

ДОДАТОК ДЛЯ РОЗПІЗНАВАННЯ ПАСПОРТНИХ ДАНИХ ОФЛАЙН

Рогоза М. С., магістрант, rogoza02@gmail.com

Донецький національний технічний університет, Луцьк, Україна

У час цифрових технологій більшість людей звикли працювати із електронними документами у їх повсякденному житті — від написання робочих звітів до створення плану на день. Однак, не зважаючи на це, багато компаній, роботодавців та окремих фізичних осіб вимагають саме фото документів як підтвердження особи, посади або прав на майно. Вважаючи це за норму, або не маючи альтернативи, люди відправляють конфіденційну інформацію за наданою їм адресою запропонованим методом зв'язку.

Таким чином, навіть не знаючи про це, ці люди можуть стати жертвами різних форм шахрайства: як і через причину недостатньої захищеності засобу зв'язку, так і через причину надмірної довіри до сторони, що запитує особисту інформацію.

Враховуючи наведену вище інформацію, існує потреба у створенні додатку, що міг би уникнути згаданої проблеми, а також надати додаткові переваги, такі як зменшення людського фактору у процесі заповнення особистих даних, збереження часу на копіювання даних до баз даних шляхом відправки тільки необхідної інформації з фото або встановлення додаткового кодування у випадку передачі через незахищений канал зв'язку.

Постановка задачі

Актуальним є завдання створення додатку, що зможе зчитувати необхідні дані із фото паспорту України (ID-карти), перетворювати їх у текстовий формат та створювати закодований файл із ними, який може бути відправлений до отримувача, котрий заздалегідь знає формат отримуваних даних і має доступ до описаного файлу.

Структура пропонованої системи розпізнавання

Система складається із трьох основних функціональних частин:

- попередня обробка зображення;
- зчитування полів;
- кодування файлу з інформацією.

Попередня обробка також ділиться на дві частини: обрізання зображення, усунення геометричних деформацій та обертання; зміна показників яскравості, насиченості, контрастності тощо для покращення результатів розпізнавання.

Першим кроком у роботі є знаходження границь паспорту за допомогою спеціальних алгоритмів знаходження границь [1]. Після цього початкове зображення набуває чорно-білого вигляду, де грані документу досить легко знайти. Якщо взяти по дві точки на кожній із таких граней та провести

через них пряму, за допомогою математичних перетворень можна знайти як точку перетину цих прямих, що буде вказувати на кутові точки паспорту, так і кут нахилу (рис. 1). Конкретні математичні формули включають у себе знаходження кута між двома векторами та точки перетину двох векторів.

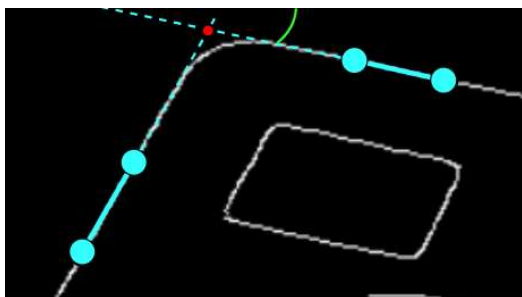


Рисунок 1. Розрахунок кута обертання (зеленим) та точки перетину (червоним)

Кут нахилу та точки перетину знаходяться для того, щоб правильно оцінити як саме застосовувати матрицю перетворень, що має загальний вигляд (1). Змінюючи відповідні параметри, можна створити матриці, які, при перемноженні із пікселями зображення, зможуть виконувати переміщення деякого діапазону точок (у даному контексті пікселів), їх масштабування, обертання на довільний кут, а також будь-яку комбінацію із цих трьох дій.

$$M = \begin{bmatrix} ax + by + c \\ dx + ey + f \end{bmatrix} \quad (1)$$

де: a, b, c, d, e, f — скалярні значення.

Тож, після знаходження кутових точок паспорту достатньо застосувати матрицю (1) з певними параметрами, щоб отримати готовий для розпізнавання документ.

У реальних задачах, однак, часто виникає ситуація, коли до описаних перетворень необхідні додаткові, що допоможуть усунути геометричні спотворення: перетворити довільний чотирикутник на прямокутник. У цьому випадку необхідно застосувати гомографічну матрицю виду:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2)$$

де: h_{ij} — скалярні значення.

Матриця (2) має схожі із (1) характеристики та застосування, але має ключову відмінність у тому, що з її допомогою на зображенні паралельні лінії можна перетворити на непаралельні та навпаки. Таким чином, за допомогою застосування (2) до зображення довільний чотирикутник можна перетворити на прямокутник.

Розрахунок обох цих матриць є досить складним завданням, але Python, а саме його бібліотека OpenCV надає відповідні функції, що полегшують їх застосування.

Безпосередньо перед початком розпізнавання необхідно виконати додаткову попередню обробку, що включає в себе зміну контрастності, яскравості, насиченості, та різкості. Значення цих параметрів, однак, залежить від початкової якості фото, тому їх бажано обирати заново кожного разу. Проте,

якщо початкове фото має прийнятну якість, невелика зміна параметрів, зазвичай, не погіршує результат.

Наступним кроком є оптичне розпізнавання символів (OCR). Так як на цьому етапі гарантовано отримується зображення, що є правильно обернене, без наявності спотворень та зайвих ділянок, розпізнавання можна здійснювати частинами, окремо аналізуючи кожний рядок документу.

Саме розпізнавання здійснюється за допомогою бібліотеки Pytesseract [2], що є імплементацією системи Tesseract для мови Python. Вона підтримує велику кількість мов, зокрема українську та англійську — єдині мови, текст яких присутній на паспорті.

Після переведення слів та дат на зображенні у текстовий формат їх бажано додатково закодувати. Це запропоновано виконувати за допомогою алгоритму кодування RSA [3].



Рисунок 2. Приклад вхідних даних для розпізнавання

```
OCR result:
ukr_s_name    РОГОЗА
eng_s_name    ROHOZA
ukr_f_name    МИХАЙЛО
eng_f_name    MYKHAILO
birth_date    09 01 2002
exp_date      [redacted]
doc_num       [redacted]
Enter 'y' if result is correct (y/n): y
Data saved to 'encoded'
```

Рисунок 3. Приклад підтвердження даних після розпізнавання

При тестуванні роботи програми, використовуючи фото, зображене на рис. 2, можна отримати результати, зображені на рис. 3. Ця інформація показується користувачу тільки для її підтвердження, після чого, за допомогою заздалегідь сформованого публічного ключа, відразу створюється закодований файл, який можна декодувати тільки на стороні отримувача за допомогою приватного ключа.

За замовчуванням програма працює для обох сторін паспорта та потребує два фото для повноцінної роботи. Всі дії, що були описані щодо попередньої обробки та розпізнавання, є релевантними і для зворотної сторони. Також за замовчуванням скануються не всі поля, а тільки ті, що є найбільш потрібними у загальному використанні.

Висновки та напрям подальших досліджень

Створено додаток, що зчитує необхідні поля з паспорта України та створює файл із закодованою інформацією про них.

Для подальшого покращення роботи можна звернути більш докладну увагу до процесу машинного навчання, а саме його застосування для знаходження документу на фото при великих рівнях шуму, підбору значень контрастності, яскравості, насиченості та різкості; а також навчання системи

безпосереднього розпізнавання на власних тестових даних з акцентом на паспорт України: його шрифт, розмір, а також збільшення бібліотеки слів з урахуванням прізвищ, імен тощо.

Література

1. Canny J. A computational approach to edge detection [Електронний ресурс] / John Canny // IEEE transactions on pattern analysis and machine intelligence. — 1986. — PAMI-8, № 6. — С. 679—698. — Режим доступу: <https://doi.org/10.1109/tpami.1986.4767851> (дата звернення: 11.11.2023). — Назва з екрана.
2. Smith R. An overview of the tesseract OCR engine [Електронний ресурс] / R. Smith // Ninth international conference on document analysis and recognition (ICDAR 2007) vol 2, Curitiba, Parana, Brazil, 23–26 верес. 2007 р. — [Б. м.], 2007. — Режим доступу: <https://doi.org/10.1109/icdar.2007.4376991> (дата звернення: 11.11.2023). — Назва з екрана.
3. Rivest R. L. A method for obtaining digital signatures and public-key cryptosystems [Електронний ресурс] / R. L. Rivest, A. Shamir, L. Adleman // Communications of the ACM. — 1983. — Т. 26, № 1. — С. 96—99. — Режим доступу: <https://doi.org/10.1145/357980.358017> (дата звернення: 13.11.2023). — Назва з екрана.

Анотація

Представлено процес розробки засобу оптичного розпізнавання символів на прикладі паспорту України (ID-картки) на основі використання бібліотек мови Python. Наведено покроковий опис перетворення зображення для операцій обрізання, зсуву та обертання. Описано процес попередньої обробки зображень із метою покращення якості розпізнавання. Вказано спосіб використання алгоритму кодування для документів.

Ключові слова: Оптичне розпізнавання символів, попередня обробка зображень, матриця перетворень, Python, Pytesseract, кодування тексту.

Abstract

The process of developing an optical character recognition tool on the example of a passport of Ukraine (ID-card) based on the use of Python language libraries. A step-by-step description of image transformation operations like crop, shift, and rotate is provided. The process of image preprocessing to improve recognition quality is described. The method of using the coding algorithm for documents is indicated.

Keywords: Optical character recognition, image preprocessing, transformation matrix, Python, Pytesseract, text encoding.