УДК 004:931

МЕТОДИ РОЗПІЗНАВАННЯ ТЕКСТУ

Репік С. І., Штогріна О. С.
Інститут телекомунікаційних систем НТУУ «КПІ», Україна
E-mail: sergrepik@gmail.com, l.shtogrina@gmail.com

Methods of text recognition

This paper presents handwriting recognition method that is based on the methods of comparison with etalon. In this method the step on which sample symbols is comparing with etalons was modified. When there is a letters comparison it used their center of mass. Contributed modification increases the recognition rate. This is possible by reducing the number of operations of overlay letters on etalon, when they compares.

Підчас повсякденної діяльності державні структури, бізнес, наукові та навчальні організації використовують велику кількість паперових документів, більшість з яких є рукописними. Велика кількість даних та знань міститься в друкованих або рукописних документах, які є архівними. Зростає потреба в оцифруванні паперових документів, з метою подальшої обробки їх вмісту автоматизованими комп'ютерними системами.

Розпізнавання тексту можна розділити на кілька напрямів, які досить суттєво відрізняються методами їх вирішення. Текст може бути друкований чи рукописний. Будь який з них може бути додатково структурований. Наприклад, формули можуть містити різні рівні записів, такі як надстрочні, підстрочні, спеціальні позначки математичних дій, тощо.

На сьогоднішній день існує ряд методів, що вирішують проблему розпізнавання друкованого тексту, однак досі не існує систем, здатних розпізнавати будь-який рукописний текст [1]. Існуючі системи можуть досить не якісно розпізнавати конкретні почерки. Отже актуальною є задача розробки методу розпізнавання рукописного тексту, який дозволить опрацьовувати рукописні документи.

Перед тим як відбувається розпізнавання тексту, завжди існує попередня обробка вхідного зображення. Першим її етапом є покращення якості зображення. На цьому етапі підвищують контрастність та різкість зображення, а також здійснюється фільтрація від шумів. Наступним етапом є сегментація [2], за допомогою якої визначається структура тексту. Сегментація дещо відрізняється для друкованого та рукописного тексту. В обох випадках виділяються строки, слова та літери, але для друкованого тексту сегментація літер набагато простіша та відбувається подібно до сегментації слів – за допомогою методу горизонтальних та вертикальних профілів [3]. Для рукописного тексту сегментація на рівні літер складніша: літери можуть зливатися у один сегмент або навпаки, одна літера розпадатися на кілька сегментів. Це суттєво ускладнює задачу розпізнавання.

Після попередньої підготовки, методи розпізнавання рукописного і друкованого тексту відрізняються більш суттєво. Для друкованого тексту

застосовується порівняння сегментованих літер з еталонами із різноманітних шрифтів. Знайшовши співпадіння з одним із них, наприклад, першої літери, залишок тексту, циклічно розпізнається за рахунок порівняння усіх виділених сегментів з літерами визначеного шрифту.

Для рукописного тексту порівняння з еталонами шрифтів принципово не можливе – кожен почерк є унікальним, отже еталону, на зразок шрифту, не існує. Спочатку необхідно створити базу еталонів конкретного почерку, з якою надалі відбуватиметься порівняння.

Існує два підходи до розпізнавання рукописного тексту – розпізнавання онлайн та офлайн [4]. Перший передбачає розпізнавання безпосередньо підчас написання тексту і використовує алгоритми написання символів, котрі враховують траєкторію руху «пера» – предмету, яким здійснюється написання. Цей підхід називають розпізнаванням онлайн. На сьогоднішній день, задачу такого розпізнавання для більшості мов можна вважати вирішеною. Сучасні електронні записники широко його використовують. Другий підхід має на меті розпізнавання рукописного тексту, який було написано заздалегідь. Розпізнавання офлайн набагато важливіше, так як кількість вже написаного тексту величезна. Проблема саме такого розпізнавання у загальному випадку досі не вирішена.

Існує два основні типи методів вирішення задачі розпізнавання тексту офлайн – структурні та еталоні, а також їх комбінації [4].

Структурні методи засновані на виділенні та аналізі різних структурних елементів символу, їх ознак та властивостей. Кожна літера розбивається на вузли та криві, що їх з’єднують. На основі набору таких даних робиться висновок, яка літера написана. Однак існує проблема, пов’язана із тим, що більшість літер написані не каліграфічно та, відповідно не мають чітких з’єднань.

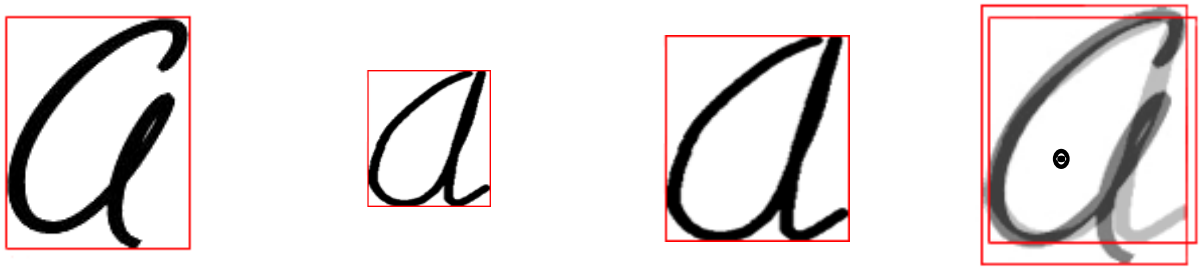
Еталоні методи передбачають порівняння заданого, не розпізнаного символу з набором деяких еталонів. Для цього використовують нейронні мережі, які необхідно заздалегідь заповнити еталонами. Існують кілька алгоритмів порівняння тексту з еталоном. Найпростіший варіант – попіксельне порівняння, однак для нього необхідні рівні по розмірам зображення, що порівнюються. Інші варіанти – накладення та накладення зі зміщенням, у яких зображення ставляться у відповідність одне одному. Методи еталонного порівняння рукописного та друкованого тексту, на перший погляд, схожі, однак суттєво відрізняється. Еталони рукописного тексту, на відміну від літер шрифту, лише грубі зразки. Це значно збільшує вірогідність помилки.

В роботі запропоновано метод розпізнавання рукописного тексту, що базується на методах еталонного порівняння, в якому модифіковано етап порівняння зразку літер з еталоном таким чином, що при співставленні літер враховуються їх центри мас. Внесена модифікація дозволяє збільшити швидкість розпізнавання за рахунок зменшення кількості операції накладання літер одна на одну при порівнянні з еталоном.

Спочатку необхідно провести попереднє навчання для кожного окремого зразка почерку, створивши для цього необхідну базу з еталонами. Літери

вписуються в прямокутник таким чином, щоб його сторони були дотичними до них. Літера, що показана на рис. 1.а – еталон, на рис. 1.б – літера, виділена з тексту за допомогою сегментації. Додатковою умовою на співставлення є однакові масштаби літер. Так як розмір виділеного сегменту для літери, яку необхідно розпізнати, та літери з бази можуть відрізнятися, то на другому кроці потрібно провести масштабування (рис. 1.в). Часто лінійні розміри виділеного сегменту та шаблону в горизонтальній та вертикальній площинах одночасно можуть не співпадати, так як навіть літери, написані однією людиною частково відрізняються. Однак достатньо співпадіння по горизонталі чи по вертикалі (рис. 1.а, 1.в).

Співставлення виділеного сегменту з еталонами відбувається таким чином, що їх центри мас співпадають (рис. 1.г). Літера, що має найбільше співпадіння із літерою з бази вважається її копією, а також може бути додана в базу для подальшого навчання.



а) б) в) г)

Рис. 1. Приклад зіставлених по горизонталі літер.

Запропонований метод надає можливість розпізнавати рукописний текст заданого зразку. Його перевагою є можливість порівняння літер, що не мають однакових лінійних розмірів. За рахунок того, що достатньо зіставити лише центри мас, підвищується швидкість точного накладання літер, за рахунок зменшення кількості накладань та взаємних зсувів зображень для правильного позиціонування.

Література

1. Кучуганов А. В., Лапинская Г. В. Распознавание рукописных текстов / А. В. Кучуганов, Г. В. Лапинская // Материалы международной научной конференции, Ижевск, 13–17 июля 2006. – С. 98 – 103.
2. Shafait F. Performance Comparison of Six Algorithms for Page Segmentation / F. Shafait, D. Keysers, T. Breuel // Image Understanding and Pattern Recognition (IUPR) research group. – 2006. – pp. 12.
3. Запрягаев С. А., Сорокин А. И. Сегментация рукописных и машинописных текстов методом диаграмм Воронного / С. А. Запрягаев, А. И. Сорокин // Вестник ВГУ, серия: системный анализ и информационные технологии, № 1, 2010. – С. 160 – 165.
4. Васильев С. Распознавание непрерывного рукописного текста в режиме off-line [Електронний ресурс]. – Режим доступу: https://geektimes.ru/post/136165/ – Електрон. тестові дані (дата доступу 02.03.2016).