ІСТОРІЯ АЛГОРИТМУ РОЗПІЗНАВАННЯ ТЕКСТІВ

Гродецький Д. І.
Житомирський державний університет імені Івана Франку, м. Житомир

Обчислювальні системи існують протягом десятиліть, їх основне призначення – це можливість замінити людину і виконання замість неї трудомісткої роботу. До таких систем належать системи розпізнавання тексту [1]. Алгоритми розпізнавання тексту використовуються в різних сферах. Їх потрібно використовувати для оцифрування старих книг, переведення тексту зображень в електронний вигляд, для полегшення процесу визначення поштових індексів та ідентифікаційних номерів паспортів.

Розпізнавання тексту – це важке для реалізації завдання. Проблеми, з якими стикаються розробники програм розпізнавання тексту є накладанням символів один на одного, їх схожості у різних мовах, низької якості зображення і також шуму на зображенні.

Останні публікації підтверджують актуальність такої роботи. Навіть відомі лідери програмних пакетів оптичного розпізнавання символів (OCR), призначені для вирішення подібних проблем, не завжди можуть впоратися з розпізнаванням звичайних зображень, хоча текст легко читається візуально. Зазначена проблема розглядається в [2]. У 1929 році Густав Таушек запатентував метод оптичного розпізнавання тексту в Німеччині, а потім Гендель, який запатентував цей метод у Сполучених Штатах у 1933 році. У 1935 році метод Таушека також отримав патент США. Машина Таушека — це механічний пристрій, який використовує шаблон і фотодетектори.

У 1950 році Девід Х. Шепард, криптоаналітик Служби безпеки збройних сил США, проаналізував проблему перетворення друкованої інформації на машинну мову, створену комп'ютером, і побудував машину для вирішення цієї проблеми. Перша комерційна система була встановлена в Reader's Digest у 1955 році. Друга система була продана Standard Oil і використовувалася для зчитування кредитних карток для обробки чеків [3].

Топологічні особливості, особливості форми та багато інших ознак, які традиційно вважаються ефективними для розпізнавання зображень, не є інформативними. Низька роздільна здатність і низька

якість можуть викликати помилки на етапі попередньої обробки, особливо пропуск рядків. Найпростіший і швидкий спосіб – відсканувати документ за допомогою сканера. В результаті виходить цифрове зображення документа - графічний файл. Текстове подання інформації зручніше, ніж графічне. Щоб алгоритм розпізнавання працював належним чином, якість вхідного зображення має бути якомога вищою. Якщо зображення шумне, розмите і низька контрастність, це ускладнює завдання алгоритмів розпізнавання.

Наразі латинські символи в друкованому тексті можна точно розпізнати лише за допомогою чіткого зображення, наприклад, відсканованого друкованого документа.

Розроблений алгоритм використовує компонентну архітектуру класифікаторів, організовану у вигляді дерева, листи якого є простими класифікаторами, а внутрішні вузли відповідають операціям, що об'єднують результати нижнього рівня (рис. 1).

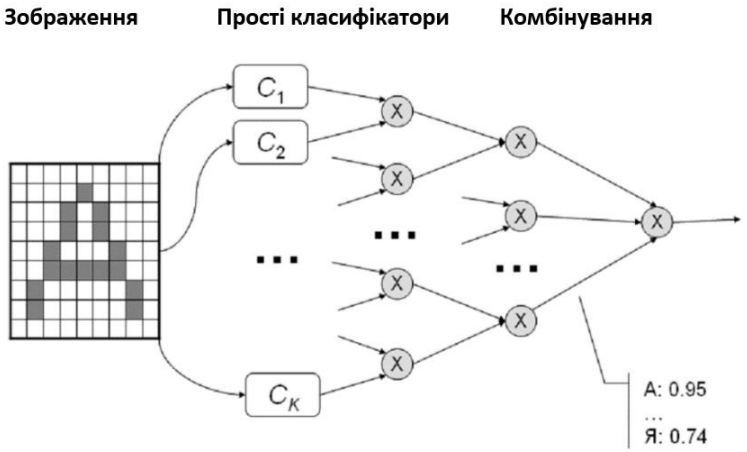

Рис. 1. Архітектура класифікатора.

Простий класифікатор працює в два етапи (рис. 2). Спочатку функції обчислюються на основі вихідного зображення. Значення кожної функції є функцією яскравості деякої підмножини пікселів зображення. У результаті на вхід нейронної мережі надходить вектор символічних значень. Кожен вихід мережі відповідає літері в алфавіті, а отримане значення сприймається як ступінь належності нечіткої множини.

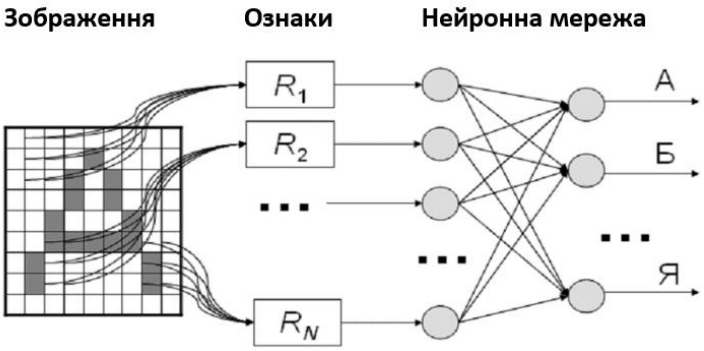

Рис. 2. Простий класифікатор

Результатом роботи класифікатора є нечітка множина, отримана в результаті комбінування на верхньому рівні. На останньому етапі приймається рішення про найбільш правдоподібний варіант прочитання слова. Для цього використовуються рівні можливості прочитання окремих літер, міжлітерної сегментації та частоти поєднань літер.

Загалом, можна зробити висновок, що, хоча запропонований алгоритм не такий хороший, як найкращі в своєму класі комерційні продукти від Abbyy, він здатний розпізнавати текст нижчої якості, ніж система з відкритим кодом CuneiForm. Тому технологія OCR дозволяє точно розпізнавати текст у зображеннях, якщо текст не спотворений безпосередньо. Це можуть бути випадки накладання символів або подібності зображення одного символу до іншого (наприклад, комбінацію літер «LI» можна ідентифікувати як «U»). Але поки що не розроблено жодної технології, яка б дозволяла автоматично розпізнавати текст без помилок, тому проблема все ще потребує втручання людини.

ДЖЕРЕЛА

1. Бройдо, В. Л. Вычислительные системы, сети и телекоммуникации / В. Л. Бройдо. — СПб.: Питер, 2004. – 703 с.

2. Богданов В., Ахметов К. Системы распознавания текстов в офисе. Компьютер-пресс. 1999, №3, с.40-42.

3. Оптическое распознавание символов URL: https://www.wikiwand.com/ru/Оптическое_распознавание_символов