ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
(повне найменування вищого навчального закладу)
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
(повне найменування інституту, назва факультету (відділення))
КАФЕДРА ПРОГРАМНИХ ЗАСОБІВ І ТЕХНОЛОГІЙ
(повна назва кафедри (предметної, циклової комісії))

Пояснювальна записка

до кваліфікаційної роботи
магістра
(освітній рівень)

на тему: «Розробка програмних інструментів розпізнавання тексту по зображеннях документів»

Виконав: студент 2 курсу, групи 5зПР2
спеціальності
121 - «Інженерія програмного забезпечення»
(шифр і назва спеціальності)

Шерстюк Володимир Григорович
(прізвище та ініціали)
Керівник д.т.н., професор Жарікова М.В.
(прізвище та ініціали)
Рецензент _____
(прізвище та ініціали)

Херсон - 2020

АНОТАЦІЯ

Загальний обсяг роботи: 85 аркушів основного тексту, 14 ілюстрації, 6 таблиць. При підготовці використовувалася література з 30 різних джерел.

Робота складається із вступу, трьох розділів та двох додатків.

Об’єкт дослідження – процес розпізнавання та вилучення специфічних даних із документа по зображенню.

Предмет дослідження – засоби автоматизації процесу розпізнавання та вилучення специфічних даних із документа.

Мета дослідження – розробка програмної системи розпізнавання та вилучення специфічних даних із зображень документів та тестування системи із метою аналізу її ефективності.

Для досягнення мети дослідження поставлено і вирішено такі завдання:

- аналіз інструментів розпізнавання тексту;
- дослідження особливостей та вимог системи розпізнавання реквізитів документу по зображенню;
- проектування та розробка системи розпізнавання реквізитів документу по зображенню.

Наукова новизна полягає у розробці новітніх ефективних алгоритмів вилучення реквізитів документа із його зображення.

Практична цінність полягає в тому, що отримана система розпізнавання реквізитів документів по зображенню на основі відкритої бібліотеки Nicomsoft OCR може бути впроваджена у будь-яку корпоративну систему електронного документообігу на комерційній основі.

Ключові слова: РОЗПІЗНАВАННЯ ТЕКСТУ, КОМП’ЮТЕРНИЙ ЗІР, ОПТИЧНЕ РОЗПІЗНАВННЯ СИМВОЛІВ, СИСТЕМА РОЗПІЗНАВННЯ РЕКВІЗИТІВ, РОЗПІЗНАВННЯ ЗОБРАЖЕННЯ.

ЗМІСТ

ВСТУП

Сьогодні будь-яка сучасна організація чи компанія вводить електронний документообіг всередині компанії та з іншими компаніями. Кожен документ має певні реквізити та ключові слова, за якими його можна швидко знайти у базі даних. В залежності від типу документу, реквізити можуть бути різні, та представлені у різних типах даних. В будь-якому випадку, сьогодні майже у всіх організаціях, введенням реквізитів та даних по документам займається співробітник. Цей процес автоматизований хіба що, для компаній, в яких використовується один тип документу. Але коли типів таких документів багато - автоматизувати цей процес стає важче.

Розвиток нейронних мереж та систем комп’ютерного зору проходить саме зараз, але вже досяг певних результатів. Багато систем вміють досить точно розпізнавати об’єкти, що зображені на фотографіях, в тому числі і текст в документах.

Мета цієї магістерської кваліфікаційної роботи – вивчити можливість використання бібліотек із розпізнавання тексту для автоматизації розпізнавання їх реквізитів та розробка такої системи.

Існує велика кількість OCR бібліотек, потрібно було проаналізувати їх та порівняти. Потрібно було також розробити певний підхід, за яким система би могла «навчатись» точно викорінювати потрібні дані із документа.

Після цього вже можна було приступати до автоматизації бізнес-процесу із виявлення реквізитів та збереження їх у базі даних. Поміж таблиці із отриманими реквізитами у базі даних також потрібно було створити схему для збереження даних, потрібних саме для точного викорінення реквізитів.

Не існує єдиної, повноцінної системи, яка здатна використовувати OCR бібліотеки для автоматизації бізнес-процесу розпізнавання та викорінення реквізитів документів. Така система дозволить скоротити час на розпізнавання даних документів та їх збереження у базі даних. А висока точність розпізнавання також дає можливість проводити автоматично аналіз

документів. У світі, де намагаються автоматизувати всі робочі процеси, така система точно знайшла б своє місце.

Відсутності таких готових систем є пояснення. Зазвичай, у кожної компанії своя політика документообігу, та свої формати документів. Поміж розпізнавання, треба зробити аналіз документу та викорінити потрібні дані. При великий кількості різних типів документів це зробити досить важко. Тому єдиним вирішенням такої проблеми є створення у системі модуля, який би дозволяв «навчати» систему точно розпізнавати реквізити в кожному типі документу.

Звичайно, якість отриманих даних, а отже й якість всієї системи в цілому напряму залежить від точності розпізнавання символів. Тому був проведений аналіз найпопулярніших бібліотек із розпізнавання тексту, та підібраний найкращий із варіантів.

Вирішення цих ключових проблем та якісна технічна реалізація й приведе до створення системи із розпізнавання реквізитів документів по зображенню.

Система розроблялася як окремий модуль, який можна буде із легкістю впровадити у будь-яку систему із організації роботи компаній.

Р