УДК 004.056

К. В. Поліщук

РОЗПІЗНАВАННЯ ДЕФОРМОВАНИХ СИМВОЛІВ ТЕКСТОВИХ ДОКУМЕНТІВ

Вінницький національний технічний університет

Анотація

Запропоновано метод визначення із цифрового зображення серії та номера паспорта за допомогою програмного продукту.

Ключові слова: оптичне розпізнавання символів, деформовані символи.

Abstract

A method is proposed to precisely determine the series and passport number from the image using the software product.

Keywords: optical character recognition, image distortion.

Вступ

Розпізнавання образів (об'єктів, сигналів, процесів, ситуацій чи явищ) – це задача ідентифікації об'єкта або визначення його властивостей за зображенням (оптичне розпізнавання), за аудіозаписом (акустичне розпізнавання) чи за іншими характеристиками. Важливою складовою класифікації об’єктів є процедура машинного навчання, метою якої є побудова вирішального правила або функції, що класифікує об’єкти за ознаками. Другою частиною процесу класифікації є процедура узагальнення, яка полягає у класифікації нової вибірки за допомогою вирішального правила, сформульованого у процесі навчання [1, 2]. Значне місце серед методів розпізнавання посідають розпізнавання друкованих символів. Розпізнавання тексту на зображеннях - дуже актуальна тема для досліджень, яка дозволяє вирішувати ряд наукових та прикладних задач. Сучасні методи розпізнавання символів використовуються для вирішення широкого кола завдань, як офісних, так і спеціалізованих. Існує велика кількість програмного забезпечення для розпізнавання текстів, наприклад, ABBYY FineReader, CuneiForm та ін. Крім того, існує багато комерційних і державних проектів по розпізнаванню тексту для вузькоспеціалізованих завдань, наприклад, системи по розпізнаванню номерних знаків автомобілів або серії та номера паспорта особи. При створенні складних, високонавантажених систем з розпізнавання стандартизованого тексту може виникнути проблема швидкості розпізнавання даних. Тому актуальним є розробка поліпшеного методу розпізнавання тексту із деформованими символами для підвищення продуктивності при обробці масивів даних із текстовими документами.

Результати дослідження

Для вирішення задачі розпізнавання символів на теперішній час виділяють три основні класи методів: шаблонний, ознаковий і структурний [3]. У шаблонному методі здійснюється перетворення вихідного зображення у растрове і потім виконується порівняння його по точкам з усіма наявними в базі шаблонами. Даний тип методів має досить високу стійкість до дефектів зображення, а також високу швидкістю обробки вхідних даних. Структурні методи представляють символ у вигляді графа, в якому множина вершин складають структурні одиниці вихідного зображення, а множина ребер - просторові відносини між ними. Під структурними одиницями в даному випадку маються на увазі складові символів. Розбиття на структурні одиниці і їх аналіз вимагає досить великих обсягів пам'яті, а також великої кількості процесорного часу. Ознакові методи аналізують набір властивих символам певних ознак. Однак це ж є і недоліком цих методів, так як об'єкт замінюється своїм спрощеним уявленням, то велика частина інформації про зображення символу втрачається. Як наслідок - зменшується ймовірність однозначного визначення символу. Тому для розпізнавання тексту із деформованими символами було вирішено використати шаблонний метод.

Для забезпечення прийнятної точності розпізнавання символів у алгоритмі шаблонного підходу виконуємо попередню обробку зображення. Основним етапом попередньої обробки даного підходу є представлення зображення символів у вигляді растрового зображення, потім виконуємо

нормалізація розміру та товщини елементів зображення, що представляють складові символів паспортних даних. Принцип роботи алгоритмів шаблонного підходу заснований на прямому порівнянні зображення символу, що розпізнається, з усіма шаблонами, що зберігаються в базі шаблонів. Найбільш підходящим є шаблон, який має найменшу кількість незбіжних пікселів, що відрізняють цей шаблон від зображення символу, що розпізнається. Так як розміщення паспортних символів має строго визначений порядок, то для розпізнавання використовуємо базу даних по символах літер та символах цифр, які мають визначену кількість елементів представлення символу. До переваг шаблонного підходу відносяться простота реалізації, висока швидкість розпізнавання і хороша стійкість до дефектів зображень символів.

Для підвищення ефективності роботи алгоритму по розпізнаванню символів паспортних даних запропоновано використати метод Хаара [4]. Метод Хаара, приймаючи на вхід зображення, визначає, чи є на ньому шуканий об'єкт, тобто виконує задачу класифікації, розділяючи вхідні дані на два класи (присутній шуканий об'єкт, відсутній шуканий об'єкт). Правильно навчений каскад Хаара має високу швидкість виконання класифікації, а також стійкість до різного роду відхилень.

Признак Хаара являє собою набір прямокутних областей зображення, які примикають один до одного і розділенні на дві групи. Можливих признаків Хаара існує велика множина (різноманітні комбінації областей різної ширини і висоти з різними позиціями на зображені). Для розпізнавання паспортних символів створюємо початковий набір ознак, що залежить від реалізації вибраних літер алфавіту та цифр та їх представлення у вигляді множини точок.

Щоб вирахувати значення конкретного признаку Хаара для якого-небудь зображення, потрібно додати яскравості пікселів зображення в першій і другій групах прямокутних областей окремо, а потім відняти із першої отриманої суми другу. Отримана різниця і є значенням конкретного признака Хаара для даного зображення.

Це досить оптимальний метод у відношенні складність реалізації та якість опрацювання із урахуванням технічних вимог. Також він дозволяє опрацьовувати деформовані об'єкти з найменшими спотвореннями.

Висновки

Встановлено, що алгоритм Хаара може бути використаний для вирішення задачі пошуку деформованих об'єктів (серія та номер паспорта) на зображенні. Запропонований підхід може бути використаний у комп'ютерних системах розпізнавання паспортних даних за отриманим цифровим зображенням.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Заяць В. М. Методи розпізнавання образів. Навч. посібник. / В. М. Заяць, Р. М. Камінський.- Львів, видав. національного університету «Львівська політехніка», 2004. – 176 с.
2. Б.Б. Круліковський, А.І. Сидор, О.М. Заставний, Я.М. Николайчук // Теоретичні основи розпізнавання багатомірних образів у Хеммінговому просторі // Науковий вісник НЛТУ України. 2016. Вип. 26.3.- С. 361-367.
3. Желтов, С. Ю. Обработка и анализ изображений в задачах машинного зрения / С. Ю. Желтов. — М.: Физматкнига, 2010. — 672 с.
4. Белых Е. А. Обучение каскадов Хаара // Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика. 2017. Вып. 1 (22). С. 41-53.

Поліщук Катерина Валеріївна — студентка групи 2КІ-15б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: 2ki15b.polishchuk@gmail.com

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

Polishchuk Kateryna V. — student group 2CE-15b, Department of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail : 2ki15b.polishchuk@gmail.com

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