УДК 004.832

ПІДХІД ДО РОЗПІЗНАВАННЯ РУКОПИСНОГО ТЕКСТУ

Савчук Тамара, Пупко Олександр

Вінницький національний технічний університет

Анотація

В даній роботі запропоновано алгоритм розпізнавання рукописного тексту, що базується на сегментації зображень рукописних документів, що може бути реалізованим з використанням нейромережевого підходу, та покращить якість розпізнавання архівних документів.

Abstract

This paper proposes an algorithm for recognizing handwritten text based on image segmentation of handwritten documents, which can be implemented using a neural network approach and improve the quality of archival document recognition.

Рукопис залишається засобом спілкування та збору інформації в повсякденному житті навіть із впровадженням нових технологій. Більш того, величезна кількість історичних колекцій не є доступна у зручному форматі. В даний час підходи та способи перетворення зображень в цифровий текст розвиваються швидкими темпами, навіть незважаючи на те, що все ще є можливості для покращення та вирішення багатьох завдань.

В останні роки системи розпізнавання друкованого тексту стали досить ефективними. Рукописні цифри та інтерактивне написання розпізнаються якісно. Однак поточні технології все ще знаходиться на обмеженому рівні, щоб розпізнавати текстові зображення різних стилів та мов рукописного вводу. Системи розпізнавання рукописного тексту можуть відрізнятися в багатьох аспектах і залежать від конкретних завдань. Бінаризація може виконуватися як перший крок або після сегментованих рядків чи символів.

Як відомо, процес перетворення документів, що зберігаються в традиційних форматах (газета, книга і т. д.), в текстовий вигляд проводиться через багато етапів (дискретизація, бінаризація, придушення шумів, блок-сегментація, витяг рядків і символів, розпізнавання і т. д.).

Сегментація зображення – один з важливих етапів в системі оптичного розпізнавання тексту. Мета цього етапу – виділення на введеному документі тієї області, в якій представлений текст, і відділення текстової інформації, графічної [1].

На поточний момент величезна частина інформації зберігається в електронному вигляді. Пошук і витяг необхідних знань відбуваються набагато простіше завдяки напівавтоматичним системам навігації по різних корпусах тестів, зображень і відео. Для навігації по рукописних документах потрібно виділити об'єкти, присутні на зображенні, такі як текст, ілюстрації, друковані вставки.

Таким чином, одним із завдань комп'ютерного зору є завдання розпізнавання рукописного тексту, що базується на сегментації відсканованого документа. Сегментація є найважливішою складовою алгоритмів розпізнавання рукописних текстів. Але якість вихідного документа, особливо, при аналізі архівних документів, часто є дуже низькою. Походження дефектів обумовлено низькою якістю паперу, зношеністю документа, низьким дозволом при електронному скануванні [2].

Серед інших, найбільшого розповсюдження набули підходи до вирішення задачі сегментації тексту, що базуються на алгоритмах, заснованих на припущенні про горизонтальну орієнтацію сторінки з формуванням вихідних даних у вигляді друкованих матеріалів [2].

У першу групу алгоритмів можна віднести методи, засновані на розбиті сторінки на однорідні прямокутні блоки. За визначеними ознаками проводиться класифікація за обраним алгоритмом. Наприклад, документ розрізається на однорідні прямокутні блоки фіксованого розміру, для кожного блоку вважається дискретне перетворення Фур'є, потім блоки кластеризуються за допомогою методу k-середніх. На виході алгоритм повертає дві бінарних маски: для тексту і для зображень [3].

У другу групу алгоритмів відносяться алгоритми, засновані на аналізі кордонів бінаризованих компонент. На вхід подається документ, текст в якому розташований під одним і тим же кутом. Проводиться визначення орієнтації тексту за допомогою проекції інтенсивності зображення. Потім відбувається розмиття вихідного документа, поділ на неоднорідні блоки і обхід в глибину по межах блоків, на підставі якого відбувається злиття блоків в різні класи. На виході маємо бінарне зображення з кордонами тексту в документі. Після порогової бінаризації зображення відбувається вертикальне і горизонтальне згладжування, пошук меж, об'єднання різних компонент з евристичних міркувань, і розділення на компоненти. На виході алгоритм повертає 2 бінарних зображення: з текстом і з ілюстраціями.

Результатом роботи нині існуючих алгоритмів сегментації є бінарна маска, що виділяє текст на зображенні. Проте, в запропонованих алгоритмах не виконується декомпозиція бінарної маски на окремі компоненти зв'язності, які необхідні для сегментації сторінки на кілька класів. До того ж вихідними даними в основному є друковані матеріали з однаковою орієнтацією, а не рукописний текст. Ця особливість вихідних даних створює складності через різноманітність сегментів, різних шрифтів авторів.

Методи бінаризації, як першого кроку розпізнавання історичного тексту, поділяються на глобальні та локальні (адаптивні) [3]. Глобальні методи використовують одиничне порогове значення для відокремлення чорно-білих пікселів, тоді як декілька значень використовуються для визначення локального порогу на основі локальної області (рис. 1б) [3]. Локальні методи визначення порогу є ефективними у випадку різноманітних фонових компонентів, різного роду кольору, текстури або яскравості зображення, видимості тексту (рис. 1в) тощо.

Рисунок 1 - Адаптивна бінаризація середнього порогу: а) вихідне зображення; б) бінаризаване; в) гаусівська адаптивна бінаризація

Після бінаризації, необхідно виокремити області, де розташований текст (рис. 2а) з подальшою сегментацією текстових рядків, яка ґрунтується на проекції гістограми (рис.2б). На рисунку 2б видно, що кожна текстова лінія відповідає піку в гістограмі. Пусті простори між піками представляють можливі області між різними текстовими рядками. Для вирішення цієї проблеми задаються порогові та мінімальні відстані.

Наступним кроком є обробка зображень кожної лінії сегментованих рядків зверху на виявлених контурах, одночасно не включаючи контури менше встановленого порогу. При отриманні контурів, символи на краях зображення можуть спричинити проблему визначення всіх символів на краю як одного контуру [4]. Це питання вирішується за допомогою простого та швидкого способу додати білі межі на 1 піксель на кожній стороні зображення.

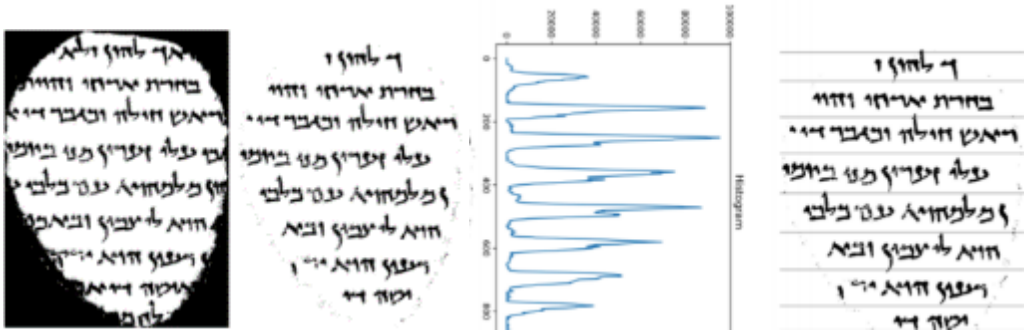

Рисунок 2 – Обрізане зображення і зображення з налагодженим фоном та сегментація ліній тексту за допомогою гістограмної проекції

Таким чином, запропонований підхід до розпізнавання рукописного тексту для історичних документів, передбачає:

- бінаризацію;
- пошук текстової області;
- коригування фону, сегментацію ліній, реконструкцію символів та обробку шумів.

При цьому, розпізнавання літер може бути реалізовано за допомогою нейронної мережі [5, 6]. Це дозволить вирішити одну із задач, що зустрічаються при обробці зображення для розпізнавання, а саме – автоматизовану сегментацію зображень рукописних документів.

Після розпізнавання може виконуватися додаткова корекція, що дозволить збільшити якість розпізнавання суперечних символів (тобто символів у яких є кілька кандидатів з приблизно однаковою оцінкою ступеня відповідності декільком еталонам) за допомогою бінаризації. Означений підхід дозволить підвищити якість розпізнавання рукописного тексту для історичних документів за рахунок автоматичної сегментації зображень рукописних документів.

Список використаних джерел

1. Алексєєв А. Алгоритм розпізнавання символів на основі структурного підходу / А. Алексєєв, В. Заяць, Д. Іванов // Вісник Нац. ун-ту «Львівська політехніка «Комп'ютерна інженерія та інформаційні технології». – 2002. – № 468. – С. 129–133.

2. Чуба Б. О. Програмний засіб розпізнавання символів на тлі завад на основі нейронної мережі / Б. О. Чуба, О. К. Колесницький // Науково-практична конференція «Сучасні тенденції розвитку системного програмування». К.: – 2016. – С. 34–42.

3. Fischer A. Character prototype selection for handwriting recognition in historical documents with graph similarity features / Fischer A., Bunke H. // Proc. 19th European Signal Processing Conference. – 2011. – pp. 1435–1439.

4. Neha Gupta, V .K. Banga, Image Segmentation for Text Extraction. – 2012.

5. Савчук Т.О. Математична модель розпізнавання символів на основі нейроних мереж/ Савчук Т.О., Ярема Є.О. Матеріали VIII Міжнародної НПК "Інтелектуальні системи прийняття рішень та інформаційні технології". м.Чернівці: Рута. – 17-19 травня 2006. – С.75-77.

6. Савчук Т.О. Використання нейронних мереж для розпізнавання символів/ Савчук Т.О., Ярема Є.О. Науково-технічний журнал «Реєстрація, зберігання і обробка даних». м.Київ, Інститут проблем реєстрації інформації НАН України, ISSN 1560-9189 – 2005. – Том 7, №4. - С.78-84.