УДК 004.93

К. В. Поліщук
М. А. Очкуров

РОЗПІЗНАВАННЯ ПОШКОДЖЕНИХ ДРУКОВАНИХ ТЕКСТІВ ІЗ ВИКОРИСТАННЯМ БАГАТОРІВНЕВОГО АНАЛІЗУ ДОКУМЕНТА

Вінницький національний технічний університет

Анотація
Запропоновано підхід по розпізнаванню пошкоджених текстових документів із використанням згорткової нейронної мережі та шаблонного методу.
Ключові слова: розпізнавання текстових символів, згорткова нейронна мережа, шаблонний метод.

Abstract
The approach to the recognition of damaged text documents using a convolutional neural network and a template method is proposed.
Keywords: text symbol recognition, convolutional neural network, template method.

Вступ

Одним з актуальних напрямків інформаційних технологій є завдання обробки та розпізнавання зображень текстових документів. Це використовується в системах для розпізнавання тексту, розпізнавання різного типу бланків статистичного обліку, бланків податкових декларацій, різного виду банківські рахунки та інших типів документів [1, 2]. Методи автоматичного розпізнавання образів та їх реалізація у системах оптичного читання текстів (Optical Character Recognition, OCR-системах) є однією із найбільш яскравих та дієвих технологій штучного інтелекту [3]. Ще складнішим стає це завдання при виконанні автоматизованого розпізнавання пошкоджених друкованих текстів, яке в даний час поки ще не вирішено у повній мірі та є однією з найбільш актуальних та складних задач розпізнавання даних. Розгляду одного із підходів по виділенню та розпізнавання текстових документів присвячений даний матеріал.

Розпізнавання текстових символів

Відомі на теперішній час підходи для розпізнавання символів текстового документу можна згрупувати у чотири групи. Це методи шаблонні, ознакові, структурні та на основі нейронних мереж [4]. Кожен із цих підходів характеризується своєю складністю та ефективністю. Самим простим є шаблонний метод, у основі якого лежить послідовне порівняння отриманого зображення символу із еталонними зображеннями. Найкращі результати розпізнавання показують методи на основі використання нейронних мереж. Але при знаходженні у текстовому документі значної кількості символів із пошкодженим станом ні один із цих методів не дає бажаного результату.

Пропонується об’єднати методи на основі нейронних мереж та шаблонні методи для підвищення ефективності розпізнавання пошкоджених текстових символів. Пошук та розпізнавання символів у цифрових зображеннях виконується за ряд етапів, загальна послідовність для виділення об’єктів буде такою.

Початковим етапом усіх алгоритмів розпізнавання є етап попередньої обробки. На цьому етапі покладається виконання таких завдань: підвищення якості зображення за рахунок фільтрації, видалення завад та інші операції, що мають на меті підвищити якість зображення.

Наступним етапом є виділення тексту на зображенні як регіону інтересу. На цьому етапі роботи на бінаризованому зображенні виділяється безпосередньо область, на якій знаходиться текст, що підлягає розпізнаванню. На цьому етапі виконуються операції сегментації та нормалізації текстових фрагментів та багаторівневий аналіз текстового документа. Текст розділяється на зручні

для аналізу складові частини. На даному етапі послідовно виконується поділ тексту на окремі рядки (сегментація рядків), потім поділ рядків на окремі слова (сегментація слів), а надалі поділ виділених слів на елементарні складові частини у вигляді символів. Потім виконується нормалізація символів до необхідних розмірів для виконання процесу розпізнавання. Для виконання ряду дій на цих початкових етапах роботи даної послідовності були використані деякі вже розроблені програми із бібліотеки із відкритим кодом OpenCV [5].

На наступному етапі виконуємо розпізнавання символів із використанням нейронної мережі глибокого навчання. Попередньо відбувається її налаштування на розпізнавання текстових символів. Нерозпізнані символи надходять на інший модуль програми, де відбувається їх порівняння із набором еталонних символів.

У модулі розпізнавання пошкоджених символів здійснюється їх розпізнавання із використанням шаблонного методу та наявної бази даних еталонних символів. Тут використані такі запропоновані підходи.

Визначаємо тип шрифту по попередніх розпізнаних символах (латиниця чи кирилиця). Виконуємо звертання до відповідного набору символів для розпізнавання. Вибираємо перший символ із визначеного набору. Визначаємо шляхом логічного множення кількість збігів по накладанню матриці еталонного символу із матрицею опису вибраного символу. Підраховуємо кількість збігів для силуету символу у вибраному еталонному зображенні. Підраховуємо кількість збігів для фону символу у вибраному еталонному зображенні. Знаходимо сумарну кількість збігів для силуету та фону. Запам'ятовуємо отримане значення. Переходимо до вибору наступного еталонного символу. Повторюємо дії попередніх пунктів по визначенню сумарної кількості збігів для вибраного нового еталонного символу. Порівнюємо результати двох попередніх обчислень. Фіксуємо більше значення із отриманого порівняння. Переходимо до наступного еталонного символу. Повторюємо дії пунктів по визначенню сумарної кількості збігів для вибраного нового еталонного символу та його порівняння із попереднім символом. Перевіряємо, чи всі еталонні символи із вибраного набору символів пройшли операцію порівняння. Виводимо силует символу, який отримав найбільшу кількість збігів при порівнянні еталонних символів із аналізованим символом.

Розробка ефективного програмного забезпечення є важливою задачею для виділення та розпізнавання об'єктів текстового зображення. Для вирішення цієї задачі створена програмна реалізація запропонованого підходу з використанням мови програмування Python та бібліотеки OpenCV, яка дозволяє здійснити процес виділення та розпізнавання символів текстових документів.

Запропонований підхід може бути використаний у комп'ютерних системах виділення та розпізнавання об'єктів за отриманим цифровим зображенням текстових документів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Шапиро Л. Компьютерное зрение. / Л. Шапиро, Дж. Штокман - М.: Бином, 2009. – 763 с.
2. Желтов С. Ю. Обработка и анализ изображений в задачах машинного зрения / С. Ю. Желтов. - М.: Физматкнига, 2010. - 672 с.
3. Оптичне розпізнавання символів [Электронный ресурс] – Режим доступа: https://ua.wikipedia.org/wiki/ Оптичне_розпізнавання_символів.
4. Жихаревич В. В. Аналіз методів розпізнавання символів тексту / В. В. Жихаревич, С. Е. Остапов, І. В. Миронів // Радіоелектронні і комп'ютерні системи. – 2016. – № 5. – С. 137–142.
5. OpenCV library [Электронный ресурс]. – Режим доступа: https://opencv.org/.

Катерина Валеріївна Поліщук - студентка групи 1КІ-19м факультету інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: 2ki15b.polishchuk@gmail.com.

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

Kateryna V. Polishchuk - students, Department of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: 2ki15b.polishchuk@gmail.com..

Mykola A. Ochkurov — Senior lecturer of the Computer Techniques Chair, Vinnytsia National Technical University, Vinnytsia.