає у порівнянні зображення окремого символу з усіма шаблонами, що є в наявності у базі, і вибір шаблону з найменшою кількістю відмінностей від вхідного зображення. Ухвалення рішення про належність зображення символу з тестової вибірки до певного класу символів здійснюються за критерієм мінімуму (максимуму) деякої метрики подібності зображення символу і його шаблону [9].

Шаблонні системи мають високу швидкість обробки вхідних даних, але чітко розпізнають лише ті шрифти, шаблони яких наявні в їх базі. Даний підхід вимагає створення шаблону для кожного шрифту. Доцільно застосовувати ці методи для розпізнавання текстових документів, які мають чіткий шрифт, малий відсоток дефектів та шумів.

Прикладом шаблонного підходу є програма розпізнавання TypeReader (http://www.expervision.com/), яка використовує машинно-залежні алгоритми, а також має понад 2600 різних варіантів накреслень символів.

Структурний метод, особливістю якого є опис об'єктів з точки зору їх структури з виділенням окремих складових елементів і зв'язків між цими елементами. У такому підході символ описано як граф, вузлами якого є елементи вхідного об'єкта, а дугами – просторові відносини між ними. Структурними елементами є лінії, що створюють символ.

До переваг структурних методів розпізнавання можна віднести інваріантність щодо типів і розмірів шрифтів. Тому ці методи будуть ефективними при розпізнаванні рукописного тексту.

Основною проблемою є труднощі з ідентифікацією знаків, що мають дефекти (наприклад, розрив лінії або злиття сусідніх ліній), а також невисока швидкодія [10].

Прикладом цього підходу є система Kofax OmniPage (https://www.kofax.com/), основою якої є пошук особливостей кожного символу на основі структурного підходу.

Ознакові методи. В цьому методі аналізуються не символи, а набір властивих їм певних ознак. Зображенню ставиться у відповідність N-вимірний вектор ознак. Розпізнавання полягає в порівнянні його з набором еталонних векторів тієї ж розмірності. Для розпізнавання символів можуть використовуватися різні системи ознак. Задача прийняття рішення про приналежність образу до того чи іншого класу на підставі аналізу обчислювальних ознак, має ряд суворих математичних рішень в рамках детерміністичного та ймовірнісного підходів [10].

Основна перевага ознакових методів – простота реалізації, хороша узагальнююча здатність та стійкість до змін форми символу. Найбільш серйозний недолік цих методів – нестійкість до різних дефектів зображення. Крім того, на етапі отримання ознак відбувається незворотна втрата частини інформації про символ [10]. Тому, ці методи доцільно застосовувати при розпізнавання тексту з низьким відсотком дефектів, або застосовуватися у парі з іншими методам.

Прикладом використання ознакового методу є програма SmartIDReader (https://smartengines.ru/smart-idreader/), яка розроблена компанією SmartEngines. Додаток орієнтований на розпізнавання паспорту та інших ідентифікаційних документів різних країн.

Останнім часом спостерігається розвиток досліджень за напрямом розпізнавання текстової інформації та поява нових підходів, зокрема, *статистичних методів* розпізнавання зображень. В цих методах аналізується зв’язок між віднесенням об’єкта до того чи іншого класу (образу) і ймовірністю помилки при вирішенні цього завдання. Ці методи базуються на теорії прийняття

рішень Байєса. Автори [11, 12] поділяють методи статистичної класифікації на параметричні та непараметричні.

Дослідження доводять, що статистичні методи в деяких випадках мають кращі результаті ніж інші методи. Так, параметричні статистичні класифікатори (MQDF) в умовах тренування з невеликими вибірковими даними мають кращі показники в порівнянні з нейронними класифікаторами [12].

Прикладом реалізації статистичного методу є програма CuneiForm від компанії CognitiveTechnologies (https://launchpad.net/cuneiform-linux) Особливістю CuneiForm є використання алгоритму адаптивного розпізнавання, який генерує внутрішній шрифт для кожного символу, базуючись на символах, що найкраще розрізняються. Також програма має вбудовану експертну систему, що дозволяє проводити аналіз оцінок альтернатив алгоритму розпізнавання та обирати найбільш оптимальний варіант.

Особливої актуальності набувають методи розпізнавання символів із застосуванням *штучних нейронних мереж.* Популярність цих методів обумовлена тим, що штучні нейромережі можуть виконувати роль класифікатора, який добре моделює складну функцію розподілу символів, в тому числі й рукописних або розташованих на зображеннях, тим самим збільшуючи точність розпізнавання порівняно з іншими методами [13].

Нейронні мережі з успіхом можуть застосовуватися в системах розпізнавання тексту, в тому числі й рукописного, в роботі (тренування) з великими вибірковими даними [12], але існують певні недоліки, які треба враховувати: це значний обсяг обчислювальних ресурсів, необхідних для організації процесу навчання нейромережі, що призводить до великих витрат пам’яті [14].

Метод нейронних мереж, який використовує принципи штучного інтелекту, широко застосовується в більшості сучасних систем розпізнавання символів. Таким прикладом є система інтелектуального розпізнавання рукописного тексту PenReader (http://www.paragon.ru/) та ін.

Таким чином, проведені дослідження дозволили узагальнити алгоритм розпізнавання тексту на зображенні у технології OCR та визначити його основні етапи (попередня обробка зображення, сегментація тексту, виділення ознак, розпізнавання (класифікація), подальша обробка отриманого результату).

В роботі розглянуто найбільш поширені методи розпізнавання тексту (шаблонні, структурні, ознакові, статистичні, із застосуванням штучних нейронних мереж). Встановлено, що для розпізнавання текстових документів з досить високою якістю (низький відсоток шумів, чіткий шрифт тощо) доцільно застосовувати шаблонні, ознакові методи. Структурні методи та методи, що базуються на штучних нейронних мережах будуть ефективними у розпізнанні складних шрифтів та рукописного тексту. Для текстових зображень з великим відсотком дефектів, складним фоном, необхідно провести попередньою обробку даних із застосуванням методів зниження рівня шумів і поліпшення якості зображень.

Якщо критерієм вибору методу класифікації є розмір вибіркових даних для тренування, то з невеликими вибірковими даними краще застосовувати параметричні статистичні класифікатори, тоді як нейронні класифікатори кращі показники демонструють з великими вибірковими даними.

Проведені дослідження не вичерпують всіх аспектів цієї проблеми. Перспективними напрямами подальших наукових досліджень є вивчення оптимальних методів та підходів у розпізнанні рукописних документів.

Список джерел:

1. Zdebska T., Andrunyk V., Kempnyk R., Chyhyra V. Optical Character Recognition. URL: http://ena.lp.edu.ua:8080/bitstream/ntb/52204/2/2020v2_Zdebska_T-Optical_Character_Recognition_90-100.pdf. (дата звернення: 15.12.2020).

2. Eikvil L. OCR Optical Character Recognition. URL: https://www.nr.no/~eikvil/OCR.pdf. (дата звернення: 25.12.2020).

3. Hamad K, Kaya M. A Detailed Analysis of Optical Character Recognition Technology Karez. URL: https://dergipark.org.tr/en/download/article-file/236939. (дата звернення: 25.12.2020).

4. Шакун В.А., Ролич О. Ч. Анализ технологии OCR распознавания текста на изображениях. URL: https://core.ac.uk/download/pdf/323161821.pdf. (дата звернення 15.12.2020).

5. Болотова Ю.А., Спицын В.Г., Осина П.М. Обзор алгоритмов детектирования текстовых областей на изображениях и видеозаписях. Компьютерная оптика. 2017. – Т. 41, № 3. С. 441-452. DOI: 10.18287/2412-6179-2017-41-3-441-452.

6. Костенко П.Ю., Василишин В.И., Слободянюк В.В. Уменьшение аддитивного шума на цифровых изображениях с использованием технологии суррогатных даннях. URL: http://www.hups.mil.gov.ua/periodic-app/article/11929. (дата звернення 15.12.2020).

7. Спицын В.Г., Болотова Ю.А., Фан Н.Х. Применение вейвлет-преобразования Хаара, метода главных компонент и нейронных сетей для оптического распознавания символов на изображениях в присутствии импульсного шума. URL: https://cyberleninka.ru/article/n/raspoznavanie-simvolov-na-osnove-veyvlet-preobrazovaniya-metoda-glavnyh-komponent-i-neyronnyh-setey. (дата звернення 15.12.2020).

8. Демин А. А. Обзор интеллектуальных систем для оценки каллиграфии 77-48211/478895. 2012. № 09. URL: https://iu4.ru/publ/2012_ing_vest_09_02.pdf. (дата звернення 15.12.2020).

9. Куксова С.А. Сравнение методов распознавания символов номерного знака автомобіля. URL: https://www.graphicon.ru/html/2015/papers/28.pdf. (дата звернення 15.12.2020).

10. Афонасенко А.В., Елизаров А.И. Обзор методов распознавания структурированных символов. URL: https://cyberleninka.ru/article/v/obzor-metodov-raspoznavaniya-strukturirovannyhsimvolov. (дата звернення 12.02.21)

11. Cheriet M., Kharma N., Cheng-Lin Liu, Ching Y. Suen. Character Recognition Systems: A Guide for Students and Practitioners. Wiley-Interscience, 2007. 360 с

12. Cheng-Lin Liu, Hiromichi Fujisawa. Classification and Learning for Character Recognition: Comparison of Methods and Remaining Problems URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.5945&rep=rep1&type=pdf. (дата звернення 28.01.21).

13. Друки А.А., Милешин М.А. Алгоритмы распознавания рукописных подписей на основе нейронных сетей. Фундаментальные исследования. 2014. № 11-9. – С. 1906-1910. – URL: http://www.fundamental-research.ru/ru/article/view?id=35866. (дата звернення 27.01.2021).

14. Жихаревич С. Е., Остапов І. В. Аналіз методів розпізнавання символів тексту URL: http://nti.khai.edu:57772/csp/nauchportal/Arhiv/REKS/2016/REKS516/Zhikharevich.pdf (дата звернення 27.01.2021).