DOI 10.31891/2307-5732-2022-309-3-50-54
УДК 004.773.2

ШАХОВСЬКА Н. Б.
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-6875-8534
e-mail: Nataliya.b.shakhovska@lpnu.ua
ШЕБЕКО А.
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-0212-8855
e-mail: andrii.shebeko.knm.2018@lpnu.ua

РОЗРОБЛЕННЯ АРХІТЕКТУРИ СИСТЕМИ ОПТИЧНОГО РОЗПІЗНАВАННЯ СИМВОЛІВ З ФОТОГРАФІЙ ДОКУМЕНТІВ

Робота спрямована на створення інформаційної системи оптичного розпізнавання символів з фотографій документів. Складність опрацювання зображень, які являють собою набір пікселів, викликають у комп'ютерів незручності у роботі з такими даними. Для розв'язання цієї задачі можна використати такі підходи: математичні алгоритми, одна складна нейронна мережа або набір з кількох простих нейронних мереж. Звичайні математичні алгоритми складно оптимізувати до необхідної точності при використанні неструктурованого набору даних, яким являються зображення. Одна складна нейронна мережа є як швейцарський ніж, який може усе, але недостатньо якісно. Саме тому ми будемо використовувати 2 типи нейронних мереж з різними архітектурами, кожна з яких спрямована на розв'язання конкретної підзадачі. Ключовими елементами такої технології є: модуль пошуку тексту, модуль розпізнавання символів української та англійської мови, модуль пошуку ключових слів, модуль пошуку необхідних даних.

Ключові слова: *оптичне розпізнавання символів, рекурентні нейроні мережі, аналіз зображення.*

Nataliya SHAKHOVSKA, Andrii SHEBEKO
Lviv Polytechnic National University

DEVELOPMENT OF THE ARCHITECTURE OF DOCUMENT OPTICAL CHARACTER RECOGNITION SYSTEM

This paper aims to develop information technology for document optical character recognition systems. The difficulty of processing images, which are a set of pixels, causes inconvenience in working with such data. This problem can be solved in different ways: usual mathematical approaches, a single complicated neural network, and a set of problem-specific deep neural networks. Usual mathematical approaches perform poor with unstructured data like images. A single neural network is like a swiss knife: it can do many tasks, but none with the best quality. So we will use two different deep neural networks, each for the appropriate part of the problem.

The critical elements of this technology are the module for text detection and segmentation of the image, the module for text recognition in Ukrainian and English languages, the module for parsing multiple keywords, and the module for searching for the final data. The first and second modules consist of several machine learning models with specific architecture, depending on their task. All trained models are tested for accuracy and noise resistance and will be used in the future for searching required data from different document images. Output data of the developed system provide speedup, automation processing images and scans of the documents, reduce the number of mistakes caused by human factor. All data is converted from image pixels into a structured text set represented in the document, which the machine can easily use.

We can use such technology in banking and insurance, where we can send images of documents and they will be automatically processed and converted into user name, surname, date of birth, serial number, and required fields for specific services.

Keywords: *optical character recognition, recurrent neural networks, image analysis.*

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

В даний час технологія ідентифікації по паспорту, водійському посвідченню, студентському квитку відіграє дуже важливу роль у нашому суспільстві. Ця технологія використовується при надані адміністративних послуг, банківській сфері та ін. В минулому дані документу вводилися вручну, що вимагає уважності та певних затрат часу. Питання щодо швидкого та точного отримання інформації з документів стало важливим завданням для науковців та бізнесу.

Автоматична ідентифікація по документах має дуже високу дослідницьку перспективу та значне практичне застосування. Ідентифікація по зображенню документів є задачею обробки зображень, яке зазвичай складається з трьох основних частин: попередня обробка зображення, сегментація зображення та розпізнавання символів. Прикладами використання даної технології є: актуалізація даних в Приват банку, студентська підписка на Ютуб а також ідентифікація на біржі Бінанс.

Незважаючи на широке застосування даної технології, існує багато нюансів, які є досі у процесі удосконалення. Прикладом може бути зчитування зашумленого тексту з зображення. Результатом статистичного дослідження виявлено, що людина, наприклад, при пошуку втручання у зображення, людина знаходить їх у 53% випадків, у той час як нейронні мережі правильно розпізнають близько 99% інформації з зображень [10].

Удосконалення нюансів даної технології може стати важливою як для бізнесу так і для держави. Наприклад, це може дозволити людині легше та швидше отримувати певні переваги, які надають онлайн сервіси ідентифікованим користувачам. А для бізнесу це є більший ступінь захисту інтересів як власних так і користувача [3, 11]. Сфера бізнесу потребує сучасних інформаційних технологій для автоматизації

процесів, покращення продаж, збільшення привабливості та зменшення собівартості продукту. Сучасні мобільні персональні пристрої надають змогу оформити страхування чи іншу необхідну послугу незалежно від місця та часу. А автоматизація обробки даних позволяє задовільнити потребу користувача в з найменшими затратами часу та людських ресурсів.

Аналіз досліджень та публікацій

З розвитком технологій виникає все більша потреба у швидкій та якісній автоматичній обробці даних. Для роботи з таблицями, текстовими або числовими файлами створено безліч систем опрацювання, перевірки, редагування та їх автоматичної генерації з мінімальним втручанням людини. З таким типом даних комп'ютерні системи справляються без проблем та значно швидше та ефективніше за людину. Зовсім по іншому склалась ситуація з обробкою фото та відео. Існує безліч книг, архівів, чеків та звітів у паперовому вигляді, які необхідно оцифрувати, або відеофайли з камер дорожнього руху, на яких необхідно виявити та зчитати номери машин. Ці всі проблеми відносяться до однієї категорії – комп'ютерного зору. А пошук розпізнавання тексту, номерів машин – до оптичного розпізнавання символів.

Ми розглянемо проблему зчитування тексту з фотографій документів з фіксованою структурою: паспорт, водійське посвідчення, студентський квиток. У практичних застосунках, зчитування тексту з зображень ускладнюється розмитістю тексту, складним та неоднорідним фоном, викривленням перспективи. Саме тому у більшості втілень система поділяється на декілька модулів, перший – це попередня обробка, наступний це виділення області де знаходиться текст і останній це розпізнавання самого тексту [1–4]. Також варто зазначити, що дана проблема є значно складнішою чим наприклад розпізнавання номерних знаків, де використовуються однотипні камери з високою роздільною здатністю та статичним фоном [9]. У нашому випадку знімок може бути зроблений з безлічі моделей телефонів, з різних ракурсів, освітленості та ін. [3].

Розглянемо перший етап – попередню обробку, яка полягає у зменшенні кількості шумів, вирівнювання контрасту, яскравості, іноді приведення до відтінків сірого, для збільшення швидкодії наступних кроків алгоритму [1, 3]. Деякі підходи передбачають приведення зображення до бінарного вигляду, наприклад за допомогою алгоритму OSTU [1], вихідне зображення містить два класи пікселів, наступної бі-модальної гістограми: пікселі переднього плану і пікселі тла, потім обчислюється оптимальний поріг, що розділяє два ці класи. Для зменшення шуму застосовується фільтр Гауса з малим (3, 5) розміром ядра [3]. Для ще більш чіткого виділення тексту на зображеннях застосовується оператор Собеля у вертикальному та горизонтальному напрямках, який дає змогу виділити границі, контури. В деяких рішеннях замість оператора Собеля використовують алгоритм Кенні. Оскільки сам документ може являти не 100% площі зображення, виконують пошук границь, далі пошук кутів прямокутника, який описує сам документ а далі відкидають весь непотрібний фон та проводять вирівнювання перспективи [4].

Відповідно до нашого підходу – ми використовуємо окремі нейронні мережі для пошуку та для розпізнавання тексту. Отже після попередньої обробки зображення необхідно знайти сам текст. До широкого розповсюдження нейронних мереж використовувались висхідні підходи: MSER, SWT [4]. Далі набули розповсюдження моделі пошуку об'єктів або сегментації: SSD, R-CNN, FCN [5]. Для пошуку тексту на зображенні ми розглянули 2 новітні підходи з використанням нейронних мереж. Перший – це пошук цілого слова, проте дані підходи показують значне зниження ефективності у складних ситуаціях: текст викривлений, спотворений, зашумлений або надзвичайно довгий і об'єднати текст в 1 обмежувальну коробку стає складно [5]. На противагу, існує підхід з пошуком окремих літер, який їх потім об'єднує в слова та речення. Пошук по окремих літерах усуває складності, які виникають при пошуку зразу цілого слова, саме тому цей підхід ми і обрали. Алгоритм CRAFT – Character Region Awareness for Text Detection (розпізнавання регіону символів для виявлення тексту) полягає у використанні згорткової нейронної мережі яка генерує оцінку регіонів символів та оцінку спорідненості (Affinity score). Оцінка регіонів позволяє локалізувати окремі символи на зображені, а оцінка спорідненості використовується для об'єднання символів у групи, які в подальшому об'єднуються в рядки та стовбці, для подальшої обробки [5].

Отримані області з текстом являють собою прямокутники пікселів, які є досі незрозумілими для більшості комп'ютерних систем. Їх необхідно перетворити у текст. При роботі з нейронними мережами ми зазвичай представляємо дані у вигляді векторів. І вихідні дані у нашій задачі являють собою набір символів, які складаються в слова. Отже, ми маємо справу перетворенням послідовності пікселів у послідовність символів (sequence to sequence). Тобто ми з одного зображення на вході, повинні отримати декілька послідовних передбачень. Довжина залежить від слова, наприклад в англійській мові від двох літер (ok) до п'ятнадцяти (congratulations) і більше. Через таку варіативність, звичайні глибокі згорткові нейроні мережі (DCNN) не можуть бути напряму використані для розв'язання таких задач [6]. Для роботи з послідовностями використовують рекурентні нейронні мережі (RNN), які не вимагають чіткої відповідності між вхідними та вихідними даними [6]. З цього випливає, що ми можемо поєднати переваги кожного з підходів, використавши елементи глибокої згорткової та рекурентної нейронної мережі (CRNN). Перевагами даного підходу є: можливість на пряму отримувати та навчатись інформативних ознак безпосередньо з вхідних зображень, завдяки DCNN на вході, має таку ж властивість, як і RNN, здатні створювати послідовність передбачень, відсутність обмеження по довжині послідовних об'єктів, вимагаючи лише

нормалізації по висоті як при тренуванні, так і при використанні, а також містить меншу кількість параметрів мережі, та відповідно менший час обробки даних чим звичайні глибокі нейронні мережі [6].

Після отримання текстового представлення того, що зображено на фотографії, необхідно знайти необхідні дані. У нашому випадку це поля з ім'ям, прізвищем та ін. Проаналізувавши структуру документів, з якими необхідно працювати, можна прийти до висновку, що розташування даних нагадує таблицю. Найсучаснішим підходом є CascadeTabNet, яка використовується для виявлення кінцевої таблиці та розпізнавання структури з документів на основі зображень. Вона використовується при оцифровуванні фотографій таблиць та певних документів, в основі алгоритму – каскадна нейронна мережа. Даний підхід є надзвичайно ефективним [7], особливо для великих наборах даних. Проаналізувавши рішення даної проблеми, ми можемо значно спростити структуру алгоритму пошуку до двох етапів: пошуку ключових слів, завдяки порівнянню відстані Левенштейна [8]. А також, маючи координати ключового слова – можна здійснити пошук відповідного значення у 2-вимірному просторі, знаючи про розташування шуканого слова відносно ключового [2].

Отже, проаналізувавши та обравши найбільш відповідні існуючі рішення, визначено мету роботи – побудувати систему оптичного розпізнавання символів з фотографій документів. Ключовими елементами такої технології повинні бути: модуль попередньої обробки, модуль пошуку тексту, модуль розпізнавання тексту, модуль пошуку ключових слів, модуль пошуку необхідних слів.

Виклад основного матеріалу

З опрацьованих джерел ми можемо виділити ключові блоки, необхідні для реалізації нашого алгоритму розпізнавання тексту з фотографій документів. Для більшої гнучкості та спрощення реалізації покращень кожен блок є незалежним, і при необхідності його можна покращити або замінити на більш ефективну або нову реалізацію конкретної підзадачі. Також для зручності використання, алгоритм обгорнуто у сервер з кількома точками доступу, який можна розгорнути на окремій робочій станції та за необхідності надсилати фотографії на обробку та отримувати результати обробки як тільки вони готові, або за запитом, а також зберігати усі дані для подальшого використання та аналізу.

Отже ми отримуємо таку архітектуру:

- Ініціалізація та завантаження моделей у пам'ять відеокарти для пришвидшення обробки зображень за допомогою технологій CUDA, cuDNN та розпаралелювання даних. Викликається один раз, при запуску сервера або алгоритму.
- Попередня обробка зображення – приведення зображення до відтінків сірого, застосування фільтра Гауса з розміром ядра 3*3 для зменшення шумів, також є можливість змінювати розмір зображення – зменшити для пришвидшення роботи, або збільшити – для покращення результатів.
- Пошук тексту на зображенні, за допомогою попередньо натренованої моделі CRAFT, яка повертає нам набір областей з текстом, які представляються координатами точок прямокутника, який описує даний текст. Приклад роботи пошуку зображено на Рис. 1.
- Розпізнавання тексту на зображенні. Даний модуль складається з двох окремих модулів: розпізнавання англійських літер та символів за допомогою попередньо натренованої нейронної мережі CRNN, яка у свою чергу представляє архітектуру VGG-BiLSTM-CTC та розпізнавання українських літер та символів за допомогою власноруч натренованої нейронної мережі CRNN, яка у свою чергу представляє архітектуру ResNet-BiLSTM-CTC. Приклад роботи зображено на Рис. 1.

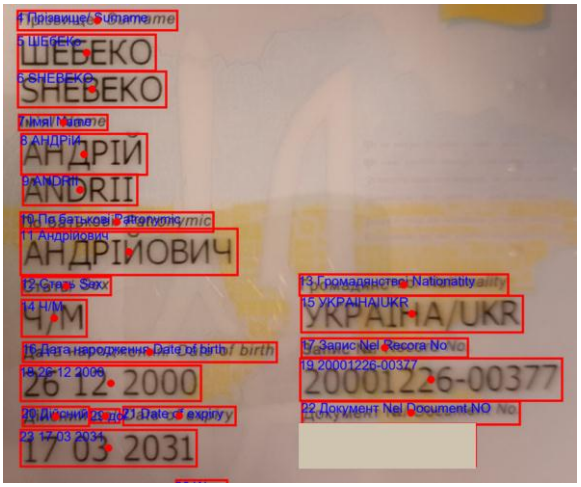

Рис. 1. Елемент документу. Приклад роботи модулів пошуку тексту (червоні рамки) та розпізнавання тексту (синій текст)

- Визначення найкращих співпадінь, пошук ключових слів – маючи набір ключових слів, характерний для певного документу (наприклад 'орган, що видав' – для паспорта або 'категорія' – для водійського посвідчення) ми можемо провести порівняння з усіма словами, які розпізнали нейроні мережі на попередньому кроці. І за допомогою обчислення відстані Левенштейна обрати найкращі співпадіння, які і

будуть нашими ключовими словами для відповідної мови, оскільки вони дублюються обома мовами (Рис. 2).

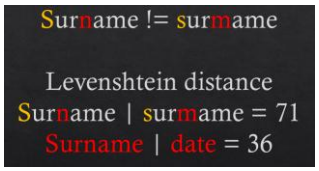

Рис. 2. Приклад застосування відстані Левенштейна для порівняння слів/рядків

- Пошук відповідних до ключових слів даних – маючи координати ключових слів, ми у 2 вимірному просторі можемо провести пошук відповідних даних (наприклад під ключовим словом ‘Прізвище’ знаходиться прізвище власника документу, або справа від ‘Категорія’ знаходиться літера, яка позначає категорію водія транспортного засобу).

Для наочного представлення, усі основні компоненти зображено діаграмі діяльності (Рис. 3).

Функціональні вимоги

Система повинна бути простою у використанні та інтеграції у існуючі застосунки, або при розгортанні на окремій машині. Для цього ми проаналізували застосунок, та виокремили основні елементи, до яких будуть відбуватись виклики ззовні. Це ініціалізація системи та завантаження моделей у пам’ять відеокарти, завантаження та обробка зображень та отримання готових результатів. Ці три елементи були виділені у окремі блоки та обгорнуті у інтерфейси веб-серверу на основі FastAPI, яка дає змогу розгорнути та запустити легкий та гнучкий сервер.

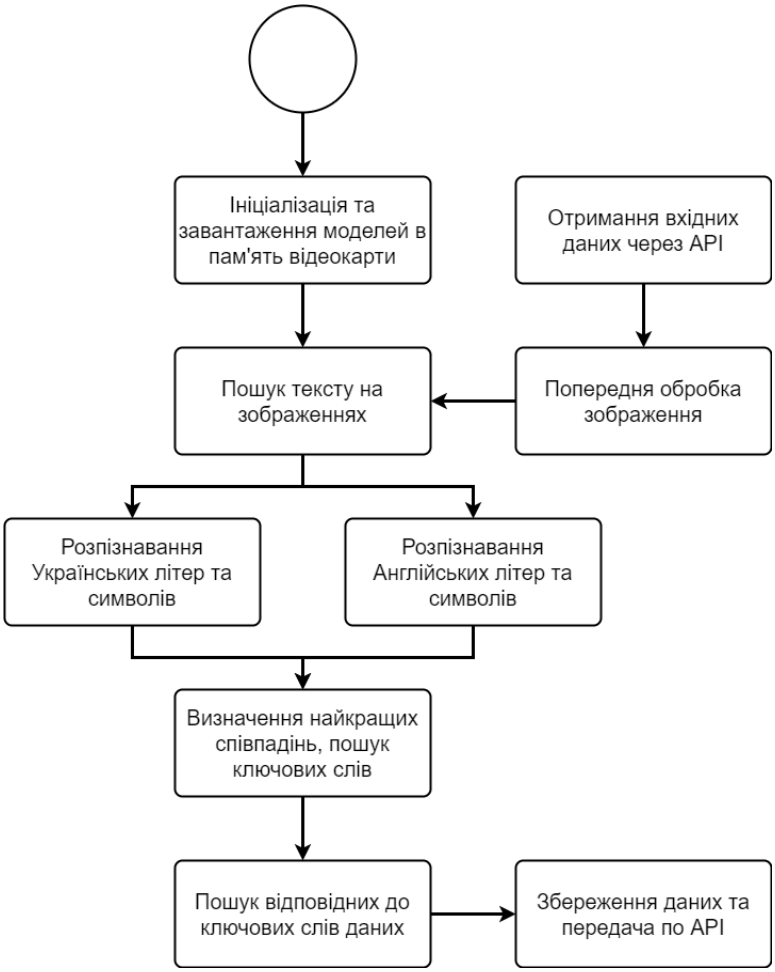

Рис. 3. Діаграма діяльності

Наш сервер реалізує доступ до наступних функцій алгоритму:

- Ініціалізація та завантаження моделей в пам’ять відеокарти – при виконанні запиту, з пам’яті комп’ютера зчитується нейронна мережа, усі необхідні алгоритми з бібліотеки cuDNN, відбувається виділення пам’яті відеокарти та туди завантажуються усі необхідні для роботи дані.
- Отримання вхідних даних – надсилання зображення або групи зображень з мітками про тип документу, які будуть автоматично опрацьовані та збережені на сервері.
- Збереження даних та передача – отримання готових опрацьованих даних з вхідних зображень.

Оскільки система розглядається як компонент для обробки даних – необхідності у витрачанні ресурсів на авторизацію та роботу зі складною базою даних немає. Через відсутність авторизації необхідно розгортати сервер лише у захищеній віртуальній приватній мережі (VPN). А база даних являє собою одну таблицю з двома полями: назва вхідного зображення та результати обробки даного зображення. Для подальшої розробки та аналізу присутня система логування, яка записує усі події у системі, можливі помилки з детальною інформацією про них, а також час виконання кроку алгоритму та задіяні ресурси процесора, відеокарти, оперативної пам’яті. При майбутніх покращеннях, це надає змогу порівняти зміни у якості роботи, стабільності та швидкодії та акцентувати подальшу розробку на найбільш критичних для бізнесу елементах.

Серверна частина була розроблена за допомогою фреймворку FastAPI, який використовує технологію Uvicorn – імплементація веб серверу з інтерфейсом асинхронного серверного шлюзу (ASGI) та за допомогою мови програмування Python. Запити до серверу дотримувалися протоколу HTTPS та були оформлені згідно зі стандартом REST. Вибір фреймворку зумовлений його швидкодією, мінімальними затратами часу на інтеграцію, стандартизацією, яка чітко дотримується специфікацій побудови веб-інтерфейсів [fastapi web site].

На даний момент система підтримує три типи вхідних документів: паспорт, водійське посвідчення та студентський квиток. Під час запиту до сервера, разом з зображенням необхідно передати ключове слово з назвою документу. Необхідність в цьому полягає у тому, що кожен документ має різну структуру та ключові слова. Завдяки модульності системи обробка усіх типів документів відбувається однаково, крім

етапу пошуку ключових слів. З цього випливає простота розширення системи – нам необхідно лише проаналізувати структуру нового документу та внести зміни лише в один модуль, добавивши дані про ключові слова, відповідні до даного документу.

При використанні системи у прикладних застосунках є змога отримати великі об'єми інформації про якість роботи системи. Знаючи нюанси та неточності, які пов'язані з певним типом документу, ми можемо передавати серверу додаткові необов'язкові параметри для корегування алгоритму обробки. Наприклад при великій кількості шумів – ми можемо передати параметр для збільшення розміру фільтру Гауса, або при малому розширені зображення – збільшити параметр масштабування, а при оптимальній якості обробки даних, але недостатній швидкодії – навпаки зменшити параметр масштабування.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Розроблено архітектуру інформаційної технології оптичного розпізнавання символів з фотографій документів. Передбачено модульність алгоритму для спрощення розробки покращень, інтеграцію веб-інтерфейсу для зручності використання з уже існуючими системами та можливості розгортання на окремому сервері. Також наявна підтримка аналізу кількох видів документів, з можливістю додавання нових, без лишніх затрат часу, завдяки модульній структурі.

Система для оптичного розпізнавання символів з фотографій документів може бути використана для мобільних застосунків, таких як банківські, страхові, фінансові. Також можна розгорнути окремий веб-сервіс, до якого клієнти будуть відправляти зображення та отримувати готові результати з оцифрованими текстовими полями відповідно до документу.

References

1. Fang, Xuwei, Xiaowei Fu, і Xin Xu. 2017. «ID card identification system based on image recognition». С. 1488–92 в 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA).

2. Nguyen, Tạn, і Trong Khanh Nguyen. 2019. «A Method for Segmentation of Vietnamese Identification Card Text Fields». International Journal of Advanced Computer Science and Applications 10:415–21.

3. Zuo, Lin, Wenyu Chen, Hong Qu, Li Huang, Zheng Wang, і Yong Chen. 2019. «An Intelligent Knowledge Extraction Framework for Recognizing Identification Information From Real-World ID Card Images». IEEE Access 7:165448–57. doi: 10.1109/ACCESS.2019.2929816.

4. Dat, T. T., L. T. A. Dang, N. N. Truong, P. C. L. T. Vu, V. N. T. Sang, P. T. Vuong, і P. T. Bao. 2021. «An Improved Crnn for Vietnamese Identity Card Information Recognition». Computer Systems Science and Engineering 40(2):539–55. doi: 10.32604/CSSE.2022.019064.

5. Baek, Youngmin, Bado Lee, Dongyoon Han, Sangdoo Yun, і Hwalsuk Lee. 2019. «Character Region Awareness for Text Detection». P. 9357–66 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE.

6. Shi, Baoguang, Xiang Bai, і Cong Yao. 2015. «An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition». arXiv:1507.05717 [cs].

7. Prasad, Devashish, Ayan Gadpal, Kshitij Kapadni, Manish Visave, і Kavita Sultanpure. 2020. «CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition from Image-Based Documents». С. 2439–47 в 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE.

8. Navarro, Gonzalo. 2001. «A guided tour to approximate string matching». ACM Computing Surveys 33(1):31–88. doi: 10.1145/375360.375365.

9. Huang, Q., Z. Cai, і T. Lan. 2021. «A Single Neural Network for Mixed Style License Plate Detection and Recognition». IEEE Access 9:21777–85. doi: 10.1109/ACCESS.2021.3055243.

10. Adobe Communications. «Adobe Research and UC Berkeley: Detecting Facial Manipulations in Adobe Photoshop». 2022 (https://business.adobe.com/blog/the-latest/adobe-research-and-uc-berkeley-detecting-facial-manipulations-in-adobe-photoshop).

11. Lin, G. S., J. C. Tu, і J. Y. Lin. 2021. «Keyword Detection Based on Retinanet and Transfer Learning for Personal Information Protection in Document Images». Applied Sciences (Switzerland) 11(20). doi: 10.3390/app11209528.