УДК 004.93

О. В. Перестюк
Т. Б. Мартинюк
М. А. Очкуров

ЗАСОБИ ОПТИЧНОГО РОЗПІЗНАВАННЯ СПОТВОРЕНИХ ДРУКОВАНИХ СИМВОЛІВ ТЕКСТОВИХ ДОКУМЕНТІВ

Вінницький національний технічний університет

Анотація

Запропонована технологія оптичного розпізнавання поєднує методи попередньої обробки зображень, сегментації та класифікації символів із застосуванням згорткових нейронних мереж і шаблонних алгоритмів. Розроблений програмний засіб забезпечує підвищену стійкість до шумів, деформацій та дефектів друку, що дозволяє покращити точність відтворення текстової інформації у порівнянні з традиційними OCR-системами.

Ключові слова: оптичне розпізнавання тексту, спотворені друковані символи, попередня обробка зображень, згорткові нейронні мережі, OCR

Abstract

The proposed technology integrates image preprocessing, segmentation, and symbol classification methods using convolutional neural networks and template-based algorithms. The developed software demonstrates increased robustness to noise, deformations, and printing defects, resulting in higher text reconstruction accuracy compared to traditional OCR systems.

Keywords: optical character recognition, distorted printed symbols, image preprocessing, convolutional neural networks, OCR

Вступ

Зростання обсягів цифрової інформації та необхідність автоматизації опрацювання текстових документів зумовлюють підвищений інтерес до технологій оптичного розпізнавання тексту. У багатьох сферах — від цифровізації архівів і діловодства до технічної документації — виникає потреба у швидкому та точному перетворенні друкованих матеріалів у машинно-читану форму [1]. Однак значна частина документів містить спотворення, спричинені низькою якістю друку, зношеністю носіїв, нерівномірним освітленням або дефектами оцифрування, що істотно ускладнює коректне розпізнавання символів. Традиційні OCR-системи демонструють обмежену ефективність у таких умовах, оскільки їхні алгоритми не здатні повною мірою адаптуватися до неконтрольованих змін структури символів. Тому розроблення засобів оптичного розпізнавання, стійких до різних видів спотворень, є актуальним напрямом підвищення точності й надійності обробки текстової інформації.

Оптичне розпізнавання спотворених друкованих символів

Реалізацію системи оптичного розпізнавання спотворених друкованих символів можна розділити на два основних функціональних блоки: попередню обробку та нормалізацію зображення, а також сегментацію та виділення фрагментів, які містять окремі символи. У першому блоці здійснюється підготовка вхідного зображення шляхом усунення шумів, вирівнювання яскравості, корекції контрасту та формування однорідної структури документа. Це дозволяє мінімізувати вплив спотворень, пов’язаних з неякісним скануванням, дефектами друку, тінями або нерівномірністю освітлення. У другому блоці виконується визначення області тексту та відокремлення символів для подальшого аналізу та класифікації.

Для оптичного розпізнавання текстових документів використовуємо поєднання згорткової мережі з рекурентною нейронною мережею. CNN-шари виконують виділення локальних просторових ознак символів, тоді як RNN- або LSTM-шари аналізують послідовність цих ознак, забезпечуючи коректне сприйняття тексту як неперервної лінії. Найбільш поширеними модифікаціями рекурентних мереж, що застосовуються в OCR, є LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Unit). Ці

архітектури вирішують проблему згасання градієнта, характерну для класичних RNN, та забезпечують стабільне навчання на довгих послідовностях. У нашому випадку використовуємо архітектуру LSTM.

Архітектура системи складається з таких модулів: модуль отримання зображення, модуль попередньої обробки, модуль сегментації та модуль візуалізації результатів. Отримання зображення реалізується із підтримкою графічних форматів RGB та стандартних джерел оцифрування, включаючи сканери та камери, що забезпечує можливість масштабування системи для роботи з документами різної якості. На цьому етапі визначаються параметри зображення, такі як роздільна здатність, орієнтація сторінки та формат кольору, що впливає на коректність подальших обчислювальних операцій. Попередня обробка включає нормалізацію геометричних параметрів, фільтрацію шумів (медіанну, гаусову), адаптивну бінаризацію та морфологічні операції, які формують структуру зображення, придатну для точного виділення символів [2].

Для забезпечення високої продуктивності та коректної роботи з документами, що містять суттєві спотворення, система виконує попередні етапи обробки таким чином, щоб кожен символ був представлений у стандартизованому фрагменті, який потім подається на етап класифікації. Застосування нормалізації розміру та яскравості забезпечує узгодженість даних, що є критично важливим для точності подальшого розпізнавання, особливо під час роботи зі зношеними, пошкодженими або архівними документами. Модуль візуалізації відображає проміжні та фінальні результати обробки, включаючи виділені текстові області, сегментовані символи та реконструйований текст, що забезпечує можливість контролю якості та налагодження системи.

Для підвищення точності розпізнавання спотворених друкованих символів у системі застосовано комбінований підхід, що поєднує методи класичної сегментації з можливостями нейронних мереж, інтегрованих у механізм розпізнавання Tesseract. На відміну від традиційних OCR-алгоритмів, які базуються переважно на статистичних правилах формування ознак, сучасні нейромережеві моделі дозволяють автоматично виділяти інформативні контури, локальні структури та текстурні характеристики символів, що значно підвищує стійкість системи до шумів, фонового забруднення, незначних деформацій та нерівномірності друку. Формування ознак здійснюється через багаторівневу обробку, у межах якої вхідне зображення перетворюється у внутрішнє представлення, придатне для подальшої класифікації [3].

Сегментація тексту виконується методом аналізу геометричної структури документа, включаючи визначення границь рядків, слів та окремих символів. У випадках спотворених або частково пошкоджених документів застосовуються додаткові процедури корекції, такі як адаптивне визначення з’єднаних компонентів, морфологічна фільтрація та виявлення областей зі злитими контурами. Це дозволяє уникнути помилкових сегментацій, що можуть виникати через дефекти друку, перекоси сторінки або непропорційність елементів символів. Після сегментації кожен фрагмент нормалізується та подається на класифікатор, який за допомогою згорткової нейронної мережі прогнозує ймовірність належності символу до певної категорії.

Особливу складність становить розпізнавання символів, у яких спотворення змінюють геометрію настільки, що окремі фрагменти символу візуально наближуються до інших класів, наприклад, «O» і «0», «I» і «l» або «S» і «5». Для вирішення таких випадків у системі використано механізм контекстної корекції, який аналізує можливі варіанти реконструкції слова з урахуванням словникових моделей та статистичних закономірностей мови. Це дозволяє значно знизити частоту помилок при класифікації складних або частково пошкоджених символів і відновлювати текст навіть у випадках, коли вхідне зображення містить суттєві дефекти.

Архітектура програми розробленого засобу побудована за принципом послідовного виконання етапів, де результат попереднього модуля є вхідними даними для наступного. Такий підхід забезпечує логічну узгодженість процесу та дозволяє контролювати якість обробки на кожному з етапів. У процесі функціонування системи спочатку здійснюється завантаження зображення документа, після чого активується модуль попередньої обробки. На цьому етапі відбувається компенсація можливих нахилів, спотворень і нерівномірності освітлення, що є типовими проблемами при роботі зі сканованими або фотографованими текстами. Застосування фільтрації та контрастного підсилення дозволяє підготувати чітке та однорідне зображення, придатне для точного розпізнавання символів.

На наступному етапі відбувається передача підготовленого зображення до модуля розпізнавання, де виконується перетворення графічного подання у текстову інформацію. Для цього використовується нейромережевий алгоритм, навчений на великих наборах друкованих символів різних шрифтів і мов. Нейронна мережа аналізує структуру символів, їхні контури та співвідношення, що дає змогу точно

відтворювати текст навіть у разі наявності незначних спотворень або дефектів друку. Отриманий результат розпізнавання передається до модуля постобробки, де здійснюється остаточне форматування тексту, виправлення типових помилок і збереження результатів у зручній для користувача формі.

Для забезпечення високої точності розпізнавання була сформована навчальна вибірка, що включала зображення символів різних шрифтів, розмірів, товщини ліній та рівнів деградації, характерних для архівних і низькоякісних документів. Набір даних було розділено на навчальну та валідаційну частини, що дозволило контролювати якість узагальнення моделі та запобігти перенавчанню. У процесі підготовки вибірки використовувались методи штучного аугментування, такі як додавання шумів, варіації контрастності, повороти та деформації, що дало змогу моделювати реальні умови, у яких можуть перебувати текстові документи. Завдяки цьому нейронна мережа здобула здатність коректно обробляти широкий спектр спотворень, що характерні для практичних задач OCR [4,5].

Навчання нейронної мережі здійснювалося за методом оберненого поширення помилки. Функція втрат включала компоненти, що відображають точність класифікації символів, зокрема різницю між цільовими мітками та прогнозованими значеннями. Оптимізатор Adam забезпечував адаптивну корекцію вагових коефіцієнтів, що сприяло швидкій та стабільній збіжності моделі. Особливу увагу приділено забезпеченню стійкості до помилкових класифікацій, які виникають при значних пошкодженнях контурів або при низькій різкості символів. Для цього застосовувалися регуляризаційні методи та оцінювання проміжних результатів на валідаційній вибірці, що дозволило визначити оптимальну конфігурацію моделі.

Після завершення навчання проведено тестування системи на незалежному наборі даних, який містив символи зі складними варіаціями форми та яскравості. Отримані результати показали, що запропонований підхід забезпечує високу точність розпізнавання навіть за умов значних спотворень, нерівномірності тону та часткової втрати фрагментів символів. Модель продемонструвала здатність коректно реконструювати текстові фрагменти, у тому числі символи зі схожою геометрією, що раніше становили найбільшу складність для традиційних OCR-систем.

Таким чином, розроблений підхід довів практичну ефективність у задачах цифровізації текстових документів, обробки технічної документації та автоматизованого введення даних у комп'ютерні системи.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Шелестов А. Ю. Методи та засоби оптичного розпізнавання текстової інформації / А. Ю. Шелестов // Вісник Національного технічного університету України «КПІ». — Серія «Інформатика, управління та обчислювальна техніка». — 2017. — № 65. — С. 88–94.
2. Afzal M. Z., Kölsch A., Ahmed S., Liwicki M. Deep learning-based document segmentation and recognition: a survey // Pattern Recognition. — 2019. — Vol. 86. — P. 106–132.
3. Xiao-Feng Wang, Zhi-Huang He, Kai Wang, Yi-Fan Wang, Le Zou, Zhi-Ze Wu. A survey of text detection and recognition algorithms based on deep learning technology. 2023, Neurocomputing, p. 126702. [Електронний ресурс]. – Режим доступу: https://doi.org/10.1016/j.neucom.2023.126702.
4. What's New in Tesseract OCR. Google Developers. [Електронний ресурс]. – Режим доступу: https://opensource.google/projects/tesseract .
5. Deep Learning for OCR: A Comprehensive Guide. PapersWithCode. [Електронний ресурс]. – Режим доступу: https://paperswithcode.com/task/optical-character-recognition .

Перестюк Олександр Вікторович — студент групи 2КІ-24м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: alexandr.perestyuk@gmail.com

Мартинюк Тетяна Борисівна — доктор технічних наук, професор кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця.

Очкуров Микола Андрійович — старший викладач кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця.

Perestiuk Oleksandr V. — student of the 2KI-24m group, faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia e-mail: alexandr.perestyuk@gmail.com

Martyniuk Tetiana B. — Dr. Sc. (Eng), Professor of department of the computer engineering, Vinnytsia National Technical University, Vinnytsia.

Ochkurov Mykola A. — Senior lecturer of department of the computer engineering, Vinnytsia National Technical University, Vinnytsia.