УДК 004.415.2

СУЧАСНІ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ РУКОПИСНИХ ТЕКСТІВ

Данило Денисенко

Національний авіаційний університет, Київ

Науковий керівник — Олександр Бобарчук, к.т.н.,
доцент кафедри КММТ

Анотація. Технології отримання якісних цифрових зображень з рукописних оригіналів постійно розвиваються та удосконалюються. На першому етапі залежно від якості та специфіки рукописного оригіналу можуть використовуються сканування та фотографування. Подальша обробка передбачає розпізнавання оцифрованих текстів. Розглянуті сучасні технології сканування та розпізнавання рукописних текстів.

Ключові слова: рукописні тексти, сканування, фотографування, розпізнавання, цифрові зображення.

Актуальність теми. В умовах сьогодення важливою стає діяльність зі збереження надбань історично-національної культури, матеріалів наукової цінності та забезпечення до них доступу без будь-яких просторових та часових обмежень шляхом створення електронних колекцій документального надбання [1]. Рукописні оригінали є надзвичайно важливими та цінними матеріалами, що зазвичай містять унікальну інформацію, переважно в одиничному екземплярі. Їх збереження для вивчення та передачі наступним поколінням є важливим завданням.

Виклад основного матеріалу. В процесі оцифрування будь якої графічної інформації відбувається перетворення нецифрової інформації або даних у цифрову форму з метою зберігання, зміни або обміну цими даними різними електронними пристроями [2].

Якщо питання оцифрування друкованих матеріалів та реальних фотозображень є достатньо вивченим, то питання оцифрування рукописних оригіналів знаходиться на етапі активного вивчення. Цифрове зображення з рукопису є обмеженим відображенням тексту, що міститься у фізичних артефактах [3]. Цифрове зображення не є ідентичним рукопису, а його візуальний вигляд не охоплює і не враховує всі якості рукопису.

У процесі отримання первинних цифрових зображень з рукописних оригіналів важлива роль належить використанню сучасних новітніх технологій оцифрування. Найпоширенішими є сканування та фотографування. Подальша обробка первинних цифрових зображень з рукописних оригіналів передбачає розпізнавання оцифрованих текстів. Для цього використовується технологія оптичного розпізнавання текстів. Існує декілька видів технологій розпізнавання символів, які можуть автоматично перетворювати рукописний або набраний текст у цифрові символи: оптичне розпізнавання символів (OCR); інтелектуальне розпізнавання символів (ICR); інтелектуальне розпізнавання слів (IWR). Розпізнавання рукописного тексту є складним завданням, адже, на відміну від однорідного друкованого тексту, рукописний може мати ряд специфічних особливостей. Наприклад, почерк може бути важким для зчитування, слова та букви можуть бути складними для інтерпретації, текст може бути безперервним, або на сторінці можуть бути відсутні абзаци або інші маркери поділу. Сторінки рукопису також можуть бути брудними, містити примітки, блиски, виправлення чи стирання [3].

З-поміж засобів, що використовуються для розпізнавання рукописного тексту можна назвати Google Cloud Vision API, Hanvon Technology, Hanwang Technology, Infrrd.ai, Microsoft Azure Read API, Mitek, MyScript, Selvasai, Unitek.ai, Vidado та ін.

Практична перевірка розпізнавання тестових рукописних оригіналів та їх перетворення в друкований текст показали, що найкращі результати досягаються в Google Keep. Цей ресурс надає можливості безкоштовно здійснювати якісне розпізнавання рукописних оригіналів з мінімальними спотвореннями символів. Практично всі слова і речення розпізнано точно і чітко, лише в окремих місцях можливе порушення структури документу. Ресурс надає можливість скопіювати матеріал в Google Документ, де можна провести подальше опрацювання тексту.

Висновки

Завдяки оцифруванню, рукописи стають візуальними об’єктами, що зберігаються на персональному комп’ютері чи в Інтернеті. У цифровому вигляді вони збережені від зникнення та доступні для ознайомлення та вивчення найширшому колу користувачів.

Список використаних джерел:

1. Павленко Т. Цифрова репрезентація книжкових пам’яток: підходи та шляхи реалізації. Бібліотечний форум: історія, теорія і практика. 2017. № 1. С. 11–13.

2. Що таке оцифрування? [Електронний ресурс] / Режим доступу: https://uk.theastrologypage.com/digitize

3. Liv Ingeborg. Digitization and Manuscripts as Visual Objects: Reflections from a Media Studies Perspective. [Електронний ресурс] / Режим доступу: https://brill.com/view/book/edcoll/9789004399297/BP000002.xml?body=fullhtml-43184