

МЕТОД РОЗПІЗНАВАННЯ СИМВОЛІВ З МАЛОЮ КОНТРАСТНІСТЮ

Суприган Олена, Дячук Тетяна

Вінницький національний технічний університет

Анотація

Дані тези призначені для надання ознайомчої інформації про дослідження, яке присвячене вдосконаленню методу розпізнавання символів з малою контрастністю. Для покращення якості та швидкості розпізнавання було застосовано комплекс методів, що включає в свій склад: морфологічні алгоритми, перетворення Хафа та векторний простір, порівняння з шаблоном.

Abstract

The current abstract is intended to provide information about the research, which is devoted to the improvement of the method of character recognition with low contrast. In order to improve the quality and speed of recognition the set of methods have been applied, which includes: morphological algorithms, Hough transform and vector space, comparison with the template.

Вступ

На даний час розвиток технологій сприяв появі таких понять як електронна документація, електронні книги, електронні бібліотеки. Вони надають більше можливостей для своїх користувачів у порівнянні з фізичними носіями даних, що робить їх все більш популярними та використовуваними. Приймаючи до уваги цей фактор, накопичена паперова документація потребує переведення в електронний варіант для подальшого її використання. Переведення паперової документації в електронний варіант досить довготривала операція при виконанні вручну. Тому для цієї задачі часто використовуються програмні засоби, що автоматизують даний процес. При цьому існуючі програмні методи не завжди є достатньо ефективними – більшість систем орієнтовано на використання одного з методів розпізнавання символів паперових носіїв, що не завжди підходить для різнорідних вхідних даних; може мати низьку швидкість розпізнавання або малу якість розпізнавання символів.

Тобто, існує проблема недосконалості сучасних методів розпізнавання символів, що не забезпечують високу якість та швидкість розпізнавання. Тому необхідно розробити метод, який забезпечить високу якість розпізнавання символів та швидкість перетворення паперових документів в електронний варіант.

Таким чином, дане дослідження полягає у аналізі існуючих методів розпізнавання символів та розробці методу, який буде забезпечувати достатню якість та швидкість розпізнавання символів з малою контрастністю.

Аналіз відомих методів розпізнавання символів

Найперспективнішими методами, які використовуються для задачі розпізнавання символів, є нейронні мережі, метод найближчого сусіда, клітинні автомати та розпізнавання за шаблоном.

Нейронні мережі використовують для вирішення поставленої задачі через їх здатність до вирішення задач при невідомих закономірностях та нечітких даних та стійкість до шумів у вхідних даних. Проте є ряд недоліків, серед яких: складність структури для проектування, вимога до значних обчислювальних ресурсів. Тоді як метод найближчого сусіда дає змогу зменшити кількість об'єктів навчальної вибірки та зменшити вимоги до обчислювальних ресурсів, проте потребує досить широкого різноманіття даної вибірки, що в умовах поставленої задачі не є прийнятним вирішенням через велику кількість ймовірних варіантів відображення символів. Клітинні автомати

також надають можливість зменшити затрати обчислювальних ресурсів через реалізацію паралельних обчислень, проте з накопиченням та збільшенням кількості правил ця перевага проявляється не настільки явно. Метод розпізнавання за шаблоном є досить простим при програмній реалізації та має перевагу у швидкості обчислень, але у той же час існують випадки, коли приналежність невідомого символу до того чи іншого шаблону не визначається однозначно, що може викликати його некоректне розпізнавання та відображення у результаті [1].

Оскільки наявні засоби розпізнавання символів не можуть забезпечити одночасно високі швидкість та якість розпізнавання, буде доцільніше використати комплекс методів до підходу розпізнавання символів з малою контрастністю.

Реалізація обраного комплексу методів

Для реалізації цієї задачі пропонується застосувати такий набір методів як: морфологічні операції, використання векторного простору та перетворення Хафа, порівняння з шаблоном.

Запропонований комплекс вирішує такі основні проблеми як: використання вхідних даних з низькою якістю; різноманіття шрифтів, що використовуються; зменшення кількості шаблонів для порівняння, що значно збільшить швидкість та якість розпізнавання.

Поєднання запропонованих методів було реалізовано в інтелектуальному модулі за описаними далі принципами роботи.

Вхідні дані у вигляді цифрового зображення подаються на вхід модуля та конвертуються у GIF формат, з яким працює модуль. Потім виконується групування усіх наявних кольорів зображення та підрахунок кількості пікселів на зображенні, що передають визначені кольори. Обираються групи кольорів з найбільшою кількістю пікселів та для кожної з них будується бінарне зображення, де колір групи передається білим, а фон чорним. До отриманих бінарних зображень застосовується перетворення Хафа, яке позбавляє зображення залишкових шумів. Після перетворення Хафа використовуються морфологічні алгоритми для відновлення втрачених даних символів, які були пошкоджені. Також застосування морфологічних операцій збільшує шанс правильного відтворення символу. Для результуючого зображення знаходяться межі між символами та виконується їх відокремлення одне від одного для виключення можливості злиття символів та їх некоректного розпізнавання. На основі цих зображень за допомогою векторного методу відтворюється векторне зображення символу та порівнюється з існуючими еталонами наявної бази даних. Після виконання цього алгоритму дій результат має, у порівнянні з іншими методами, зазначеними раніше, досить високу точність розпізнавання, що визначається кількістю правильно розпізнаних символів серед набору всіх символів зображення.

Запропонований метод розпізнавання символів з малою контрастністю було реалізовано на мові програмування Python, що дозволила виконати об'єднання зазначених методів у комплекс при невеликому обсязі програмного коду, та проведено його тестування. Дослідження роботи програмного засобу проводилось на репрезентативній вибірці в 50 текстів технічного, наукового та художнього стилю з різними показниками їх якості. Із цих текстів: 36 було розпізнано з мінімальною кількістю помилок (3-7 помилок на 100 символів) та 24 тексти з допустимою кількістю помилок (8-13 помилок на 100 символів). Отже, результат тестування довів раціональність використання даного підходу та подальше застосування у якості системи або модуля розпізнавання символів.

Список використаних джерел:

1. Способы распознавания (оцифровки) текста [Електронний ресурс]. – Режим доступу: <http://mcgrp.ru/sposobyi-raspoznavaniya-otsifrovki-teksta-article.html> - Назва з екрану.